Genomic and transcriptomic signatures of inflammation and malignancy in the intestinal tract

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Daniela Esser

Kiel, Februar 2019

Erster Gutachter: Prof. Dr. Philip Rosenstiel

Zweiter Gutachter: Prof. Dr. Thomas Roeder

Tag der mündlichen Prüfung: 30.04.2019

Table of contents

I List of figures

II List of tables

III Abbreviations, units, and symbols

Abbreviation	Meaning
5′meC	5-Methylcytosine
A	Adenine
ACF	Aberrant crypt foci
AJ	Apical junction
AJC	Apical junction complex
AOM	Azoxymethane
AS_DGN	Antisense downstream gene neighborhood
ASE	Antisense exonic
ASI	Antisense intronic
AS_UGN	Antisense upstream gene neighborhood
ATP	Adenosine triphosphate
Avg.	Average
bp	Base pair
BLAST	Basic local alignment search tool
BLAT	BLAST-like alignment tool
C	Cytosine
CAC	Colitis-associated colorectal cancer
CD	Crohn's disease
CIN	Chromosomal instability
Cosmic	Catalogue of somatic mutations in cancer
CNV	Copy number variation
cPI	Cancer proliferation index
CRC	Colorectal cancer
CRI	Cancer-related inflammation
DAI	Disease activity index
DALM	Dysplasia-associated lesions or masses
dbSNP	Single nucleotide polymorphism database
DGI	Downstream gene intersection
DGN	Downstream gene neighborhood
DNA	Deoxyribonucleic acid
DSS	Dextran sodium sulfate
e.g.	For example (abbreviation of Latin ‘exempli gratia’)
EBV	Epstein-Barr virus
ECM	Extracellular matrix
ECS	Exonic gene conservation score
EBV	Epstein-Barr virus
ELD	Exon-linked downstream
ELU	Exon-linked upstream
ePCR	Emulsion polymerase chain reaction
eQTL	Expression quantitative trait locus
ESP	Exome sequencing project
et al.	And others (abbreviation of Latin ‘et alii / et aliae’)
ExAc	Exome Aggregation Consortium
FA	Focal adhesions

FATHMM	Functional analyses through hidden markov models
FPKM	Fragments per kilobase of exon per million fragments mapped
G	Guanine
GC	Gastric cancer
GO	Gene ontology
GWAS	Genome-wide association study
HGMD	Human gene mutation database
H.pylori	Helicobacter pylori
IBD	Inflammatory bowel diseases
ICGC	International cancer genome consortium
i.e.	That is (abbreviation of Latin 'id est')
IEC	Intestinal epithelial cell
IGE	Intronic gene element
InDel	Insertion / deletion
INTER	Intergenic element
IQR	Interquartile range
ISE	Intron-spanning element
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAM	Methylazoxymethanol
MDS	Multidimensional scaling
mg	Milligram
ml	Milliliter
MMR	Mismatch repair
miRNA	Micro ribonucleic acid
mRNA	Messenger ribonucleic acid
MSI	Microsatellite instable
MSS	Microsatellite stable
NCBI	National center for biotechnology information
NGS	Next generation sequencing
NLR	NOD-like receptor
NLM	Non-negative matrix factorization
NMF	Non-negative matrix factorization
NO	Nitric oxide
nt	Nucleotide
nTar	Novel transcriptionally active regions
O^6-meG	O6-methylguanine
OMIM	Online mendelian inheritance in man
p	P-value
PCA	Principle component analysis
PCoA	Principle coordinate analysis
PBS	Phosphate buffered saline
PSI	Percentage spliced in / Percentage splicing index
RefSeq	Reference sequence database
RNA	Ribonucleic acid
ROS	Reactive oxygen species
RNA-Seq	Ribonucleic acid sequencing
rRNA	Ribosomal ribonucleic acid

RSS	Residual sum of squares
SCNA	Somatic copy number alteration
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SV	Structural variant
T	Thymine
TJ	Tight junction
TCGA	The Cancer Genome Atlas
UC	Ulcerative colitis
UCSC	University of California, Santa Cruz
µg	Microgram
UGI	Upstream gene intersection
UGN	Upstream gene neighborhood
µl	Microliter
UTR	Untranslated region
vs	Versus
WES	Whole exome sequencing
WGS	Whole genome sequencing

1 Introduction

Cancer causes the highest number of deaths among all diseases occurring in economically developed countries and is on the second rank in developing nations [1]. In recent decades, major advancements in the understanding of epidemiology, pathology, and pathogenesis of different cancer types as well as in the development of novel treatments have been achieved [2–4]. The course of disease and therapy success are, amongst others, influenced by molecular factors such as transcriptional and mutational patterns [3]. Therefore, a better understanding of genomic mechanisms underlying tumor development in e.g. gastric cancer (GC) and inflammation-triggered colorectal cancer could improve the patient's outcome. The application of next generation sequencing (NGS) has enabled researchers to acquire genome-wide insights into the mutational landscape of an individual tumor type at nucleotide level [4–7]. Thereby, all published large-scale data sets from cancer studies revealed an unforeseen complexity of the molecular landscape and highlighted the extensive dynamics behind tumor initiation and progression.

1.1 General characteristics of cancer

A tumor is a population of abnormal cells, which are characterized by continuous growth and uncontrolled cell divisions [8]. Tumor development is a multistep process caused by a combination of environmental and genetic factors [3, 9, 10]. At least three to seven mutations are necesssary, whereby patients with germline mutations can be predisposed to cancer [10]. In the first phase of cancer, genetic modifications as well as altered activation of processes associated with the cell cycle cause an increased proliferation rate and extended lifespan of a cell. During clonal expansion, further variants supporting the selective advantage may arise. This results in benign adenomas of increasing size, which can later develop to malignant carcinoma. In the final phase, cancer cells invade surrounding normal tissues and organs, grow in the mesenchyme at the primary site, and may start to metastasize by infiltrating blood and lymphatic vessels [8, 11]. More than hundred distinct tumor types affecting all kinds of tissues exist. These tumor types can vary in their pathological and molecular characteristics, aggressiveness as well as in their response to specific therapies [11]. However, many features of pathogenesis are common between all tumor types, such as altered genes and processes involved in survival, proliferation, and metastasis. Six common characteristics were reported as hallmarks of cancer by Douglas Hanahan and Robert Weinberg in 2000 [8]. These include (i) 'self-sufficient in growth signal', (ii) 'insensitivity to growth-inhibitory signals', (iii) 'evasion of programmed cell death (apoptosis)', (iv) 'limitless replicative potential', (v) 'sustained angiogenesis' meaning the formation of new blood vessels from pre-existing vessels, and (vi) 'tissue invasion and metastasis'. In the year 2011, Hanahan and Weinberg extended these

features by the following four characteristics [2]: (i) metabolism reconstruction to enable continuous cell growth and proliferation, (ii) immune system evasion by cancer cells demonstrating that the immune system can support or suppress tumor development and progression, (iii) genomic instability including random mutations and chromosomal rearrangements to create genomic diversity and to strengthen cancer cells by increasing the activity of cellular proto-oncogenes or reducing the power of tumor suppressor genes [12], and (iv) infiltrating cells of the innate and adaptive immune system to enhance tumorigenesis and progression by providing cancer-supporting substances to the tumor microenvironment, such as growth, proangiogenic, and survival factors as well as extracellular matrix-modifying enzymes that enable invasion and metastasis [2].

To characterize tumors, Sobin *et al.* developed the most commonly used TNM system for cancer staging [13], which was also applied for the GC samples during my studies (box 1).

Box 1 TNM cancer staging system

In the TNM cancer staging system, each letter followed by a number describes one specific feature of the cancer stage: 'T' followed by a number between zero and four describes the size and extent of the primary tumor. Thereby, a higher value indicates a larger tumor. The letter 'N' refers to the number of lymph nodes that are affected by cancer and located nearby, while 'M' delivers information about the existence of metastasis. The nomenclature can be optional extended by further tumor characteristics: The parameter 'G' describes the histopathological grading of the cancer cells using a scale ranging from one to four. One refers to a tumor cell structure, which is similar to normal cells, while four characterizes poorly differentiated cancer cells. The letters 'L' or 'V' indicate whether lymphatic vessels or veins are invaded. Finally, an 'R' indicates whether tumor removal was complete or whether resection-boundaries still harbor cancer lesions after surgery.

Example: 'T4 N1 L0 V0 R0 G2' refers to a tumor, which is growing into the subserosa layer, has spread to one or two nearby lymph nodes, but neither lymphatic vessels nor veins are invaded. The tumor cells would be moderately differentiated.

In summary, cancer describes a group of complex diseases caused by environmental and genetic factors. For many kind of tumors, the molecular background is not completely understood yet and therefore, successful therapies are missing. Two examples for solid cancers are sporadic gastric carcinoma and inflammation-associated colorectal cancer, which were investigated in my studies.

1.2 Gastric carcinoma

1.2.1 Characteristics of gastric cancer

Gastric cancer (GC), also named as stomach cancer, is a complex disease, which is caused by environmental and genetic factors [14, 15]. The main cause for tumor development is an infection with the bacterium *Helicobacter pylori* (*H. pylori*). This factor is involved in 65-80% of all GCs [16], while heritability contributes to around 28% of the susceptibility to GC [15]. Also obesity and lifestyle influence the development of GC. While smoking and consumption of some food products, such as fat, red meat, alcohol, chili pepper, and salt lead to a higher GC risk, intake of fresh fruits, vegetables (allium family), and certain micronutrients (selenium, vitamin C) have a protective effect [14]. Other risk factors, including infection of Epstein-Barr virus (EBV) or JC virus (human polyomavirus) and radiation, have probably only a small impact on GC [14].

Almost one million new cases of GC are diagnosed every year worldwide. Thus, GC is still ranked together with cancers of lung, colorectum, and breast among the four most common malignancies in the world and causes 10% of all cancer deaths worldwide [1, 17], although GC rates have decreased in most parts of the world since the early 1980s [18]. Diagnoses of GC are more abundant in Western and Northern Europe, Australia, New Zealand, and North America, while GC is less often observed in Asia and Saharan Africa [1]. Geographical variations are mainly based on dietary differences and the incidence of *H. pylori* infection [1], but GC is also unequally distributed within each population. In particular, GC is two times higher in males than in females [1]. The cancer risk steadily increases after the age of 40 years with the main diagnosis peak in the seventh decade of life [17]. Four treatment approaches are currently used in cancer therapies: Removal of the tumor and lymph node dissection is the most effective option in early stage GC. This surgery is often combined with a more unspecific chemotherapy, which regulates cell division and growth signals. As alternative, GC can be treated with targeted therapies, in which drugs or antibodies are applied to identify and attack cancer cells. This method is very precise and reduces damage to healthy tissue. A similar principle is used in immunotherapy, in which the immune system of the patient is triggered with immune tumor vaccines or antitumor antibodies. Besides the options of surgery and drug treatment, radiation can be applied to attack cancer cells [19]. Although all described therapy forms are applied and continuously improved, GC shows a 5-year survival rate of only 10-30% [17]. The low treatment success of GC is mainly due to a first diagnoses at an advanced, mostly metastasized stage [17]. One reason for late detection is a lack of specific symptoms in early stage GC. Complaints in an early stage include only features such as an impaired digestion (dyspepsia) [17]. In a more advanced stage, GC is characterized by weight loss, the

eating disorder anorexia, vague abdominal pain, nausea, vomiting or bloody stool, and early satiety [17].

In summary, although the number of GC diagnoses decreases in developed countries and treatment approaches are improving, GC is still one of the malignancies with highest incidence rate and a poor survival prognosis [14, 17, 19]. GC is often first diagnosed at a late stage, for which appropriate therapies are still lacking. A better understanding of underlying molecular processes contributing to GC would help to improve existing therapies and to develop novel treatment approaches [17]. Thereby, it is important to consider also differences between tumor types and stages.

1.2.2 Tumor types and stages in gastric cancer

Gastric malignancies can be categorized in six classes according to histological characteristics: While the relative abundances of gastric lymphomas (1-5%), gastrointestinal stromal tumors (2%), carcinoids (1%), adenocanthomas (1%), and squamous cell carcinomas are low, most tumors belong to the group of adenocarcinomas (90-95%) [20]. Based on histological and growth patterns, the latter can be further subdivided into the two main categories of the intestinal (54%) and the diffuse type (32%) [17, 21]. Intestinal type tumors feature a decreased amount of stroma as well as an irregular flat cell shape. The cell structure is similar to those in a healthy mucosa [17, 22]. This tumor type is mainly triggered by environmental factors, is often a complication of multifocal atrophic gastritis and is more frequent among elderly males [17, 22]. In most cases, intestinal type tumors arise from intestinal metaplasia as preliminary stage. The tumor origin is more frequently located in the proximal part, but the tumor can spread from the surface epithelium to the antrum, which is an expansion of the esophagus near the stomach [17, 23]. In general, intestinal-type tumors grow by extension rather than by infiltration [17]. In the diffuse type, cells lack cohesion. This leads to reduced gland formation and enables the tumor to invade into all layers of the stomach wall [17, 22, 23]. In contrast to the intestinal type, the diffuse type arises equally often in both genders, is associated with an onset in younger patients, and has mostly a genetic origin. The diffuse type is more aggressive and characterized by a worse outcome than the intestinal type [17, 23, 24].

The pathogenesis of sporadic GC is mainly based on either the microsatellite stable (MSS) or the microsatellite instability (MSI) pathway. Microsatellites are defined as repeats of up to six nucleotides, which can occur in coding sequences of genes involved in structural and functional processes as well as in non-coding regions [25]. In MSI tumors, polymerase slippage during DNA replication causes aberrant lengths of microsatellite DNA [25]. Depending on the study cohort and the investigated markers, MSI occurs in 5% to 50% of the individuals with sporadic GC [25]. The MSS type is characterized by chromosomal instability and inactivation

of tumor suppressors, while the MSI tumor type features a defective DNA mismatch repair system resulting in an increased number of somatic variants especially in repeat regions [25, 26]. MSS and MSI differ not only by alteration types but also in the set of affected genes and pathways as well as in the etiopathology [25, 26]. In contrast to MSS type cancer, MSI tumors arise more often in elderly female individuals, in antral regions, and are more frequently characterized by larger size, intestinal Laurén histotype, lower tumor stage at diagnosis, and lower number of affected lymph nodes [26]. Thus, MSI is associated with a better survival for patients with stage II tumors and can be used as a prognostic marker [26].

In summary, the most common GC type is an adenocarcinoma of the intestinal type, which can be further distinguished by the microsatellite status. Therefore, a comprehensive investigation of one sporadic MSS and one sporadic MSI adenocarcinoma of the intestinal type was performed during my studies. Besides revealing novel insights into GC, the molecular background of inflammation-associated colorectal cancer was analyzed in a mouse model.

1.3 Link between chronic inflammation and cancer

1.3.1 Role of inflammation in the tumor development

Inflammation is a key factor of the immune system and involved in the defense against microbial infections as well as in tissue repair and regeneration [12, 27]. However, chronic inflammation can also contribute to tumor development, including malignant transformation of cells and tumor expansion [28]. Thereby, inflammatory pathways can activate processes involved in proliferation, cell survival, and angiogenesis [12, 28]. The connection between carcinogenesis, host defense, and tissue repair is, amongst others, based on genes playing a role in several of these processes. As example, cytokines and proteins involved in the Wnt / β-catenin pathway contribute to tumorigenesis as well as to proliferative processes used for tissue repair and maintenance [12]. Moreover, chronic inflammation can facilitate genomic alterations such as DNA strand breaks, aberrant DNA cross-linking, and somatic mutations. In turn, this can cause activation of proliferation, cell survival pathways, and oncogenes, while tumor suppressors get dysfunctional [12, 28]. The mentioned alterations are e.g. caused by reactive oxygen species (ROS), which are accumulated by inflammatory processes as part of the host's defense mechanism [12]. ROS result from the normal cellular metabolism and are involved in the adaptation to changing environmental conditions. They are for example produced by host enzymes, which have an especially high activity during inflammation [12, 29]. This imbalance is defined as oxidative stress and can potentially lead to a cancer predisposition or can contribute to tumor progression [12, 28].

Besides inflammation as part of the immune system, inflammatory processes are also upregulated in the microenvironment of most neoplastic tissues to contribute to tumor

development and progression [30]. Thereby, cancer cells can modify the microenvironment by soluble mediators, which influence the innate and adaptive immune system and lead to the ability to escape T cell attack [27]. As example, extreme upregulation of chemokines and cytokines lead to a desensitization of receptors and in turn to missing immune response of the host [31]. Moreover, inflammatory processes impact growth and differentiation of all cell types in the tumor microenvironment, including neoplastic cells, fibroblasts and endothelial cells. Thus, cancer-related inflammation (CRI) was suggested as the seventh hallmark of cancer [30]. CRI is characterized by infiltration of leukocytes, tumor-associated macrophages as well as the occurrence of mediators like cytokines and chemokines [30]. Later in the tumorigenic process, tumor cells use signaling molecules of the innate immune system, such as selectins, chemokines, and matrix metalloproteases (MMPs) to support survival, proliferation, and spread of the malignant cells [12, 31]. These tumorigenic processes are accompanied by destruction of the adaptive immune system and increasing resistance against drugs [30, 31]. As further tumor-supporting factor, CRI might contribute to genetic instability, damaged repair mechanisms, and removal of cell cycle checkpoints, which lead to an increased number of genetic variants in the cancer cells [27, 31].

Taken together, inflammation and cancer are closely linked to each other. On the one hand, tumors are characterized by increased inflammation in the tumor microenvironment [30], on the other hand, chronic inflammation supports the development and promotion of cancer [28]. The highest incidence of malignant transformations triggered by inflammation exists in the large intestine, where colorectal cancer develops as complication of inflammatory bowel diseases (IBD) [31].

1.3.2 Pathogenesis and pathophysiology of inflammatory bowel disease

Inflammatory bowel diseases are chronic inflammatory disorders in the gastrointestinal tract, which are characterized by alternating periods of inflammatory episodes and inactive phases. The two main subtypes comprise Crohn's disease (CD) and ulcerative colitis (UC) [32]. CD can occur in inflammatory patches in the complete gastrointestinal tract but is most commonly present in the terminal ileum (small intestine) and the rectum [32]. The inflamed parts are characterized by thickened submucosa, transmural inflammation, granulomas as well as deep ulcers, which can extend into the complete mucosa. In CD, complications such as strictures, abscesses, and fistulas may occur [32]. In contrast to CD, UC affects only the colon as well as rectum and usually spreads from the proximal part in distal direction [32]. Thereby, the whole area is affected without uninflamed part in between, but inflammatory lesions are limited to mucosa and submucosa [32]. The colon wall is thinner and may harbor ulcers (sores) without extension beyond the mucosa [33].

In clinic, CD and UC are considered as distinct IBD subtypes, although both disease forms have common symptoms such as fever, abdominal cramping pain, chronic flare-ups of inflammation, signs of malnutrition, weight loss, diarrhea (often bloody in UC), and eventually anemia caused by gastrointestinal bleeding [34]. IBD can develop at any age, but the diagnosis peak for CD as well as for UC is in the early adulthood [33]. Up to now, it exists no medical therapy to cure IBD. Thus, only symptoms can be addressed as well as remission phases achieved and extended. In some cases, it is necessary to remove parts of the colon or even the complete colon (colectomy) by surgery to control the disease or to treat complications [33]. One serious complication of both IBD types is the development of inflammation-triggered colorectal tumors introduced in the next chapter (section 1.3.3) [35, 36].

Internationally, between 2.2 and 14.3 per 100,000 individuals get affected by UC per year, while the incidence for CD ranges from 3.1 to 14.6 cases per 100,000 people per year [33]. The number of diagnoses steadily increased in the last years [37], which might contribute to the high differences in reported IBD prevalences. IBD occurs more often in urban than in rural areas and the risk of disease depends on the lifestyle [33, 37]. In particular, environmental triggers, such as nutrition and smoking, influence the risk of IBD development [32, 37]. The prevalence of IBD is also associated with the country and is e.g. higher in the developed world [33]. This can be explained by the so-called ‘hygiene hypothesis’, which postulates that exposure to infectious agents can reduce the susceptibility to IBD by activating the immune system development during childhood [37].

Besides the mentioned factors and the composition of the intestinal bacterial flora, genetic factors influence the risk of IBD [32, 37]. This was shown by familial clustering differences [38]. The chance that both siblings of monozygotic twins are affected by CD was estimated at 20-50%, while dizygotic twins grown up together show a concordance rate of less than 10% [38]. In UC, the corresponding probabilities are 14-19% (monozygotic twins) and 0-7% (dizygotic twins), respectively [38]. Although already 163 susceptibility loci have been identified for IBD (110 shared by UC and CD), only 13.6% and 7.5% of the total disease variance can be explained for CD and UC, respectively. [39]. Known risk loci are mainly associated with the immune system. This include e.g. cytokine production, lymphocyte activation, response to microorganisms, and the JAK-STAT signaling pathway. Remarkably, all functional groups and pathways enriched for UC- or CD-associated genes are affected in both diseases [39].

In summary, IBD is a group of chronic diseases, whose origin is connected to genetic and environmental factors. One serious complication of IBD is the development of colitis-associated colorectal cancer (CAC).

1.3.3 Colitis-associated colorectal cancer

Chronic inflammation in the intestinal tract increases the susceptibility to CAC, which is responsible for ~15% of all deaths in IBD patients [40]. Though the link between chronic inflammation and CAC development were reported in many studies [41], little is known about the underlying molecular mechanisms of CAC.

Around 3% (8%) of CD and 2% (18%) of UC patients develop colorectal cancer within 10 years (30 years) after disease onset [35, 36]. In individuals with longstanding IBD, the risk of CAC is similar between CD and UC [42]. Thereby, patients in a remission phase have the same risk to develop CAC like patients within an active stage [40]. The incidence of CAC depends on the duration of symptoms, severity of colonic inflammation and the extent of the disease, with the highest risk for patients whose entire colon is affected (pancolitis) [40]. In addition, age at onset of IBD, genetic factors, and affection by other chronic inflammatory diseases, such as primary sclerosing cholangitis, influence the risk for CAC development [40, 42, 43]. Interestingly, some environmental factors might have a contrary effect on UC and CD. As example, it was suggested that smoking protects from CAC in 50% of the UC patients, while the incidence of CAC is four times higher in smoking CD patients than in non-smokers [29]. The risk for the development of CAC might be also different between populations, but it has to be proven that this association is not based on distinct treatment approaches [29, 40]. In comparison to sporadic CRC (mean age: 65), IBD-associated cancer occurs at younger age (mean age: 54.5 (CD) and 43 (UC)) and causes a more aggressive cancer type with higher invasion potential as well as poorer prognosis in the advanced stage. No differences regarding the survival of the patient were observed in the early disease phase [44, 45].

In 73% of the UC and 79% of the CD cases, CAC develops via dysplasia [40]. Features of dysplasia include a decreased number of epithelial crypts, an altered crypt structure characterized by distortion and branching, and larger hyperchromatic cell nuclei [42]. ‘Dysplasia-associated lesions or masses’ (DALMs) have a raised mucosa and are considered as precancerous lesions, because adenocarcinoma develop preferably in these areas [42]. Therefore, it was suggested to classify DALMs as malignant and treat them like tumors [42]. In contrast to CAC development via DALMs, sporadic CRC is characterized by precancerous polypoid adenomatous polyps, which undergo mutations and may progress to invasive carcinoma [42]. Besides the different precancerous lesions, both CAC and sporadic CRC are multistep mutational processes and molecular changes like chromosomal instability (CIN), MSI, loss of heterozygosity, and aberrant methylation patterns occur in both cancer types. However, while the first two listed characteristics appear with the same incidence between the two cancer types (85% CIN, 15% MSI), the onset and frequency for the latter two alteration types are different [29, 46, 47].

Tumors of CAC are usually nonpolypoid, flat, and discrete [29, 40, 42]. In general, CAC is characterized by a higher fraction of mucinous and signet ring cell histology [29]. Tumors in CAC can arise in all areas of the colon and often several primary tumors develop independently in different sections at the same time [44]. However, tumors tend to arise in colonic parts that are influenced by IBD and thus, the predominant location differs between UC and CD [29, 44]. All other known clinicopathological features of carcinoma occur in UC and CD. This indicates that tumor development in UC and CD is driven by similar carcinogenic processes [44].

Improvements of medical and surgical techniques as well as higher rates of colonoscopy as part of a preventative examination accounted for a better prognosis of CAC in the last decades. The treatment for CAC includes the use of chemopreventive agents like 5-amino-salacylic acid or ursodeoxycholic acid preparations, folic acids, statins, early use of steroids in flare-up episodes, and proctocolectomy as last possibility of intervention. Although therapies for CAC exist, this IBD complication is not curable without colectomy yet [29, 40, 42].

Taken together, CAC is a serious complication of IBD and is characterized by a more aggressive tumor type than sporadic CRC. Nevertheless, the underlying genetic processes are mainly unknown. To fill this gap, inflammation-associated colorectal cancer was investigated with the AOM/DSS mouse model in my studies.

1.3.4 AOM/DSS mouse model

Inflammation-associated colorectal cancer can be investigated with a well-established mouse model, in which the synergetic effect of azoxymethane (AOM), a colon procarcinogen in rodents, and the tumor-promoting substrate dextran sodium sulfate (DSS), which has toxic effects on the colonic epithelium, are applied [48, 49].

The murine AOM/DSS model is characterized by high potency as well as reproducibility and shares many featured with human colorectal cancer [48]. Like sporadic CRC in humans, AOM/DSS-induced tumors occur very often in the distal part of the colon, initiate with polypoid growth, and harbor similar histopathological aberrations [48, 50]. CAC is better reflected in a chemically induced than in a genetic knockout model, because environmental factors promote this cancer type development in humans. However, metastasis and mucosal invasiveness occur only in very rare cases of AOM-induced cancer [29, 37, 50].

In the AOM/DSS model, the procarcinogen AOM has to be activated by metabolic processes to cause DNA damage [50, 51]. In detail, after intraperitoneal injection, AOM is transported via bloodstream to the liver, where it is hydroxylated to methylazoxymethanol (MAM) [50, 51]. Afterwards, MAM is excreted from the liver and reaches the intestine. After absorption by the colon, further metabolic steps can lead to the addition of a methyl group obtained from MAM compounds to a guanine in the DNA, which might lead to mispairing of guanine with thymine

[51, 52] (Figure 1-1). DSS is a non-genotoxic agent with proinflammatory impact on epithelial cells in rodents [50, 53]. Dissolved in drinking water, DSS causes severe colitis and bloody diarrhea [50]. The proinflammatory DSS can strengthen the carcinogenic effect of AOM resulting in increased tumor growth and decreased latency time [48].

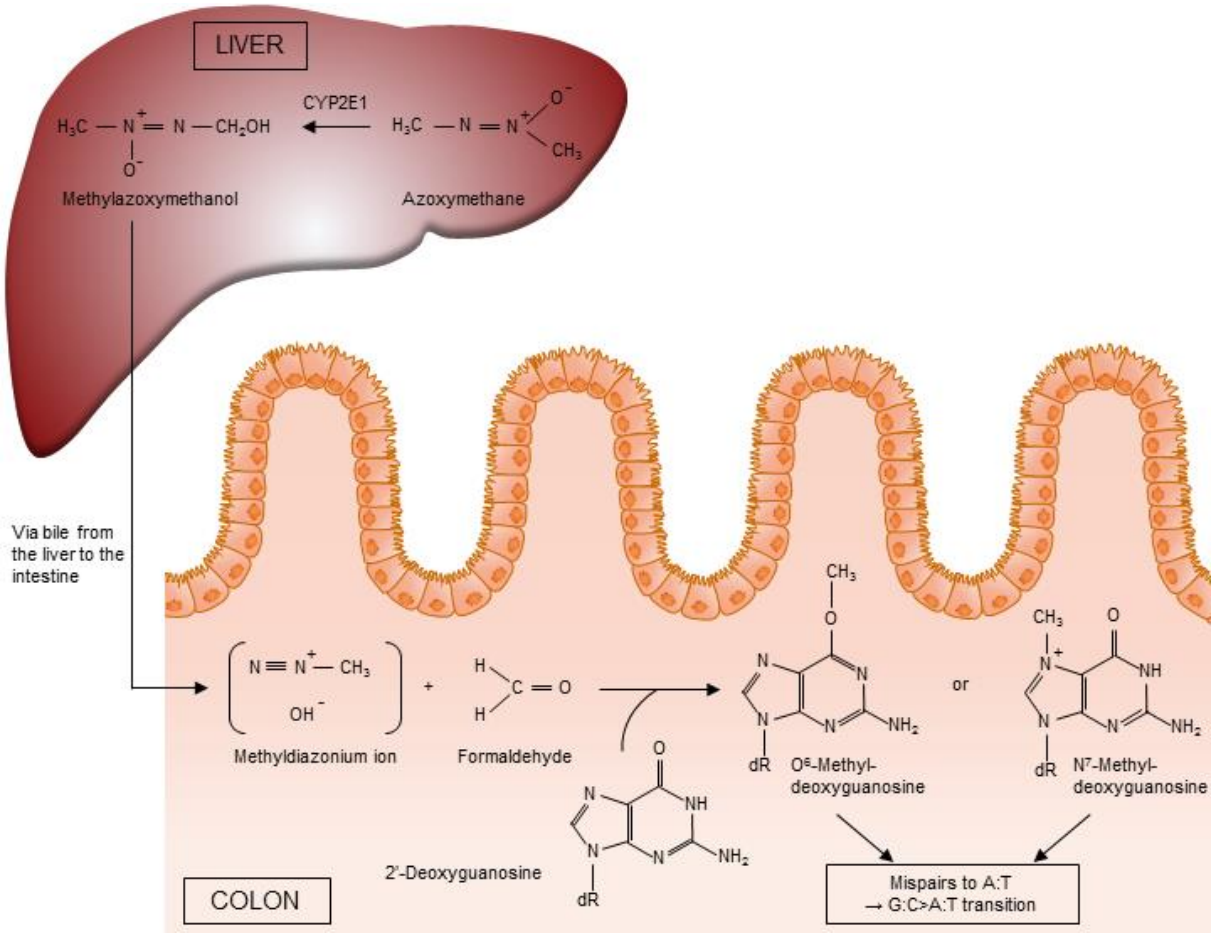


Figure 1-1 Metabolic activation steps of AOM to form DNA-reactive products
First, AOM is transported via the bloodstream to the liver, where it is hydroxylated by the alcohol-inducible cytochrome P450 isoform CYP2E1 (cytochrome P450 2E1) to MAM [50, 51]. Subsequent to the excretion, MAM gets to the intestine via blood or bile. Bacterial β-glucuronidase can additionally increase the amount of free MAM in the colon [50, 52]. After absorption by the colonic epithelium, MAM is metabolized or spontaneously decompose to a methyldiazonium ion [50–52]. The methyldiazonium ion is able to alkylate macromolecules, such as guanine. This can lead to N^7-methyl-deoxyguanosine or O^6-methyl-deoxyguanosine by adding the methyl group to the N^7 or O^6 position of guanine, respectively [51]. The latter mentioned modification can cause a mispairing during DNA replication, which leads to a base substitution from G:C to A:T [51].

Taken together, the AOM/DSS mouse model is a common approach to investigate inflammation-triggered colorectal cancer, although not all features of human CAC are reflected. To interpret results of experiments based on the AOM/DSS model correctly and to enable inferences on the human disease, deeper insights into involved molecular processes would be necessary.

1.4 Genetic factors of gastric cancer and inflammation-associated colorectal cancer

Many published NGS-based studies provided a comprehensive molecular characterization of different cancer types including genome-wide insights into the mutational landscape of

individual tumors at base-pair resolution [3, 5, 54, 55]. Most of these studies were based on whole genome sequencing (WGS) or whole exome sequencing (WES) technologies. This led to the detection of a huge number of potential cancer-driving genomic alterations, including single nucleotide variants (SNVs), small insertions or deletions (InDels), copy number variations (CNVs), and structural variants (SVs). Besides common cancer-associated molecular changes, this enabled also the identification of tumor type specific signatures described more detailed in section 1.6 [3, 5, 54, 55]. However, most of these studies did not include CAC, why the knowledge about the molecular background of this disease is still limited. Additionally, GC studies reflected a high heterogeneity of GC and pinpoint the importance of a deeper genomic characterization of GC.

Four major studies investigating the genetic basis of GC were published. A very comprehensive study comprising matched sample pairs from 295 patients with gastric adenocarcinoma was performed by The Cancer Genome Atlas (TCGA) research network [7]. The authors suggested to assign GC into four subtypes differing by molecular characteristics: (i) Tumors with EBV infection that harbor recurrent *PIK3CA* mutations, overexpression of *JAK2*, *CD274*, and *PDCD1LG2,* as well as DNA hypermethylation; (ii) MSI tumors characterized by a diagnosis in older patients compared to other subtypes as well as high mutation rates resulting in alterations of genes that encode for relevant signaling proteins with oncogenic features; (iii) Tumors with stable genomes harboring often a diffuse histological subtype and overexpression of cell adhesion pathways, mutated *RHOA* as well as fusions of RHO-family GTPase-activating proteins; (iv) Chromosomally unstable tumors (CIN) characterized by a high mutation rate in *TP53*, aneuploidy as well as overexpression of receptor tyrosine kinases.

In a second study, Wong et al. [56] used WGS to investigate 49 GCs with diffuse and intestinal histological subtypes as well as matched peripheral blood samples. They observed frequent mutations in *TP53*, *ARID1A*, *TGFBR2, CDH1*, *SYNE1,* and *TMPRSS2* as well as in Ephrins and SLIT/ROBO signaling pathway genes. Moreover, three mutational signatures with alterations at TpT, CpG, and TpCp[A/T] nucleotides were identified. Intestinal type tumors featured a higher degree of clonality and ploidy. These samples were also often affected by *TP53* mutations as well as overexpression of receptor tyrosine kinases, cell signaling and cell cycle-related genes. Besides the intestinal type, Wong et al. described two major subtypes of the diffuse type: The first one is characterized by less genetic changes and low clonality, while the second one harbors a comparatively high clonality and mutations at TpT dinucleotide positions.

A third major GC study was published by Wang et al. [57] using WGS of 100 tumor-normal pairs. They confirmed previously known (e.g. *TP53*, *ARID1A* and *CDH1*) and reported new (e.g. *MUC6*, *CTNNA2*, *GLI3*, *RNF43*) altered driver genes. Thereby, mutations occurred most

often in genes associated with adherens junctions and focal adhesion (FA). Interestingly, mutations in *RHOA* were exclusively observed in tumors of the diffuse type, while they played obviously no role in the intestinal type. Additionally, Wang et al. described a higher mutation rate in MSS compared to MSI cancers, while EBV tumors featured a reduced number of demethylated positions and increased hypermethylation in the entire genome.

The largest study combining transcriptome and exome sequencing with SNP arrays and a meta-analysis of previous GC studies was published by Liu et al. (2014) and revealed clear differences between MSI and MSS tumors [58]. While alterations of cancer drivers, such as *ARID1A*, *APC*, and *CTNNB1,* were more often observed in MSS tumors, oncogenes, such *KRAS* and *ERBB2,* were mostly affected in MSI tumors [58]. All four main studies are summarizes in Table 1-1.

Study	Main findings
TCGA [7]: Methylation profiling, mRNA + miRNA sequencing, WES, SNP array of matching samples from 295 patients	• Classification into four subtypes: EBV-infected tumors, MSI tumors, tumors with stable genome, and CIN tumors
Wong et al. [56]: WGS of 49 GCs with diffuse and intestinal histological sub-type + matched peripheral blood samples	• Mutations in *TP53*, *ARID1A*, *TGFBR2, CDH1*, *SYNE,1*, *TMPRSS2*, Ephrins, and genes of the Slit / Robo signaling pathway • Division into three mutational signatures • Classification into the intestinal type and two diffuse subtypes
Wang et al. [57]: WGS of 100 tumor-normal pairs	• Confirmation of *TP53*, *ARID1A*, and *CDH1* as cancer driver genes • Identification of *MUC6*, *CTNNA2*, *GLI3*, *RNF43* as cancer driver genes • Association of *RHOA* mutations with diffuse-type tumors • Top perturbed pathways: adherens junction and focal adhesion • Mutation rate: MSS > MSI • EBV tumors: demethylation ↓, hypermethylation ↑
Liu et al. [58]: WES, SNP array, and RNA-Seq of 51 tumor-normal pairs and 32 cell lines, meta-analysis	• 170 genes differentially spliced in GC • 55 splice site mutations associated with different isoform usage • Molecular differences between MSI and MSS tumors

Table 1-1 Summary of main findings reported in the four largest NGS-based GC studies

NGS-based projects have rarely assessed human CAC. To the best of my knowledge, only two high-throughput studies were published so far: (i) Robles et al. [59] performed WES with colorectal tumor and matching control tissues from 31 patients with CAC. While the mutation rate of *TP53* was similar to the one in sporadic CRC (63%), *APC* and *KRAS* were less often affected in CAC. Also genes of the Rho and Rac GTPase network were affected by recurrent mutations pointing to an involvement of the noncanonical Wnt signaling pathway in the CAC pathogenesis. This was supported by frequent alteration of the Wnt pathway components *SOX9* and *EP300*. Further potentially CAC-associated genes included the ERBB ligand *NRG1* and the cytokine *IL16*. (ii) Watanabe et al. [60] compared the expression patterns of non-neoplastic rectal mucosae between UC patients with and without colorectal cancer using the microarray technology. They reported 40 genes with different expression level between these groups. These genes were enriched in the categories signal transduction, receptor

activity, receptor binding, and lipid metabolism. Genes with an elevated expression in cancer patients included *CDKN1C (*cyclin-dependent kinase inhibitor), *STK39 (*serine threonine kinase), *NOL3* (apoptosis inhibitor), as well as *LRP5* and *LRP6* (low-density lipoprotein receptor–related proteins). The last-mentioned genes are involved in the transduction of intracellular signals in the Wnt pathway. The activity levels of *LRP5* and *LRP6* are also linked to cancer cell proliferation, why these genes were suggested as oncogenes [60].

Besides human CAC studies, inflammation-associated colorectal cancer was investigated in the AOM/DSS mouse model. Only one study was based on NGS, which was published during the time of my own experiments [61]. In short, controls from untreated mice, tumor, aberrant crypt foci (ACF), and control samples from mice treated seven weeks with AOM/DSS as well as tumor and matching controls from mice treated ten weeks with AOM/DSS were investigated with WES [61]. The most striking mutated pathways were extracellular matrix (ECM) receptor interaction and FA, which are both involved in cell-cell adhesion. The authors reported a high diversity between the samples and sample groups, which was not a result of mutated mismatch repair (MMR) genes. The genomic landscape of the investigated tissues from AOM/DSS-treated mice was also different from those observed in human CRC. Especially noticeable was a lack of mutations in genes associated with human CRC including *Apc*, *Trp53*, and *Kras* [61].

Several other studies investigated the molecular background of AOM/DSS-induced cancer with the microarray technology. These studies revealed upregulation of the Wnt inhibitory factor *Wif1,* the plasminogen activator *Plat,* the oncogene *Myc*, the phospholipase *Plscr2* as well as CRC-associated pathways, ubiquitin and spliceosome pathways, and Wnt signaling that regulates the expression of inflammatory genes [62–64]. Downregulation was observed for the peroxisome proliferator activated receptor binding protein *Pparbp*, the growth factor *Tgfb3* as well as for *Vegf* signaling and the apoptosis pathway [63]. Further differentially expressed genes and processes associated with murine AOM/DSS-triggered colorectal cancer include *Asprv1*, *Slc16a10*, *Pacsin3*, *Sycn*, *0610005c13Rik*, *Orc2*, *Orc5* as well as inflammatory cytokines and chemokines playing a role in the tumor microenvironment [64]. Not surprisingly, also known cancer-associated pathways were reported to be altered [62–64]. Moreover, metabolic imbalance was observed in inflammation-associated cancer [64]. This points to a close interaction between metabolic and inflammatory substrates in the tumor microenvironment [64]. Moreover, Tang et al. [65] could show that not all processes are involved in all steps of the tumor development of AOM/DSS-associated colorectal cancer. For example, p38 Mapk and Wnt / β-catenin signaling are only involved in tumor initiation and a late stage, respectively, while NF-κB and the STAT3 signaling pathway are overexpressed in all phases of CAC [65]. All main results from the major high-throughput studies investigating inflammation-associated colorectal cancer in humans or in the AOM/DSS mouse model are summarized in Table 1-2.

Study	Main findings
Robles et al. [59]: WES of colorectal tumor and non-neoplastic tissues from 31 patients with CAC	• Mutation rate of *TP53* similar between CAC and CRC • *APC* and *KRAS* less often mutated in CAC than in CRC • Recurrent mutations in Rho and Rac GTPase network • Novel associations with CAC: Wnt pathway (*SOX9* and *EP300*), ERBB ligand *NRG1*, and cytokine *IL16*.
Watanabe et al. [60]: Microarray expression analyses of rectal samples from 53 UC-patients (10 with CAC)	• Altered activation of signal transduction, receptor activity, receptor binding, and lipid metabolism in CAC patients • Upregulation of *CDKN1C*, *STK39*, *NOL3*, *LRP5,* and *LRP6* in CAC patients
Pan et al. [61]: WES of tumor, ACF, and matching controls from 10 AOM/DSS-treated mice	• ECM-receptor interaction and FA pathways enriched for mutations • High heterogeneity between samples and sample types • Different molecular mechanisms compared to human CRC
Suzuki et al. [62]: DNA microarray analysis on non-tumorous mucosa from ICR mice that received AOM/DSS, AOM alone or DSS alone, and untreated mice	• Upregulation of *Wif1, Plat, Myc*, *Plscr2* • Downregulation of *Pparbp* and *Tgfb3*
Gao et al. [63]: mRNA and miRNA microarray analyses of control, DSS-treated, AOM-treated, and AOM/DSS-treated mice (three samples each)	• Upregulation of Wnt signaling, ubiquitin and spliceosome pathways as well as CRC-associated pathways • Downregulation of apoptosis and VEGF signaling pathways • Altered activation of cellular metabolism and immune system
Li et al. [64]: Microarray expression analyses of control, DSS-treated, AOM/DSS-treated mice (12 samples each)	• Altered activation of *0610005c13Rik*, *Asprv1*, *Slc16a10*, *Pacsin3*, *Sycn*, *Orc2*, and *Orc5* • Altered activation of cancer-related pathways, inflammatory cytokines, and chemokines involved in the tumor microenvironment • Upregulation of Wnt signaling • Interaction between metabolic and inflammatory substrates in the tumor microenvironment in CAC suggested
Tang et al. [65]: Microarray expression analyses of AOM/DSS-treated ICR mice; three mice each were samples after 14, 28, 42, 56, and 140 days	• Overexpression of NF-κB and STAT3 signaling in all stages of CAC • Upregulation of P38 Mapk and Wnt / β-catenin signaling only at an early and late stage of CAC, respectively

Table 1-2 Summary of major studies investigating inflammation-associated colorectal cancer

In several studies, molecular differences between human CAC and sporadic CRC were reported (Figure 1-2). Alterations in the tumor suppressor gene *APC* as well as in the oncogenes *KRAS* and *BRAF* are often initiating events in CRC, while these events are less frequent or at a later stage in CAC [66, 67]. Additionally, although variants in *TP53* were observed with a high frequency in both sporadic CRC and CAC, they arise at different points in time [67, 68]. Alterations in *TP53* develop at a late stage in CRC, while intestinal epithelial cells (IECs) are often affected by *TP53* mutations before malignant transformation in CAC [67–69]. Mutant p53 supports constitutively activated NF-κB, which triggers chronic inflammation and tumor development [68]. A further gene, which is activated at different stages in CAC and CRC, is *TNFα.* While the gene is first upregulated in an advanced stage in CRC [70], early activation was observed in CAC [71]. Anti-TNFα drugs are medicative for the treatment of IBD patients. Also in AOM/DSS-treated mice, administration of anti-TNFα leads to smaller tumors

and a reduction of the tumor count [66, 72]. Selected known differences between CAC and CRC are described in Figure 1-2.

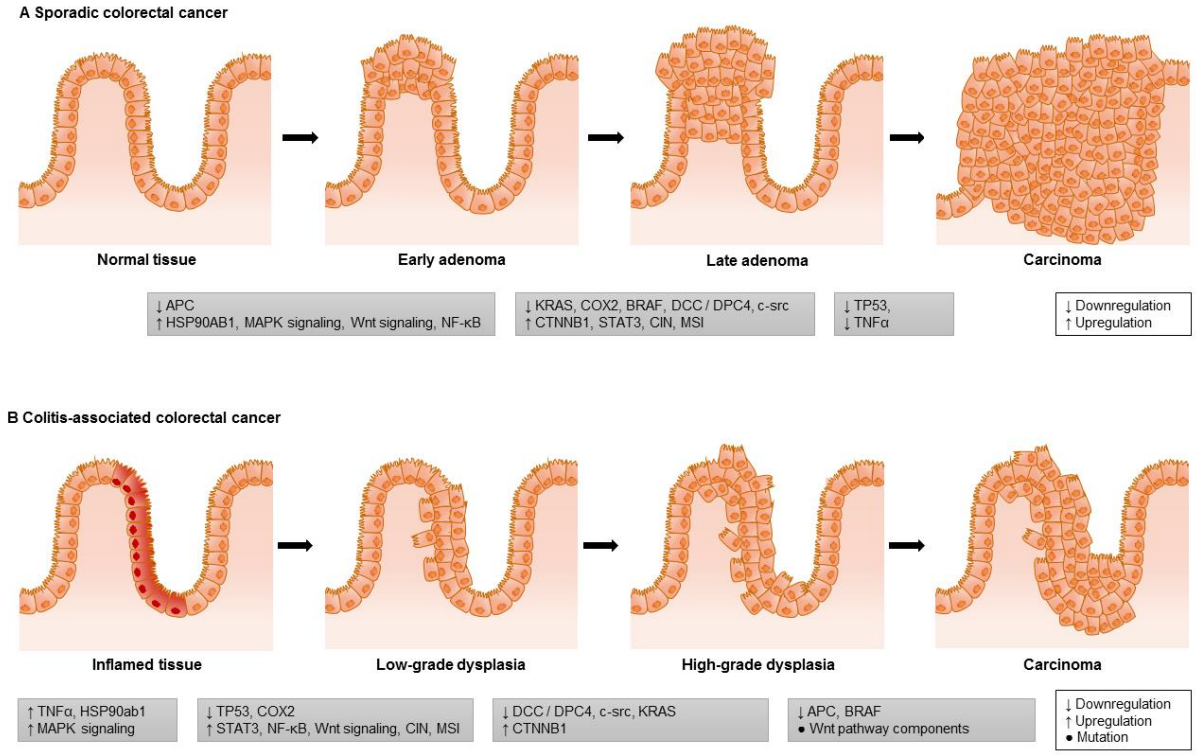


Figure 1-2 Molecular differences between CRC and CAC
The figure shows the point in time of molecular changes in (A) sporadic CRC and (B) CAC. The figure is based on Lasry et al. [73] and was modified with data from Subramaniam et al. [66], Itzkowitz et al. [74], von der Kraak et al. [71], Kim et al. [75], Rhodes et al. [47], and others [65, 76–81]. Besides differences of genes and processes, also CIN and MSI occur at different stages.

In summary, several studies investigating the genetic background of GC or inflammation-triggered cancer, which features many differences compared to sporadic CRC, were already published. However, publications based on an unbiased NGS approach are rare. In particular, only one WES and no RNA-Seq based study investigating murine AOM/DSS-induced colorectal cancer exist. In my studies, it was also the first time that two different NGS technologies were applied to get insights into the molecular background of GC.

1.5 Next generation sequencing

The Human Genome Project finished the sequencing of the first draft of the human genome in 2001 [82]. 20 institutes and companies in six countries collaborated for eleven years to reveal the first view on the human genetic code. Thereby, the primary applied approach was shotgun BAC (bacterial artificial chromosomes) for Sanger sequencing [82, 83]. The first NGS machine was produced in the mid-2000s. Due to parallelization of the sequencing reactions, the sequencing costs were 50,000-fold cheaper than in the human genome project. Further massive price reductions and output increases led to the possibility to nowadays sequence a human genome for around US$ 1000 in less than one week [84, 85].

Notable advantages of NGS are a high reliability of each called nucleotide, high robustness, and low background noise [84, 85]. Low sequencing costs and rapid investigation of genomic changes contributed to the fact that NGS became routine part of biological and medical research in the last years [84]. Furthermore, due to the avoidance of bacterial cloning of DNA, a cell free system can be used [83], which reduces the amount of required work and time. Limitations of NGS include the time-consuming and complex bioinformatic analyses, which result e.g. from the relatively short read length and the huge amount of produced data [83, 84]. Besides the need of fast and effective computational algorithms, this leads to a considerable higher need of data storage and computing power than previously developed technologies such as SNP-array, microarray or Sanger sequencing [83]. A further disadvantage of NGS is based on the need of expensive lab equipment. Nevertheless, NGS is nowadays established as key technology in basic science and translational research such as clinical diagnostics and personalized medicine [83, 84].

NGS can be divided into two main categories, namely sequencing by ligation and sequencing by synthesis [84]. During the course of this thesis, the Illumina's HiSeq 2500 (Sequencing by Synthesis) and Applied Biosystems / Life Technologies' SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequencing technologies were used. In both methods, the first step is the library preparation, which starts with the fragmentation of the DNA followed by size selection and the ligation of specific adapters to both fragment ends. Illumina uses an approach, in which the cluster generation is performed with bridge amplification in the first step (Figure 1-3). This results in one cluster of around 1000 identical sequences, which are all bound on a glass surface, for each original DNA molecule. The sequencing starts with the binding of a sequencing primer to the adapter. Labeled nucleotides are added and cause a fluorescence signal during incorporation. Due to the sequence clusters, the signal is intensified and thus strong enough for detection. After cleaving off the 3' terminator and the fluorescent group of the nucleotide, the cycle is repeated [83–85].

In the SOLiD system (Figure 1-3), an emulsion polymerase chain reaction (ePCR) on the surface of beads is performed in water-in-oil emulsion droplets after library preparation. This leads to a clonal amplification of the oligonucleotide fragments resulting in many copies of one DNA template immobilized on each bead. The subsequent sequencing process starts with the hybridization of a sequencing primer to the adapter at one fragment end. In the second step, an octamer consisting of a fluorescence labeled probe and an anchor sequence binds to the 5' end. Thereby, each color encodes for two adjacent nucleotides. Thus, each base is covered twice resulting in a higher sequencing accuracy. After color imaging, the ligated oligonucleotide is cleaved and the fluorophore removed. Cycles consisting of ligation, detection and cleavage are performed ten times. Thereby, two out of every five bases are investigated. Afterwards, the process is repeated with new offset anchors [83–86].

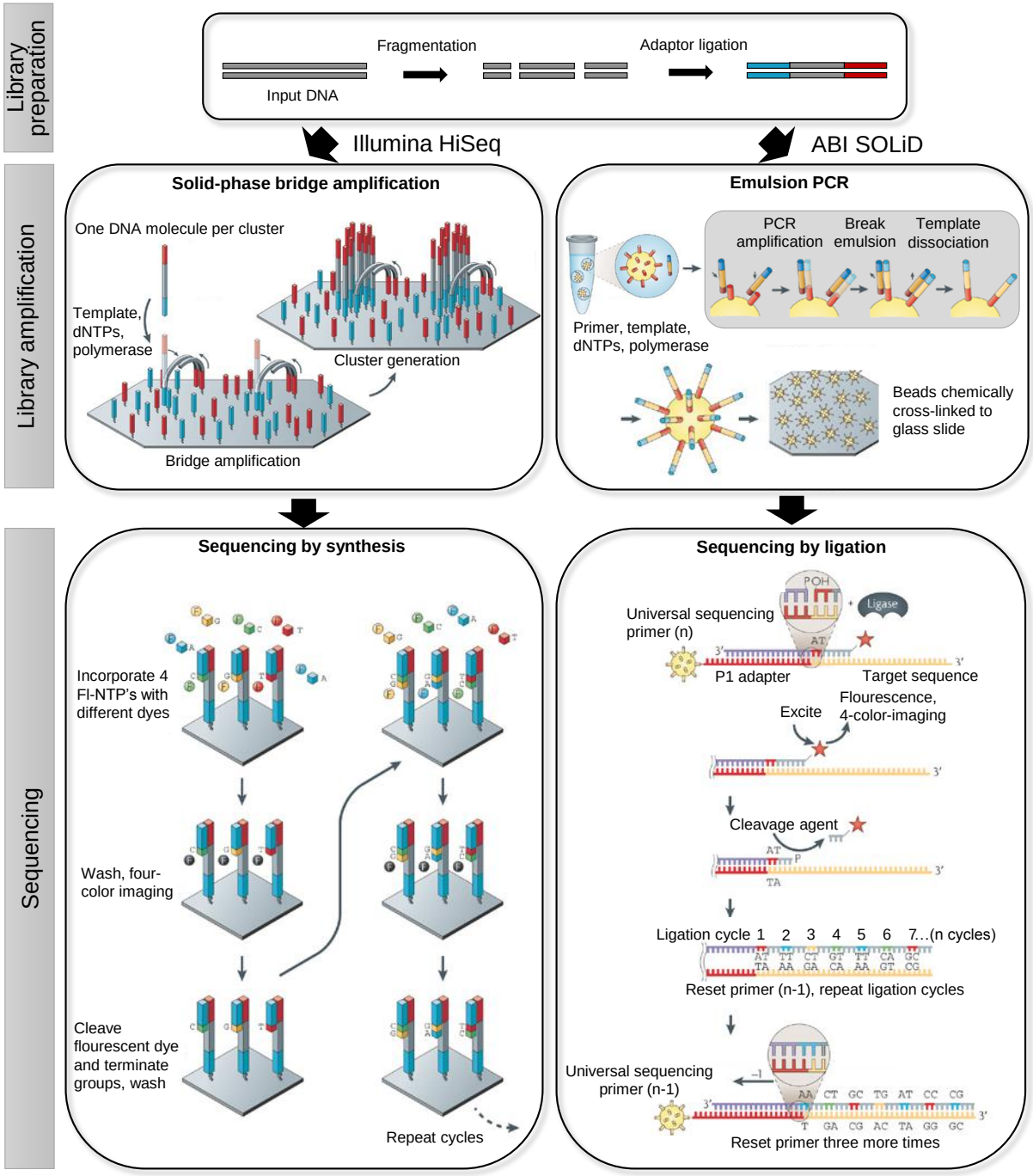


Figure 1-3 Comparison between the sequencing technologies ABI SOLiD and Illumina HiSeq
The figure was modified after Metzker *et al.* [86]. First, a sequencing library has to be generated in both technologies. Afterwards, a library amplification is performed followed by the actual sequencing. The left part shows the sequencing workflow for the Illumina HiSeq: The library amplification is based on a bridge amplification on a solid surface followed by sequencing by synthesis. In contrast, the sequencing procedure for ABI SOLiD shown on the right side is based on an emulsion PCR for library amplification, while the sequencing is performed by ligation.

Taken together, NGS is a widespread used high-throughput method to investigate the genetic background of different species. While per base prices dropped, robustness and output increased in the last years. Three of the main applications of NGS are transcriptome, whole exome, and whole genome sequencing.

1.5.1 Transcriptome sequencing

The transcriptome represents a set of RNA molecules in a cell or tissue at a specific point in time and includes messenger RNA (mRNA), ribosomal RNA (rRNA), microRNA (miRNA) and other non-coding RNA types [87]. RNA-Seq is a widespread approach to investigate the transcriptome. In particular, RNA-Seq is applied to investigate the molecular background of cancer and other complex diseases [16, 87, 88].

Less than 5% of the human DNA encodes for annotated transcripts. However, hints exist that more than 90% of the genome is transcribed, whereby it is still discussed to which extend these fragments are only transcriptomic noise or functional RNAs [89]. In contrast to the microarray technology, RNA-Seq enables the detection and investigation of novel transcriptionally active regions (nTars). This is especially important in tumor tissues, which are characterized by altered gene expression patterns including changes in non-coding RNA abundance levels [16, 64, 87]. Transcripts with high expression in diseased tissue but very low or no RNA concentration in healthy individuals might be missed with the microarray approach, because probes for microarrays are mostly based on transcriptomes from healthy individuals. Despite this advantage of NGS, no publicly available tool existed to detect and comprehensively characterize nTars at the time of my studies. To fill this gap, a novel program was developed for this purpose as part of my work. Besides the investigation of species with an incomplete annotation, this will enable a comprehensive study of phenotypes such as inflammation-triggered colorectal cancer.

Taken together, the transcriptome describes the current state of a cell and can be investigated with RNA-Seq, which enables the investigation of annotated transcripts as well as nTars. In addition, studies of the genome, especially the exome, are important to reveal the potential of a cell.

1.5.2 Whole exome sequencing

In WES, DNA sections covering exonic regions are enriched for sequencing. Thus, the focus of WES is, like for RNA-Seq, on sequencing of exonic regions. However, for the following reasons only WES is suitable for variant calling: (i) Due to expression differences in RNA-Seq, a uniform coverage distribution across all exonic regions is missing [90]. As consequence, it is not possible to detect variants in unexpressed genes [91]. In genes with low RNA level, variants can only be detected, if very deep sequencing is applied. This in turn would have a huge impact on the costs. (ii) In most cells, expression of two alleles with different mRNA levels occurs in ~20% of the transcripts. Allelic imbalance exists especially often in cancer cells, in which genes associated with cancer risk and tumor development are affected by different RNA levels of the alleles [92, 93]. This complicates the variant calling, the filtering process, and the determination of sequencing errors and thus influences the detection and classification of biological variants.

(iii) Alternative splicing variants exist for more than 90% of the genes [91]. Mismapping of reads occurs especially often at splice-junctions, which makes it difficult to distinguish between real existing variants and mistakes in the alignment process. (iv) With RNA-Seq it is hard to distinguish between genomic variants and RNA editing [94]. Because of the mentioned reasons, transcriptome sequencing is not appropriate for variant calling. Moreover, in contrast to transcriptome sequencing, probes of most WES labor protocols cover not only exons, but enable also the investigation of selected functional elements, such as promoter or some regulatory regions [95].

It is estimated that 85% of genomic variants underlying diseases can be found in exonic or functional relevant regions, although these areas account for only 1% of the genome [96]. Therefore, WES can reveal most of the somatic variants of potential interest, but is much more efficient regarding costs, time, and resources than WGS would be [16, 97, 98]. This has the advantage that the most important regions can be investigated for more samples with the same coverage depth and same costs [90].

In my studies, WES was performed in addition to transcriptome sequencing to analyze the molecular background of murine inflammation-associated colorectal cancer. Moreover, WES was applied to increase the coverage in the most important regions for the investigation of GC. In both projects, a target enrichment approach based on in-solution hybridization was applied. In short, probes (biotinylated RNA baits) with a sequence complementary to the regions of interest are provided in a liquid medium. These probes can bind to fragments from a whole-genomic DNA library. The probes carry a biotin to enable the isolation of the DNA fragments of interest with streptavidin-coated magnetic beads [95, 99]. Afterwards, the enriched DNA library fragments can be purified. Advantages of this enrichment approach compared to other techniques include (i) a uniform distributed sequence coverage, (ii) easy application, and (iii) a lack of a GC bias due to avoidance of PCR amplification [90, 98].

Taken together, WES enables to investigate the genomic sequence of exonic regions and functional elements with a cost-efficient approach. However, in contrast to WGS, WES does not cover all intergenic regions and depends on the knowledge of the underlying sequence.

1.5.3 Whole genome sequencing

WGS is the most comprehensive approach to investigate genomes and a widely used NGS method [84]. WGS is not limited to specific regions and not biased to any previous knowledge. Moreover, WGS enables the detection of structural variants, such as large insertions, large deletions, and translocations [97].

Till date, over 14,000 genomes were published in the US National Center for Biotechnology Information (NCBI) genome database and new genomes are submitted regularly [84]. These

huge data sets decipher information about genomic variance between different individuals and populations and enable the investigation of the relationship between genomic alterations and phenotypes [83]. These results can serve as reference set for clinical applications. Also in cancer research, WGS has been applied to identify cancer causing alterations on genome level, including SNVs, genomic rearrangements, CNVs, and tandem duplications [3–5].

To summarize, WGS is the most comprehensive sequencing approach and allows the detection of all DNA sequence alterations in exonic and intergenic regions.

1.6 Variant types

Genetic variation describes all genomic differences between individuals of a certain species. This includes an aberrant number of chromosomes as well as alterations, such as CNVs, InDels, tandem duplications, inversions, translocations, and SNVs [100]. Common base substitutions having a frequency above 1% within the entire population are defined as single nucleotide polymorphisms (SNPs) [101, 102]. Besides modifications of the protein structure, genomic alterations can influence the expression level of a protein or can be without obvious effect [103]. Based on a comparison between humans and chimpanzees, an average neutral alteration rate of 2.3×10^{-8} variants per nucleotide position per generation was suggested for humans [100]. This is similar to ~40-80 mutations per generation identified by the investigation of Icelandic parents and their offspring [104]. Around 10% of all SNVs have an effect on a protein function resulting in on average six new deleterious mutations in a newborn baby [104]. The human base substitution rate is roughly twice as high as the one in wild-type C57BL/6 laboratory mice, which is predicted to be around 5.4×10^{-9} alterations per nucleotide per generation [105].

Mutations can be a result of the insertion of a wrong nucleotide during DNA replication [103]. Thereby, the mutation rate rises with increasing DNA replication time. Because of a transcription-coupled repair mechanism, the mutation rate is additionally negatively correlated with the gene expression level [106]. At non-CpG sites, transitions defined as base substitution between purines or pyrimidines (purine (A/G) ↔ purine (A/G) or pyrimidine (C/T) ↔ pyrimidine (C/T)) appear nearly twice as often as transversions defined as base substitution between one purine and one pyrimidine (purine (A/G) ↔ pyrimidine (C/T)), although only two transition types but four transversion types exist [107, 108]. The mutation rate of transitions is 1.6×10^{-7} at CpG sites and 1.2×10^{-8} at non CpG sites, while transversions occur with a frequency of 4.4×10^{-8} at CpG sites and 5.5×10^{-9} at non-CpG sites [107]. One reason for the higher number of transitions than transversions is that transitions lead more likely to a synonymous SNV without amino acid change, which has in most cases no effect on the function of a gene and is therefore tolerated [108, 109]. Moreover, SNVs often occur spontaneously due to deamination of methylated cytosine to uracil leading to a C:G to T:A base pair exchange. This explains also the ten times

higher transition rate at CpG sites than the average mutation rate in mammalians, because methylated cytosines occurs more often at CpG sites than at non-CpG sites [103, 107, 110]. Besides spontaneous DNA changes, each tumor bears a specific variant pattern, which depends on the cancer type, the age of the patient at disease onset and dysfunctions in DNA repair mechanisms [5]. Moreover, alterations can be induced by lifestyle factors and environmental conditions, such as sunlight (ultraviolet light), cigarette smoking, and exposure to chemical carcinogens [9, 103]. Thereby, each mutagen leaves a specific mutation pattern as a footstep. In general, while chromosomal abnormalities are aggregated by ionizing radiation, point mutations are generated by chemical mutagens (Box 2) [103, 111]. The variant pattern of each cancer type can consist of several signatures and reflects the underlying mutational processes and, in certain cases, contributing mutagens [5]. Even if the origin of some mutational patterns is still unknown, analyses of variant distributions are important, because the results can be used as biomarker for specific phenotypes, including clinical features [84, 112].

In addition to the SNV type distribution, the SNV sequence context is important. These two features together lead to a clustering of specific cancer types and known carcinogenesis mechanisms [106]. Colorectal cancer types, which feature a microsatellite stable or a low level microsatellite instable status, are characterized by a high amount of T(C>A)T and T(C>T)G base substitutions. In contrast, colorectal tumor samples with a high level of MSI or mutant POLE, which is involved in DNA repair, harbor mainly C(C>A)N and G(C>T)N variants [113]. However, Greenman et al. [54] observed an increase of C>T variants occurring mainly at CpG sites in colorectal cancer. In contrast, an increase of T>G, C>T, C>A variants and InDels was reported for GC [54].

Like in CRC, genomic and epigenetic features play an important role in CAC. Additionally, inflammatory cells cause oxidative stress by producing reactive oxygen, nitrogen species, and mediators such as cytokines and metalloproteinases [30, 69]. This in turn influences the formation of e.g. C>A base substitutions [69, 114]. Moreover, inflammatory mediators can contribute to an increase of random genetic alterations due to downregulation of DNA repair pathways, induction of double-strand breaks and defective cell cycle checkpoints as well as altered activation of homologous recombination [30].

In summary, genomic variants can either occur spontaneously, are caused by environmental factors or triggered by altered intracellular processes. Thereby, each factor leaves a specific variant pattern.

Box 2 Mutational patterns associated with different mutagens and cancer types

The following examples demonstrate how the connection between cancer type and mutational patterns can provide insights into causing environmental factors and underlying molecular processes [9, 110, 115]:

- Non-melanoma skin cancer harbors a high number of C>T and CC>TT base substitutions at dipyrimidine sites. This mutation type is caused by ultraviolet light, which is one of the main risk factors for skin cancer [9, 110, 115].
- Lung cancer harbors a high ratio of G>T transversions resulting from the formation of bulky adducts on guanines caused by tobacco [9, 110, 115].
- In many urothelial carcinoma, the number of T>A transversions in a CTG context is increased. This pattern is e.g. caused by aristolochic acid, which is used in some products in the traditional Chinese medicine [110, 115].
- In some hepatocellular cancer types, an increase of C>A transversions was observed. These might arise by consumption of aflatoxin, which is often contained in food from southern Africa and Asia [110].
- Some liver cancer types bear an increased amount of SNVs at A:T positions caused by vinyl chlorides, which are important for the production of plastic materials and a risk factor for liver cancer [115].
- In lung and skin cancer, alterations are predominantly observed on the non-transcribed strand and in lowly expressed genes. This demonstrates that, besides the mutagen, a defect transcription-coupled nucleotide excision repair might foster the observed SNV pattern [9, 110, 115].
- Nearly all cancer types feature a high number of C>T variants at CpG sites [9, 110, 115]. This mutation type is also the most frequent alteration during evolutionary processes and caused by deamination of 5’meC [9, 110, 115].

1.7 Project aim

A better understanding of molecular mechanisms underlying a specific cancer type is essential to devise novel therapies and to improve the patients' outcome. Therefore, my studies aimed to get deeper insights into genomic factors involved in the development of two gastrointestinal cancer types, namely human GC and murine AOM/DSS-triggered colorectal cancer, which is a mouse model for human CAC.

GC causes 10% of all cancer deaths and is among the most common malignancies worldwide [1]. Nevertheless, at the point in time of the project start, the knowledge of underlying genomic signatures, including structural variants, was limited. Thus, the molecular basis of a microsatellite stable as well as a microsatellite unstable gastric carcinoma was investigated. WGS and WES were performed to enable a comprehensive characterization of genes and processes involved in the development of GC. Besides revealing novel insights into the molecular background of GC, this project was intended to develop a bioinformatic analysis pipeline, which is suitable for cancer projects based on a low sample number.

Human CAC is a serious complication of IBD. Although, compared to sporadic CRC, patients affected by CAC are younger and prognoses are worse in the advanced stage [42, 44, 45], far less is known about the molecular signatures of CAC than those of sporadic CRC. Understanding the genomic and transcriptomic background of CAC would be the key for the development of a suitable therapy. Therefore, CAC is often investigated with the murine AOM/DSS model, although underlying molecular alterations of this mouse model were not systematically analyzed yet. However, knowledge about modified genes and processes in the mouse model is of crucial importance to correctly interpret results and to draw conclusions for human CAC in a proper way. To fill this gap and to identify advantages and limitations of this mouse model for the investigation of human CAC, the molecular patterns from early and late stage murine inflammation-associated colorectal tumor samples were compared with formerly inflamed colonic tissues from mice treated with DSS or AOM/DSS as well as with healthy colonic tissue samples from untreated mice. Whole transcriptome sequencing and WES were combined in an integrative approach to reveal tumor-specific variant patterns as well as genes and processes potentially associated with AOM/DSS-induced cancer.

Both investigated cancer types are based on multistep mutation processes, but they differ in affected genes, variant patterns, and onset of some shared alterations [47, 73, 74]. This demonstrates the importance of a cancer-specific therapy, which can only be developed with a deep knowledge of underlying processes. Therefore, the aim of the thesis was a comprehensive characterization of these two cancer types and to suggest novel analysis pipelines for the investigation of NGS cancer data. A scheme of the workflow is shown in Figure 1-4.

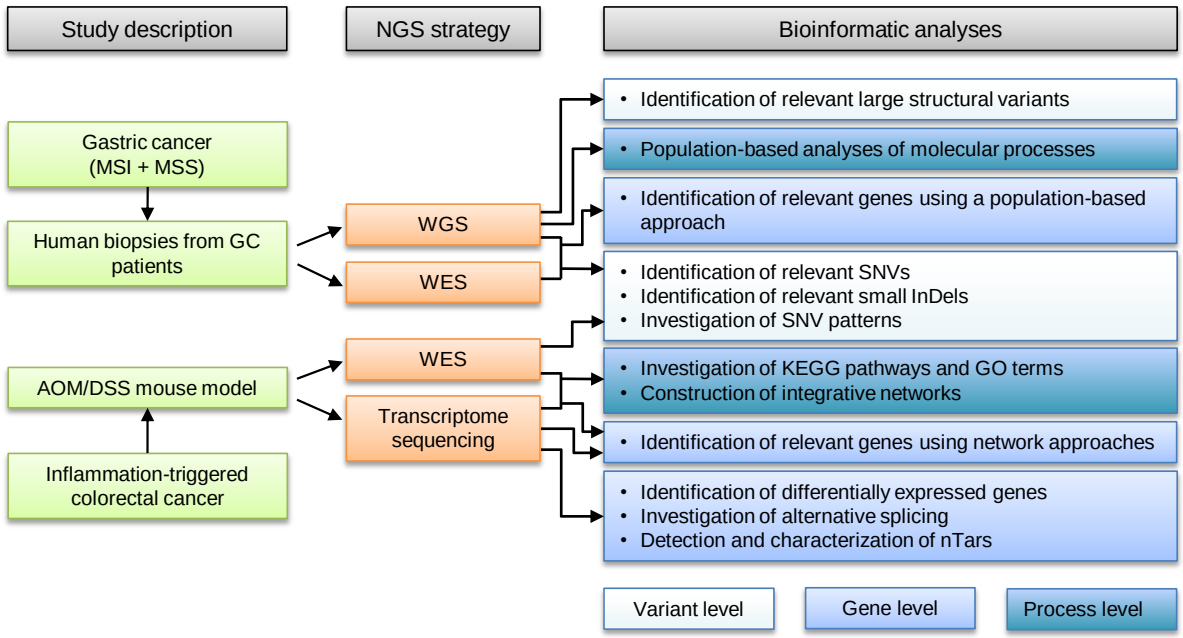


Figure 1-4 Workflow of my studies
The molecular background of two different cancer types, namely human GC and murine inflammation-associated colorectal cancer, were analyzed in the course of my studies. Using WGS, WES, and transcriptome sequencing, alterations on variant, gene, and process level were investigated.

2 Material and Methods

All chemicals, solutions, kits, and devices are described in the supplementary chapter 6.1.1.

2.1 Sample preparation

2.1.1 Sample collection from human patients with GC

Tissue samples from two female patients (74-years old Caucasian women) who had died from an intestinal GC type were retrieved from the archive of the Institute of Pathology at the University Hospital Schleswig-Holstein in Kiel. Samples from the primary tumor and the corresponding non-neoplastic gastric mucosa had been collected immediately after surgery, were fresh frozen and stored at -80 °C until use. The microsatellite status was determined as described in the supplementary chapter 6.1.2. The tumors were classified as shown in the supplementary section 6.1.3.

The patients have given written informed consent to prospective tissue sampling of excess tissue material, which was no longer needed for diagnostic or therapeutic purposes. The project was approved by the local ethics committee of the University Hospital in Kiel, Germany (reference numbers AZ 140/99 and D 453/10). All patient data were pseudonymized before study inclusion.

2.1.2 AOM/DSS colitis

The mouse experiments were performed with 40 10-13 weeks old, male C57BL/6N wild type mice from Charles River Laboratories. A well-established inflammation-associated colorectal carcinogenesis model in mice was applied [53], in which the animals were treated with a combination of the inflammatory agent dextran sulfate sodium (DSS, Sigma Aldrich, Munich, Germany) and the genotoxic colonic carcinogen azoxymethane (AOM, Sigma Aldrich, Munich, Germany). To compare the results from early and late stage cancer, three experimental settings with different DSS concentrations and treatment durations were applied (Figure 2-1). In all three sets, 10 µg AOM dissolved in 10 µl 0.9% phosphate buffered saline (PBS) per gram body weight was intraperitoneally injected (AOM/DSS group). One week later, drinking water with 2% DSS was administrated for five to seven days followed by 9-14 days without any treatment. The second cycle started again with an AOM injection using the same concentration as applied for the first administration and was followed by four to five days with oral exposure of 1-2% DSS after one week. In order to see differences between chronic inflammation and inflammation-associated colorectal cancer, a group of animals (DSS group) was treated with DSS like the AOM/DSS group. However, the injection was performed with pure PBS instead of AOM. The control mice received an intraperitoneal application with pure PBS and normal drinking water without DSS.

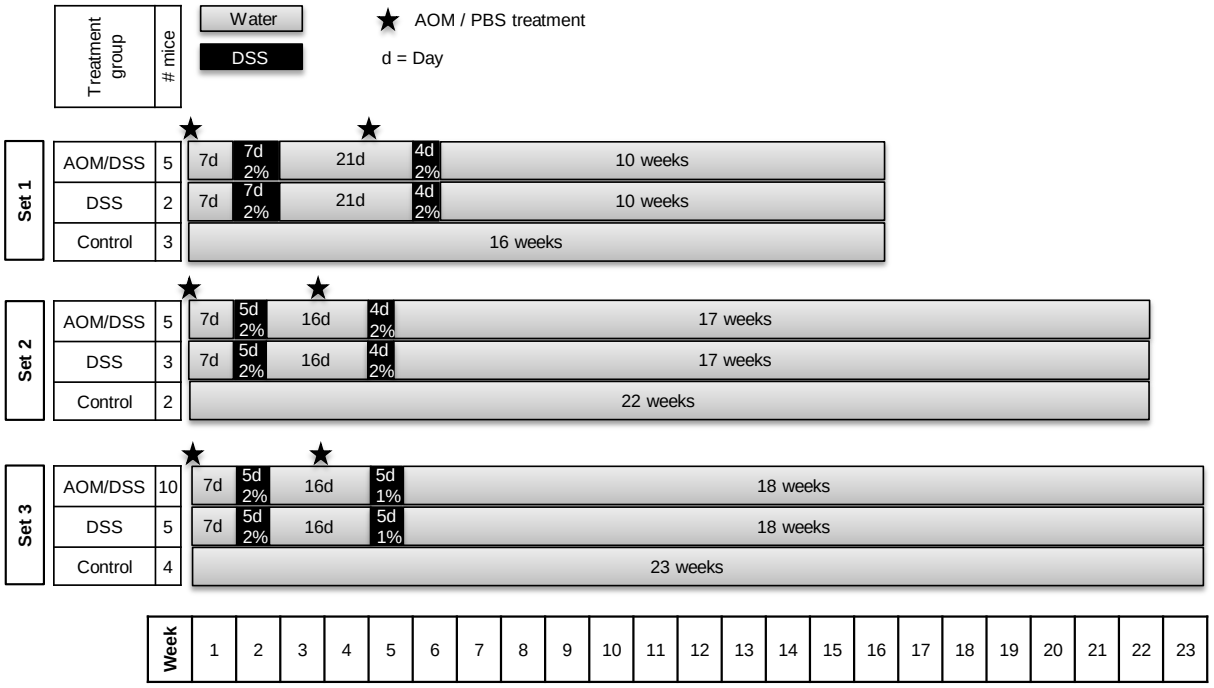


Figure 2-1 AOM/DSS colitis model
The figure shows, at which days AOM or PBS was injected and when DSS was applied with the drinking water. The first set of mice was treated with high DSS dose, the second with a medium DSS dose, and the third set with a low DSS dose. No differences existed between the concentrations of the AOM injections.

Mice with weight loss above 20% were excluded from the experiment. The disease activity index (DAI) was calculated as a sum of the scores for weight loss, rectal bleeding, and stool consistency (Table 2-1). At the end of the experiment a colonoscopy was performed under anesthesia.

Score	Weight loss	Rectal bleeding	Stool consistency
0	0%	Hemoccult test negative	Formed
1	1-5%		
2	6-10%	Hemoccult test positive	Smeary, unformed
3	4-20%		
4	>20%	Rectal bleeding	Liquid

Table 2-1 Scores for the calculation of the disease activity index

After organ removal at the end of the experiment, the colon was rinsed with PBS, opened longitudinally, and cut vertically. The first part of the colon was divided into four pieces. All sections as well as samples from the tumor tissues were snap frozen in liquid nitrogen and stored at -80°C until further use. The subsequent examinations were performed with the most proximal part from the colon of all animals as well as with tissue from up to two tumors per mouse. The most proximal section was chosen, because in none of the mice, it was affected by a tumor and therefore comparable between all treatment groups.

From the second part of the colon, ‘Swiss Rolls’ [116] were prepared, placed in cassettes, and fixed in 4% paraformaldehyde buffered formalin (Sigma Aldrich, Munich, Germany) at 4°C for more than 24 hours followed by dehydration and embedding in paraffin wax. Tissues were cut

into ~3.5 µm sections using the Leica RM 2255 (Leica Microsystems, Wetzlar, Germany) microtome and mounted on microscope slides.

The mouse experiments were performed in collaboration with Dr. Maren Falk-Paulsen. The colonoscopy was kindly carried out by Dr. Konrad Aden.

2.1.3 Hematoxylin and Eosin staining

The samples were deparaffinized and rehydrated using a decreasing xylene (Sigma Aldrich, Munich, Germany) and ethanol (Merck, Darmstadt, Germany) dilution series. After washing steps with distilled water, the slides were stained with hemalum (Carl Roth, Karlsruhe, Germany) for 2-5 minutes followed by exposing to running tap water for blueing of the nuclei (13 min). After counterstaining of the cytoplasm with a 1% eosin solution for two minutes, slides were washed for 1-2 seconds with distilled water and dried using an increasing alcoholic series. The slides were coverslipped with Roti Histokit (Carl Roth, Karlsruhe, Germany) and analyzed with a transmitted light microscope Axio Imager Z1 (ZEISS, Oberkochen, Germany).

The hematoxylin and eosin staining (HE staining) was performed by Maren Reffelmann and analyzed in collaboration with Dr. Maren Falk-Paulsen.

2.1.4 DNA and RNA extraction

After tissue homogenization, the extraction of genomic DNA from human samples was performed with the QIAamp DNA mini kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. The extraction of genomic DNA from the murine samples was conducted with the DNeasy® Blood & Tissue kit (Qiagen, Hilden, Germany) following manufacturer's recommendations.

For RNA extraction, 1 ml TRIzol (Invitrogen / Life Technologies, Darmstadt, Germany) was added to 50-75 mg pestled tissue followed by vortexing, five minutes incubation at room temperature, and addition of 200 µl chloroform. After mixing, further incubation at room temperature for 2-3 minutes, and centrifugation (12,000 g) at 4°C for five minutes, the clear supernatant was mixed with 500 µl isopropanol (Merck, Darmstadt, Germany) followed by incubation at room temperature for ten minutes. After further centrifugation (12,000 g) at 4°C for ten minutes, the supernatant was discarded and the pellet inserted into 1 ml cold 75% ethanol followed by vortexing and centrifugation (7,500 g, 4°C, 5 minutes). In the final step, the pellet was dried and dissolved in RNase-free water.

2.1.5 Library preparation

Library preparations for human WES were performed with the Sure Select Target Enrichment Human All Exon v2 kit (Agilent, Santa Clara, USA), while libraries for the human WGS were created with the TruSeq DNA sample prep kit (Illumina, SanDiego, USA).

The murine DNA samples were used to create sequencing libraries with the Illumina TruSeq DNA sample prep kit (Illumina, SanDiego, USA). After pooling four samples together, the SureSelectXT Mouse All Exon kit (Agilent, Santa Clara, USA) was applied for the enrichment of the murine whole exomes.

For the preparation of the murine transcriptome libraries, rRNA was removed with the Ribo-Zero™ Human/Mouse/Rat rRNA Removal Kit (Illumina, SanDiego, USA) followed by the library preparation with the TruSeq Stranded mRNA Sample Prep Kit (Illumina, SanDiego, USA).

Library preparations were kindly performed by the NGS laboratory of the Institute of Clinical Molecular Biology in Kiel.

2.1.6 Sequencing

Cluster generation for human WES was performed with an emulsion polymerase chain reaction (ePCR) using the SOLiD PCR Kit (Applied Biosystems / Life Technologies, Darmstadt, Germany). Afterwards sequencing was conducted with a paired 50/35 run on the SOLiD 4 System (Applied Biosystems / Life Technologies, Darmstadt, Germany). Each sample ran on a quarter slide, respectively.

Cluster generations for human WGS, murine WES, and murine RNA-Seq were performed on the cBot (Illumina, SanDiego, USA) with the TruSeq PE Cluster Kit (Illumina, SanDiego, USA). Thereby, the kit version v2.5 was applied for the human MSI tumor and non-tumor sample from the first patient as well as for the MSS tumor sample from the second patient. The kit version v3 was used for the control sample from the second patient and all murine WES and transcriptome samples. After cluster generation, the sequencing was performed with the TruSeq SBS Kit (200 cycles, paired end modus, Illumina, SanDiego, USA) on the Illumina HiSeq 2500 (Illumina, SanDiego, USA). Each human WGS sample was placed on a complete flow cell, while each pool containing four murine samples were sequenced on one lane.

The sequencing was kindly performed by the NGS laboratory of the Institute of Clinical Molecular Biology in Kiel.

SNVs in the genes *DROSHA*, *MSH4*, *RERE*, *ROS1*, *TACC2*, and *TYRO3* were additionally verified by pyrosequencing on a Pyromark Q24 device (Qiagen, Hilden, Germany). Primers for amplification of the respective gene regions and the corresponding sequencing primers are shown in Table S 5. The validations were kindly performed by the research group of Prof. Dr. Christoph Röcken at the Institute of Pathology in Kiel.

2.2 Bioinformatic analyses

2.2.1 Databases and references

The following databases were used to include known information about genes and variants in the analyses:

The reference sequences hg19 and mm10 were received from the University of California, Santa Cruz (UCSC) Genome Browser Database [102]. For human samples, the variant annotation was based on all features described in the Reference Sequence database (RefSeq) provided by the NCBI (downloaded from http://hgdownload.cse.ucsc.edu in August 2011) [117]. In the murine samples, variant annotation and expression analyses were performed with all transcripts described in the UCSC database [102]. All transcripts annotated in RefSeq, which did not overlap with any UCSC gene, were added to the murine reference. The UCSC rRNA and tRNA positions were used for the calculation of the region distribution and the sense/antisense noise. All murine annotation references were downloaded from the UCSC homepage (http://genome.ucsc.edu/) on 05/08/2013.

All protein-protein interaction networks were based on connections annotated in the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database [118]. The interactions included the following eight different sources: text mining (occurrence of both protein names in the same abstract of a publication), co-expression (similar mRNA expression patterns), co-occurrence (similar absence/presence of the genes in the phylogenetic profile), database (relation derived from the databases Biocarta, BioCyc, GO, KEGG, and Reactome), homology (high degree of sequence homology), experiments (interaction based on experimental data like affinity chromatography, information is extracted from the databases BIND, DIP, GRID, HPRD, IntAct, MINT, and PID), gene neighborhood (adjacent genomic position), and gene fusion (known gene fusion). All types of information were used to calculate a combined confidence score for each connection. The STRING protein-protein interaction networks created in this thesis were based on confidence scores larger 0.4, which is equivalent to medium reliance.

All pathway analyses were based on reaction networks annotated in the database Kyoto Encyclopedia of Genes and Genomes (KEGG) [119]. In addition, gene set enrichment analyses were based on functional groups annotated in the Gene Ontology (GO) [120] database. Exclusively the domains ‘Molecular Functions’ and ‘Biological Process’ were considered. The GO database is structured as a directed acyclic graph, in which the root is the most general and the leaves are the most specific terms. Enrichment analyses were performed on all terms, but only the highest significant branch levels referring to the most general terms were completely reported in all analyses.

SNVs and InDels, which were also observed in healthy individuals, were identified and annotated using the 'Single Nucleotide Polymorphism Database' (dbSNP) [121] hosted by the NCBI in collaboration with the National Human Genome Research Institute.

Different variant sets were compared with the following databases to identify disease-related alterations: (i) The COSMIC (Catalogue Of Somatic Mutations In Cancer) [122] database contains information about cancer samples including mutations, tumor type, tissue, and patient. Moreover, a prediction about the functional consequence and the cancer driver potential calculated with FATHMM version 2.3 (Functional Analysis through Hidden Markov Models) exists for each annotated variant [123]. Genes with a cancer driver mutation in at least two samples described in the COSMIC database were defined as cancer-associated in my studies. Additionally, the COSMIC cancer gene census list were used as reference for known cancer-associated genes. (ii) The applied variant database is HGMD (Human Gene Mutation Database) [124] maintained by the Institute of Medical Genetics in Cardiff represents published gene lesions responsible for human inherited diseases. (iii) The used ClinVar database [125] contains reports of variants found in patient samples including assessment of their clinical significance and information about the submitter. (iv) OMIM [Online Mendelian Inheritance in Man, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), http://omim.org] is a catalog of human genes and genetic disorders with particular focus on the molecular relationship between genetic variation and phenotypic expression. It is updated by the McKusick-Nathans Institute of Genetic Medicine of the Johns Hopkins University School of Medicine. (v) The used GWAS (genome-wide association study) [126] catalog of the National Human Genome Research Institute includes primary GWAS analysis, which are defined as array-based genotyping and analysis of at least 100,000 SNPs to highlight variation across the genome.

Besides the usage of the mentioned databases, the cancer proliferation index (cPI) [127] was used to filter for genes potentially associated with CAC or GC. A negative cPI value indicates that the gene might be a tumor suppressor, while a positive cPI value points to potential oncogenes. Genes having a neutral cPI value are very likely not associated with the development of cancer.

2.2.2 Statistical tests

Pairwise comparisons were either performed with the Wilcoxon Rank-Sum test [128] for comparisons of independent sample groups, such as samples from different mice or with the Wilcoxon Signed-Rank test [128] for comparisons of matching sample groups, such as samples received from the same mice (e.g. tumor vs. proximal samples from AOM/DSS treated mice). Statistical significance of multivariate approaches were proven with the Kruskal-Wallis test [129]. In all tests, the following parameter were applied for the calculation of the p-values:

In case of a known hypothesis about the relationship of the sample groups, a one-tailed test was applied. Without any previous information, a deviation in both directions was considered as possible and a two-tailed approach was used. An unpaired test was performed for all comparisons between different sample types, e.g. between tumors and controls. In contrast, tumor and proximal tissue from the same AOM/DSS-treated mice were dependent and a paired test was applied to assess significance.

The Fisher-Exact test [130] was used for all analyses of nominal variables, while comparisons between the distributions of two sets were performed with a binomial test. A linear model was used to test for significant differences between states described by several explanatory variables: First, the model was applied to test for a significant dependency between SNV count and gene expression (chapter 2.2.7.4). Second, the method was used to quantify the influence of different factors, including animal treatment and technical bias, on the mRNA levels (chapter 2.2.5). For both analyses, the function 'lm' in R [131] was applied.

For all investigations with more than one test, p-values were corrected with the Benjamini-Hochberg method [132] to control for multiple testing. In all analyses a p-value < 0.05 was considered as significant (*), a p-value < 0.01 as strongly significant (**) and a p-value < 0.001 as highly significant (***).

2.2.3 Quality filter and mapping

2.2.3.1 Human samples

The Illumina WGS sequences from patients with GC were mapped against the human genome reference build hg19 [133] using BWA version 0.5.9 [134]. The seed length was set to 35 with a maximum of two differences in this subsequence, while for all other parameters the default options were applied. This included a mismatch penalty of three, a gap opening penalty of four and a maximum edit distance of 0.04. The SOLiD WES reads were aligned with Bioscope version 1.2.1 (Applied Biosystems™) using the default parameters. Reads with mapping/pairing quality less than eight or an alignment-length/read-length ratio less than 0.85 were excluded from further analyses.

Read pairs that have identical 5' coordinates and orientations were called duplicates except the read pair with the highest sum of quality scores. Thereby, only quality scores of 15 or larger were considered. Duplicates were marked with the Picard tool MarkDuplicates version 1.41 (downloaded from http://picard.sourceforge.net) and removed from further analyses. The coverage for the WES and WGS was calculated with all mapped reads excluding duplicates.

2.2.3.2 Murine samples

The following quality filter steps were applied for all reads before performing the mappings: First, reads classified as bad quality by the Illumina CASAVA version 1.8.2 pipeline (flag 'Y') were excluded. This step was performed with the script fastq_illumina_filter-Linux-x86_64.bin (downloaded from http://cancan.cshl.edu). Afterwards, all bases with a quality score lower than 12 were trimmed from both ends of the sequences. Next, all reads having eight or more undefined bases ('N'), a mean quality score below 15 or a length smaller 20 were discarded using the tool prinseq-lite version 0.20.3 [135]. Exclusively pairs, in which both reads passed the quality filter, were considered for further analyses.

The quality filtered read pairs were mapped against the mouse genome reference build mm10 including random chromosomes (unlocalized and unplaced contigs). Unplaced sequences are defined as contigs with unknown position in the assembly. Unlocalized contigs are associated with a specific chromosome but cannot confidently be placed on the chromosome. The mappings were performed with BWA version 0.6.2 for the WES data. The seed length was set to 35 with maximum two differences in this subsequence. For all other parameters, the default options were applied. Duplicates were marked with the Picard tool MarkDuplicates version 1.41 and excluded from further analyses of the WES data. The transcriptome data were aligned with TopHat version 2.0.8b using the sensitive option [136].

2.2.4 Quality control of transcriptome data

For each sample, the numbers of k-mers with a length of nine bases were counted for all mapped reads with the program Jellyfish version 2.1.3 [137]. Before each pairwise comparison, the k-mer counts were scaled to correct for differences in the sequencing depth. In detail, the ratio of the total amount of k-mers between two samples was used to scale all values to the smaller profile. The distance between two samples was calculated with the function $f(x,y) = |x-y| / ((x+1)*(y+1))$. The median of all pairwise comparison of one sample was defined as K-dist.

The D-dist was based on the expression values of all transcripts and calculated as the median of all pairwise spearman correlation coefficients between one and all others samples. The outlier threshold for the K-dist and the D-dist was set to 1.5 times of the interquartile range [(Q1 - 1.5 x (Q3 - Q2)) and (Q3 + 1.5 x (Q3 – Q2))].

A test for a systematic 3' or 5' bias of the transcripts was performed with the script CollectRNASeqMetrics, which is part of the program Picard (http://picard.sourceforge.net). The same function provided also information about the regional distribution of the reads and the percentage of sense/antisense noise.

2.2.5 Factors influencing the variation in mRNA expression and SNV formation

The contribution of different factors like treatment or technical bias to the variation in transcriptome and exome data was calculated as described by van 't Hoen et al. [138]. The analysis was performed for the following features: expression level per gene, SNVs, genes affected by at least one SNV, and number of SNVs per gene. In short, for each feature (e.g. expression level), a standard linear model was fitted using the R function 'lm' followed by an ANOVA [139] analysis. The percentage explained by each factor such as treatment, library batch, and sequencing run was calculated from the resulting ANOVA tables by dividing the sum of squares by the total sum of squares. All comparisons with warning message indicating an unreliable F-test were excluded. Genes expressed in less than three samples were not considered for the investigation of factors influencing the mRNA level. In the WES data, the analysis was exclusively based on SNVs or affected genes, which existed in at least three samples. The distribution of the percentage of variance explained by different factors across transcripts was visualized with boxplots.

2.2.6 Test for bacterial or viral infection

To test for bacterial or viral infection, from each tumor sample over 20,000 unmapped reads with more than 95 bases having a quality score larger 35 were blasted against the NCBI nucleotide collection database ‘nt’ [140]. Furthermore, all reads with a quality score larger than 20 at 80 or more positions were assembled with Velvet version 2.06 [141] and the resulting contigs again blasted against the NCBI database ‘nt’.

2.2.7 Investigation of variants

2.2.7.1 Calling and annotation of SNVs and small InDels in human samples

SNV calling was performed with GATK version 1.3 [142], diBayes (Bioscope version 1.2.1), and Samtools version 0.1.16 [143] in the WES data and with Samtools as well as GATK in the WGS mappings. Small InDels were detected with diBayes (Bioscope version 1.2.1) in the WES and Samtools version 0.1.16 in the WGS mappings. The applied parameters are described in the supplementary chapter 6.1.5. All variants called on chromosome Y were excluded, because exclusively samples from female patients were investigated. Thus, reads mapping to chromosome Y were most likely caused by homolog regions.

Annovar (version Jun 2011) [101] was applied for the annotation of SNVs and small InDels. The genomic region types are described in the supplementary Table S 6. A base substitution was classified as damaging, if either PolyPhen-2 [144] or SIFT [145] predicted an effect on the function of the protein (PolyPhen-2 score ≥ 0.43 or SIFT score < 0.05). The SIFT scores were calculated as 1-SIFT. The PhyloP [146] score rescaled by dbNSFP [147] to [0.1] was applied for conservation information of each genomic position. Positions with a value above 0.9995

were defined as highly conserved. SNVs that were not annotated in dbSNP build 132 were classified as novel. The allele frequency information for known SNVs was based on the 2011 May release of the 1000 Genome project data for the CEU population (Utah residents with northern and western European ancestry) [148]. All variants detected by the NHLBI Exome Sequencing Project (ESP) version 2 or by the Exome Aggregation Consortium (ExAc) version 0.3 were excluded from the final reported variant tables. The ExAc_Aggregated_Populations frequencies applied in the GC study were based on sequencing data sets of 60,706 unrelated individuals, which were part of disease-specific or population genetic studies of ExAc. The databases OMIM, HGMD [124] and GWAS [126] were checked for known cancer-associated SNVs.

The allele counts for each position were calculated with the Samtools pileup tool version 0.1.16. Exclusively reads with a minimum mapping quality of 60 and bases with a quality score of at least 13 were considered. A threshold of 5% allele support was set for the existence of an SNV.

The investigation of base substitutions potentially associated with GC were performed with the union of SNVs called with Samtools (WES, WGS) and diBayes (WES), while SNVs called with GATK were only used for pathway analysis. Only SNVs and small InDels supported by five or more percent of the reads in the WES as well as in the WGS data were considered for further filter steps on variant level. An SNV was defined as somatic, if neither in the WES nor in the WGS data of the matching control sample five or more percent of the reads supported the variant. A small InDel was classified as somatic, if in the WES and WGS data of the matching control sample, less than five percent of the reads supported the identical InDel or another InDel at the same or an adjacent position. InDels were also excluded from the final InDel lists described in the supplement, if the exact same InDel type was called in one of the six adjacent base pair positions in the matching control (WES or WGS).

The investigation of SNV patterns described in chapter 2.2.7.4 was based on all SNVs called with Samtools in the WGS data.

2.2.7.2 Calling and filtering of SNVs and small InDels in murine samples

In the murine WES data, SNVs were called with GATK version 2.3.9 [142], Samtools version 0.1.19 [143], and VarScan version 2.3.5 [149], while small InDels were detected with the following programs: (i) Breakdancer version 1.2 [150] followed by Pindel version 0.2.4d [151] to increase sensitivity and specificity of the calls, (ii) Samtools version 0.1.19, and (iii) VarScan version 2.3.5 [149]. The applied parameters for variant calling are described in the supplemental chapter 6.1.7. All SNVs, which were called with at least one tool, were considered for further analyses. To classify the genomic region type (e.g. exonic, UTR) (Table S 6) of each variant, the program Annovar version 2013-08-23 was applied. The dbSNP

build 137 was used to remove variants, which were already found in untreated wildtype mice. PhyloP scores rescaled by dbNSFP to [0.1] were applied to determine the conservation of each position. The allele counts for all filter steps were based on an mpileup file created with Samtools version 0.1.19. Here, the filtering for base and mapping quality was switched off, while all other parameters were set to default.

A base substitution was classified as somatic, if it was not supported in more than one control sample by five or more percent of the reads. Somatic variants were investigated to reveal mutation patterns and SNVs potentially associated with murine inflammation-triggered colorectal cancer. In contrast, an SNV, which was supported by ≥ 5% of the reads in at least two control samples, was defined as germline. These variants were used to investigate differences between the strains C57BL/6N and C57BL/6J, which in turn deciphered SNV patterns formed by evolutionary processes. A different mouse strain was used for the mappings and annotation (C57BL/6J) compared to the mouse experiments (C57BL/6N), because a reference for the mouse strain C57BL/6N is missing so far.

Besides the direct annotation and filter process in the murine genome, a second approach was used to identify SNVs potentially associated with inflammation-triggered colorectal cancer. Here, the positions of all SNVs were converted to the human build hg38 using the liftover tool from UCSC [152]. All base substitutions were annotated with Annovar version 2013-08-23 based on the NCBI RefSeq genome annotation [117]. The thresholds for the damaging prediction of PolyPhen2 and SIFT as well as the PhyloP conservation were based on the classification provided by Annovar. In contrast to variant annotation in the GC project, the SIFT scores were not transformed. As a result, an SNV was classified as deleterious with a pathogenicity larger 0.95. SNVs were defined ‘known’, if they were listed in dbSNP build 138.

2.2.7.3 Definition of the exonic gene conservation score

An additional indicator was defined to identify potentially clinically relevant exonic variants in small human sample sets. Based on 1092 samples from the 1000 Genomes Project, the exonic gene conservation score (ECS) compared the non-synonymous variant rate (germline and somatic) within the exonic regions of the gene of interest with the average non-synonymous variant rate (germline and somatic) across all exonic regions of the genome. To ensure comparability between genes, the score was normalized by the exon lengths. Thus, the smaller the ECS, the higher is the conservation of the exonic regions of the gene and the more likely is an association of alterations in the gene with a disease. Genes with ECS < 0.01 were defined as conserved. The ECS was computed with the following formula:

$$\mathrm{ECS} = \frac{< \#\ \text{variants in gene} > \ \mathrm{x}\ < \text{Total length of exonic regions} >}{< \text{Length of exonic regions of the gene} > \ \mathrm{x} < \text{Avg. total \# exonic variants} > \ \mathrm{x}\ < \#\ \text{samples} >}$$

2.2.7.4 Investigation of SNV patterns

The investigation of the SNV type distribution was based on the six SNV classes C>A (G>T), C>G (G>C), C>T (G>A), T>A (A>T), T>C (A>G), and T>G (A>C). The analyses of the SNV patterns including the flanking bases were performed as described in Nik-Zainal et al. [153]. In short, a genomic heatmap, which was based on the counts of each SNV trinucleotide type corrected for the frequency of the three base sequence in the reference genome, was constructed using the R command 'heatmap.2' [154]. In addition, an Euclidean distance matrix predicated on the normalized trinucleotide incidences was calculated using the R function 'dist' [131] followed by an unsupervised hierarchical clustering with the R method 'hclust' [131].

Besides the SNV type distribution, it was investigated whether the SNV pattern of AOM/DSS-induced colorectal cancer was characterized by a strand bias. First, the number of mutated genes were compared between the Watson and the Crick strand. Thereby, the reference strand (Watson strand / forward strand) was defined as plus strand in the UCSC database. This means that the Watson strand harbored the 5'-end on the short arm of the chromosome, while the Crick strand (minus strand / reverse strand) had the 5'-end on the long arm. In addition, differences between coding (sense) and non-coding (antisense) strand were investigated in intragenic regions. The coding strand referred to the DNA strand whose nucleotide sequence is identical with the sequence of the produced RNA transcript (except the replacement of thymidine by uracil), while the non-coding strand was defined as the transcribed strand. Only exonic SNVs were considered for the investigation of a strand bias. Mutation positions belonging to several genes/transcripts on the same strand were counted just once. If two overlapping genes were on different strands, SNVs in the shared region were counted for both strands.

Principle coordinate analyses (PCoA) were performed with the R command ‘pcoa’ of the R package ‘ape’ (version 5.2) [155] to test for sample clustering. Therefore, the distance matrix based on Jaccard indices was calculated with the method veggdist of the R package ‘vegan’ (version 2.5-3) [131]. A PCoA was performed for each of the following three data sets: (i) exonic SNVs and InDels, (ii) genes affected by at least one SNV or InDel, and (iii) number of mutations per gene. All three categories were based on either all or only non-synonymous variants. To test whether two groups within a PCoA were significantly different, all distances to samples within the same group were compared with all distances to samples from the other group with a Wilcoxon Rank-Sum test (Wilcoxon Signed-Rank test for the comparison between proximal samples from AOM/DSS treated mice and tumor samples from AOM/DSS treated mice).

The occurrence of hypervariable regions was tested by the construction of rainfall plots, in which distances between adjacent substitutions have been visualized with the R function 'plot' [131]. Moreover, a sliding window approach was applied to receive a more detailed description

of the base substitution clusters. In this method, the number of SNVs within a specific nucleotide window were counted. The first investigated region started at the first position of the first chromosome. Then, the window was slid over one nucleotide and the number of SNVs was counted again. If for one region the SNV count was higher than the value defined in the parameters, the region size was consecutively extended to the next SNV until less than the initial number of variants was detected in the range of the initial window size. The window was slid along the complete genome until the end of the last chromosome. Based on the results of previous studies [5, 156], the following start parameters were chosen for the initial tested region: (i) eight somatic SNVs in a 160 bp window to detect small clusters and (ii) 50 somatic SNVs within a 5000 bp window to detect large clusters with low density.

To test for a relationship between gene expression and mutation rate, the genes were divided into six FPKM-classes: unexpressed genes (< 0.01), lowly expressed genes (0.01-1), three classed of moderately expressed genes (1-2 / 2-5 / 5-10), and highly expressed genes (≥ 10). A normalized SNV count was calculated for each class. This was defined as the total length of the coding parts of all genes within the expression class divided by the number of SNVs. Overlapping regions from different isoforms of the same gene were merged.

In the human samples, underlying molecular signatures of the SNV patterns were investigated with the R package SomaticSignatures [157], while the influence of signatures annotated in the COSMIC database (http://cancer.sanger.ac.uk/cosmic/signatures, [122]) to the observed GC mutational patterns were calculated with the R package deconstructSigs [158]. The decomposition was performed with non-negative matrix factorization. In the murine samples, the comparison with known mutational signatures reported in the COSMIC database and the assembly of novel signatures were performed with the R package MutationalPatterns [159].

2.2.7.5 Detection of large structural variants in the GC study

Large deletions, inversions, tandem duplications, and interchromosomal translocations were called with Breakdancer version 1.2 [150]. Sensitivity and specificity of the first three listed structural variant (SV) types were increased by the application of Pindel version 0.2.3 [151]. Two translocations were merged to one SV, if the called positions of the respective breakpoints were smaller than 25,000 bp.

At the time of my analyses, no publicly available tool for the detection of large insertions was able to include mapped as well as unmapped reads. Therefore, a novel pipeline for the detection of large insertions was developed in the course of my studies (Figure 2-2). In the first step, all unmapped reads with a quality score larger than 20 at 80 or more positions were assembled with Velvet version 1.2.06 [141] in the tumor samples. The resulting contigs were aligned with the BLAST-like alignment tool (BLAT) version 34 [160] provided by UCSC Genome Browser Database to the human reference build hg38 to find putative insert sites. If

at least 50 nt of the contig start and 50 nt of the contig end matched to genomic regions within less than five bases distance, the position was considered as a putative insert site. Thereby, the aligned sequences had to be less than 10 nt away from the contig start / end. Insertions smaller than five bases were excluded. Next, all unmapped reads of the corresponding control samples were mapped against all Velvet contigs of the tumor sample. Contigs, which were covered at 90% or more positions, were defined as germline variant and not considered for further analyses. All putative insert sites were annotated with custom script, while BLAST analyses of inserted sequenced were performed.

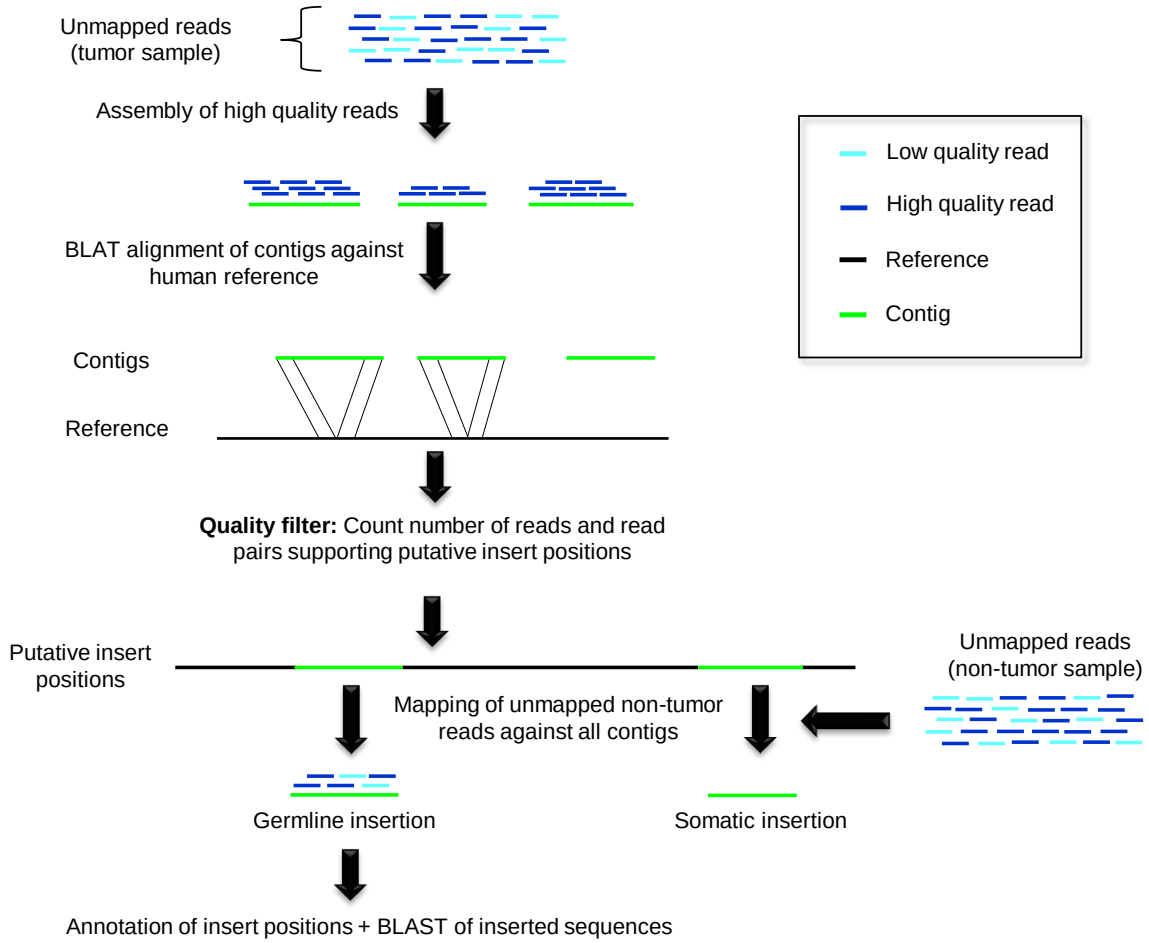


Figure 2-2 Workflow for the detection of large insertions

All structural variants were filtered with the help of split reads and the paired end information of mapped sequences. Moreover, unmapped reads were investigated to enable the identification of split reads, in which the read sequence fitted either to positions on different chromosomes or had a large gap between the aligned positions. Therefore, all high quality unmapped reads were aligned against the human reference hg19 with BLAT version 34. Thereby, high quality reads were defined as sequences with a quality score greater than or equal to 35 at 95 or more positions. All reads, which had one hit with at least 92 mapped bases without gap, were excluded from further analyses. Exclusively reads with exact two alignment parts were considered for further analyses. These two genomic positions were required to

cover together at least 92 bases of the read with a maximum of five bases existing in both read parts. The principle of split reads and read pairs to screen SVs, including the required numbers to pass the quality filter, are described in Figure 2-3. The detailed criteria, which split reads and read pairs had to fulfill in order to support a large SV, are described in the supplementary methods (chapter 6.1.8).

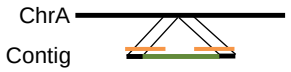
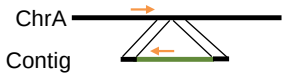
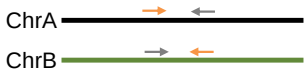
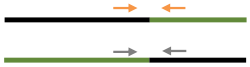
SV type	Feature	Passed filter	Reference	Sample
Insertion	Split reads	25 (BLAT)		
	Read pairs	5%		
Interchromosomal translocation	Split reads	15 (BLAT)		
	Read pairs	5%		
Inversion	Split reads	50 (Pindel) or 5 (BLAT)		
	Read pairs	10%		
Deletion	Split reads	40 (Pindel) or 15 (BLAT)		
Tandem duplications	Split reads	30 (Pindel) or 30 (BLAT)		
	Read pairs	10%		

Figure 2-3 Principle of split reads and read pairs to filter structural variants
The figure visualizes, how split reads or read pairs can support the existence of an SV. The number of supporting split reads and read pairs, which were required to pass the quality filter, are mentioned in the column “Passed filter’. Split reads were either detected using a BLAT alignment (unmapped reads) or reported by the program Pindel. The percentage of the read pairs was based on the number of read pairs supporting the SV divided by all read pairs in the relevant region multiplied with 100.

To identify somatic variants, the quality filtered SVs called in the tumor samples were compared with the unfiltered SVs detected in the controls. An interchromosomal translocation was defined as somatic, if no breakpoint of an interchromosomal translocation detected in the controls was located in the neighborhood of the interchromosomal translocation breakpoints called in the tumor samples (distance < 100 bp). A deletion, inversion or tandem duplication was classified as somatic, if none of the respective SVs detected in the controls overlapped with the SV regions called in the tumor samples.

The visualization of large SVs was performed with Circos version 0.62 [161]. All steps of the variant calling and filtering pipeline applied in the GC study are summarized in Figure 2-4.

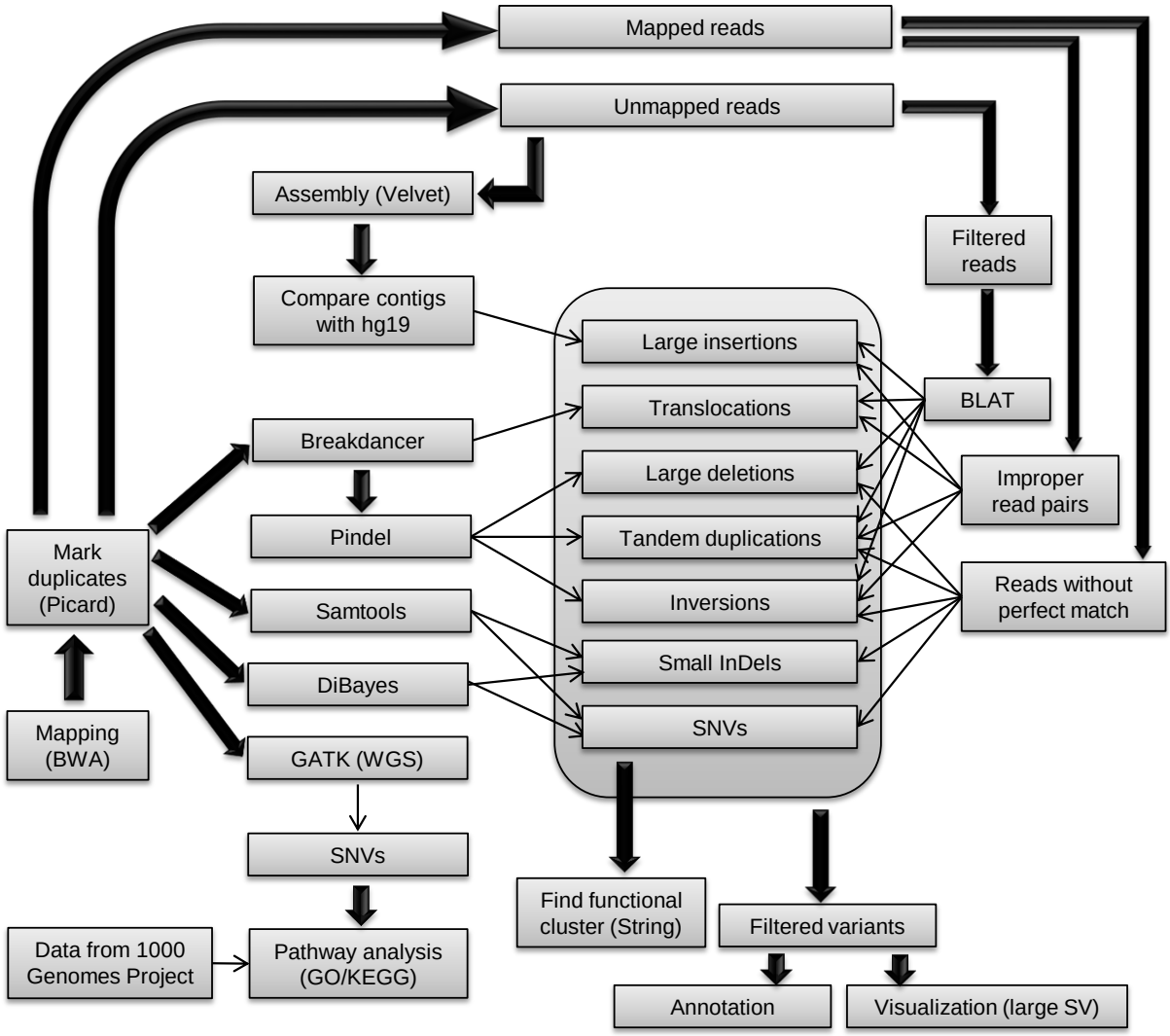


Figure 2-4 Variant calling and filter pipeline for human samples (GC project)

2.2.8 Investigation of differentially expressed genes

The read counts per transcript were calculated with the Python script HTSeq version 0.5.4 [162] for reverse stranded libraries. Reads overlapping with more than one feature were handled with the mode ‘intersection-strict’. This setting uses exclusively sequences that are completely aligned within the exon boundaries and that can unambiguous be assigned to one gene. Reads with an alignment quality lower 20 were skipped. The p-values for differentially expressed genes were calculated with DeSeq2 version 1.4.5 [163]. The applied methods provided by the DeSeq2 package were based on (i) the estimation of size factors, which control for differences in the library sizes of the sequencing experiment, (ii) the estimation of dispersion for each gene, and (iii) negative binomial generalized linear models (‘nbinomWaldTest’). To improve stability, a shrinkage estimation for dispersions and fold changes were applied. The function ‘replaceOutliersWithTrimmedMean’ was used to detect and remove outliers for each gene.

The distances between samples were calculated using the R function 'dist' with Euclidean distance based on regularized logarithm transformed (‘rlog’) read counts per gene. The R function 'rlog' creates log2 scaled values, which have been normalized with respect to the library size. With this method differentially expressed genes of all expression levels are supposed to have the same influence on the distance calculation. Afterwards, classical multidimensional scaling (MDS) was performed with the function 'cmdscale' [131] using a two-dimensional space for representation of the data. The MDS plot was created with the resulting eigenvalues using the R function 'ggplot' [164]. This analysis was also used for the detection of sample outliers. After outlier removal, all steps after counting the reads per gene were repeated.

FPKM (fragments per kilobase of exon per million fragments mapped) values were computed with Cufflinks version 2.0.2 [165]. A bias detection and correction algorithm was applied as well as an initial estimation procedure to weight reads, which mapped to multiple locations in the genome. Exclusively fragments that aligned to a position of an annotated transcript were used for normalization.

Heatmaps were created by applying the R function 'heatmap.2'. The dendrograms were based on hierarchical clustering performed with the R function 'hclust'. The input matrix for the hierarchical clustering was calculated with the R function 'dist' using Euclidean distance. For the investigation of specific disease characteristics such as tumor stage, tumor localization, and intestinal prolapse, a k-means clustering using the R function ‘kmeans’ was applied instead of a hierarchical clustering.

2.2.9 Detection of novel transcriptionally active regions

Novel transcriptionally active regions (nTar) are expressed genomic loci, which are not annotated yet. This include e.g. extended UTR regions, additional exons, and undetected coding or non-coding genes. At the point in time of my analyses, no publicly available tool was able to detect, characterize, and filter nTars in detail. To fill this gap, a new program was developed in the course of my studies.

The novel nTar detector was written in Perl and runs on all Windows as well as Linux system with an installed Bio-Samtools version (version 5.10.1 or higher). The program requires the annotation of known transcripts in Gene Transfer Format (gtf) or General Feature Format 3 (gff3) and the alignment results in mpileup (only available for unstranded sequencing data), sam or bam format as input. The workflow and features are summarized in Figure 2-5. After mapping, the tool detects all covered regions, which have not been annotated yet. Thereby, strand-specific as well as unstranded RNA-Seq data can be handled. This initial step is followed by a quality control to distinguish between mapping artifacts and real nTars. This includes the calculation of a mean and median alignment penalty score, which is based on the

number of mismatches, inserted or deleted bases of a read. Moreover, the average mapping qualities provided by the alignment tool and the average base qualities can help to exclude false positive nTars caused by mismapped reads in e.g. repeat regions. The credibility of an nTar is further assessed by the average coverage, the length of the nTar as well as the number of supporting reads and start points.

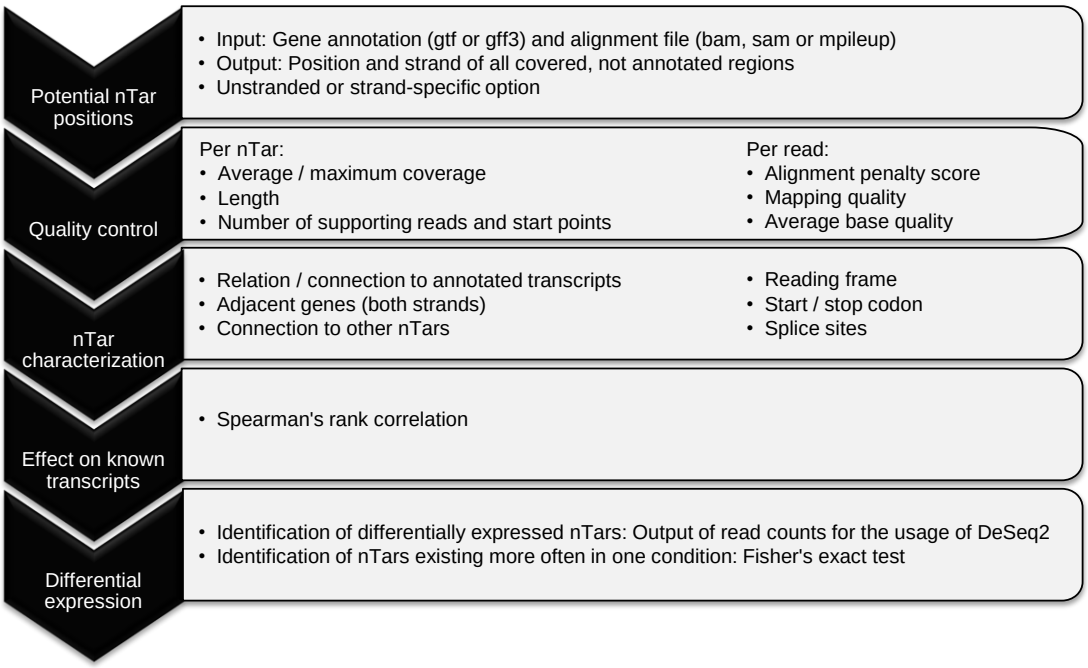


Figure 2-5 Workflow and features of the nTar detector

After the quality filtering, the program offers the possibility to characterize the nTars, i.e. to check for reading frames, start / stop codons, and splice sites. Furthermore, read pairs and split reads are used to investigate whether the potential nTar is connected with any other known or novel transcripts. As described earlier in Klostermeier et al. [166], nTars are assigned to the following region classes (Figure 2-6): (i) The regions upstream gene neighborhood (UGN) and downstream gene neighborhood (DGN) include nTars, which have a distance of less than 10,000 bp to the next annotated gene but do not overlap with any known exons. (ii) Up- and downstream gene intersecting events (UGI/DGI) start directly at the UTR region of an annotated gene without any gap. (iii) Intronic nTars, which are immediately adjacent to a known exon without covering the whole intron, are classified as exon-linked downstream (ELD) or exon-linked upstream (ELU). (iv) An nTar is defined as intron-spanning element (ISE), if the nTar covers the whole intron, while (v) an nTar that is located in an intron region without direct contact to a known exon is called intronic gene element (IGE). (vi) If an nTar does not belong to any of these classes, it is described as intergenic element (INTER, not shown in Figure 2-6). In case of stranded RNA-Seq data, the nTar region must have the same orientation as the annotated transcript to be classified into the respective category. Thus, four additional classes were defined for nTars on the opposite strand: (a) antisense exonic (ASE), if the nTar is antisense to an exonic region, (b) antisense intronic (ASI), if the nTar is antisense to an intron

and (c) antisense upstream/downstream neighborhood (AS_UGN / AS_DGN), if the distance to a known gene on the other strand is smaller than 10,000 bp without any overlap to the exon, respectively. Due to different isoforms and genes, a single nTar event might belong to several of the defined classes. Thus, the program shows for each nTar the classes of all affected isoforms separately as well as a merged version for each affected gene. Regardless of the region category, the closest genes on both strands in both directions are shown in the output file of the tool.

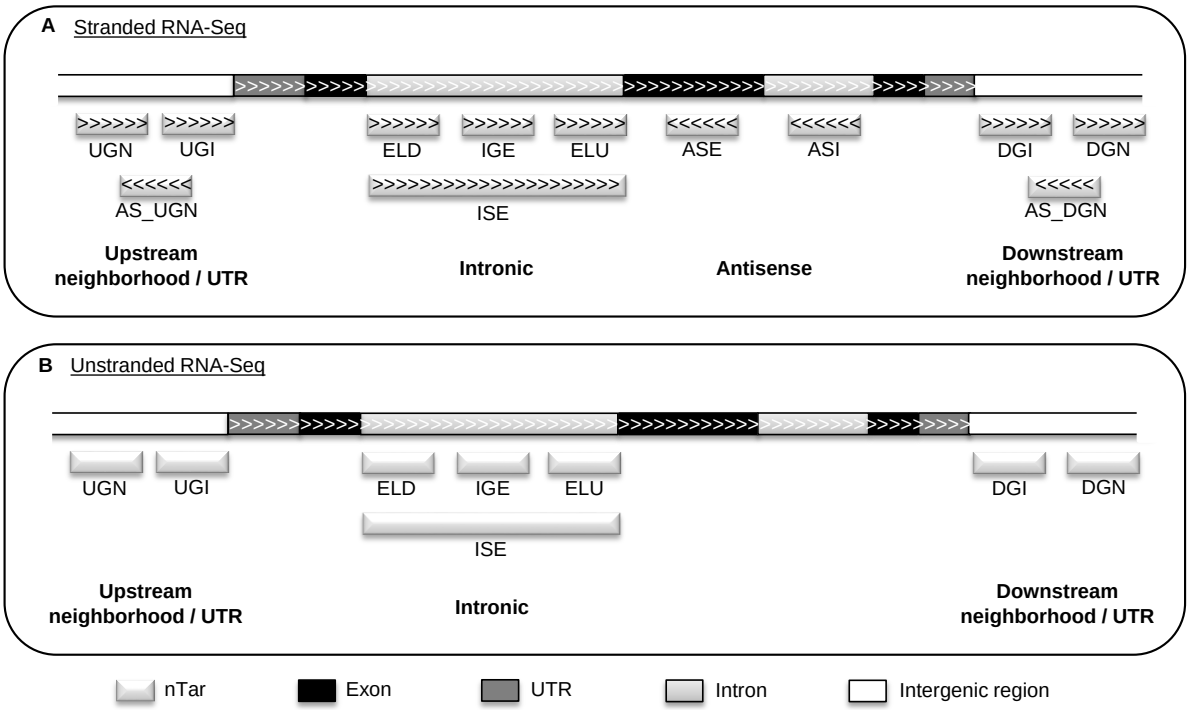


Figure 2-6 Description of possible nTar regions
(A) Possible nTar regions after sequencing with a strand-specific library protocol. (B) Possible nTar regions after sequencing with an unstranded library protocol. UGN = upstream gene neighborhood, UGI = upstream gene intersection, AS_UGN = antisense upstream gene neighborhood, ELD = exon linked downstream, IGE = intronic gene element, ELU = exon-linked upstream, ISE = intron-spanning element, ASE = antisense exonic, ASI = antisense intronic, DGI = downstream gene intersection, DGN = downstream gene neighborhood, AS_DGN = antisense downstream gene neighborhood.

Another feature of the presented program includes the support of intersample comparison of nTars. The first option enables to investigate the relation of the nTar expression level to known transcripts. For each nTar, a Spearman's rank correlation is performed with all genes, which are adjacent, overlapping (intronic or antisense) or connected. The second option of the intersample comparison helps to detect differentially expressed nTars between different sample groups. Therefore, the program offers two possibilities: (i) For each nTar, the read counts are provided as possible input for other publicly available tools such as DeSeq2. (ii) Additionally, it is possible to apply a Fisher's exact test to investigate whether significantly more samples of one condition harbor the nTar in comparison to another group. For all intersample comparisons, a strict and a merged alternative exists (Figure 2-7). The strict

version compares exclusively nTars having the exact same position. In the merged model, all overlapping nTars of all samples are merged together to one ‘meta-nTar’.

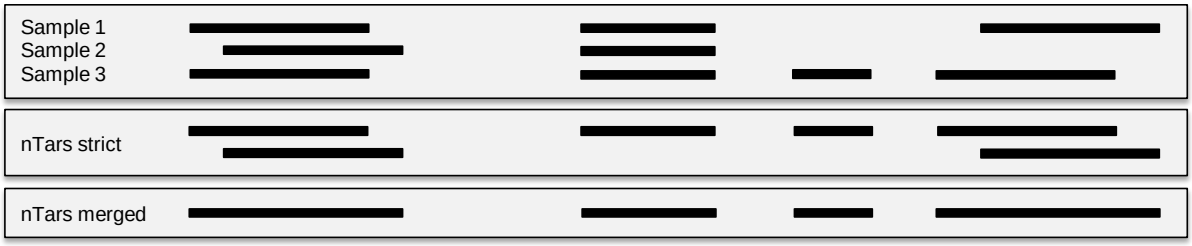


Figure 2-7 Differences between strict and merged nTar positions

The following parameters were used to detect nTars associated with murine AOM/DSS-induced colorectal cancer: The minimum length for nTars was set to 50 bp, while the average coverage had to be larger or equal to ten. A minimum of ten supporting reads with different start points was required. The threshold for the median alignment penalty score was set to three. The average base quality score had to be larger than 15 and the average mapping quality at least 30. For the investigation of differentially expressed nTars, the merged version was applied

2.2.10 Investigation of splice variants

A direct investigation of transcript abundances might be inaccurate, because in regions covered by several genes or isoforms, an unambiguous allocation of the reads to the transcripts is impossible and thus, the expression level can only be approximated. Therefore, each splicing type was considered separately. First, an inclusion level of each internal exon was calculated with the PSI (Percent Spliced In) value. The measure was computed according to Wang et al. [167], for which three types of read positions were used: (i) number of reads that mapped completely to the investigated exon, (ii) number of split reads that supported the inclusion of the exon, (iii) number of split reads that skipped the exon and thus supported the exclusion of the exon (Figure 2-8). The PSI value was defined as the number of reads supporting the inclusion of the exon divided by the number of reads supporting either the inclusion or the exclusion of the exon. In particular, a PSI value of one means that the exon is fully included, while zero stands for an exclusion of the exon.

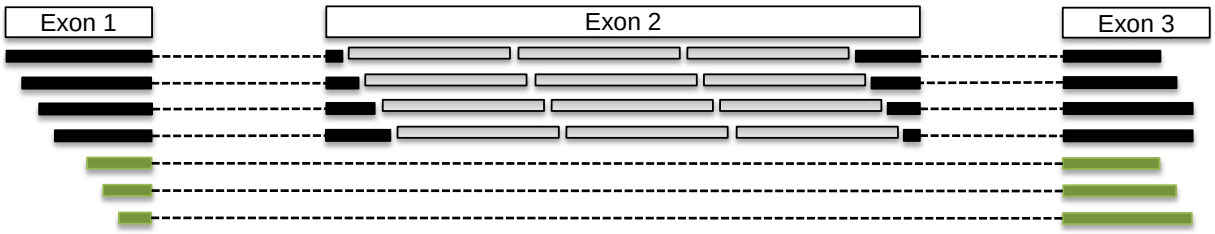


Figure 2-8 Calculation of Percentage Spliced In values
The calculation of the PSI value was based on three read types: (i) reads mapping completely to the exon of interest (grey), (ii) split reads connecting an adjacent exon with the exon of interest (black), and (iii) split reads connecting the adjacent exons (green). The PSI was defined as the number of reads supporting the inclusion of the exon (grey + black) divided by the number of reads supporting the skipping of the exon (green).

As second splice form, intron retention events were investigated. Therefore, the number of reads mapping to a specific intron was divided by the number of reads mapping to the adjacent exons. To avoid artifacts, the average coverage of the intron had to be larger than 2x in at least one sample type, while the ratio between intron and exon had to be larger 0.2 at the same time. A similar method was applied to test for skipping of the first or last exon. In this case, the ratio between reads mapped to the outer exon and the adjacent exon was calculated.

In the third approach, shifts of the donor or acceptor splice sites were investigated. Therefore, the ratios between the two major donor sites across all samples were calculated at all known acceptor sites. Splice events and donor sites were defined using the existence of split reads. Reads covering the splice site and at least 20 bp of the intronic region supported the case that no splice event exists at the investigated position in the library. The same approach was vice versa applied for all annotated donor sites.

For all computed ratios and PSI values, a pair-wise comparison between tumor and control samples was performed with a two-tailed Wilcoxon Rank-Sum test followed by a Benjamini-Hochberg correction for multiple testing.

2.2.11 Gene set enrichment analyses based on SNVs and small InDels

Most of the publicly available tools for pathway analyses were designed for transcriptome studies and do not consider parameters like gene length or gene conservation that bias the number of variants per gene. Therefore, for each pathway, the number of non-synonymous SNVs and InDels as well as the number of genes affected by at least one non-synonymous SNV or InDel were compared between tumor samples and controls.

To perform gene set enrichment analyses despite the small group sizes in the GC study, germline SNVs of 1092 samples out of the 1000 Genomes Project were annotated with Annovar version 7 and used as control set. For each pathway, the number of genes affected by a non-synonymous variant as well as the maximum and average number of all SNVs were computed for each sample. Besides the investigations based on all non-synonymous variants, the analyses were repeated using only SNVs, which were predicted to be damaging to protein function. Due to the investigation of just two tumor samples, the power of the study was too low to calculate a reliable p-value. Therefore, the number of affected genes/SNVs per pathway was compared between the tumor samples and the average/maximum count in the samples of the 1000 Genomes Project. This strategy revealed pathways, which were more often affected than in healthy individuals. To ensure the comparability between samples investigated in this study and those of the 1000 Genomes project, exclusively SNVs called with GATK using the same parameters like in the 1000 Genomes project were considered for this part of the study. The analyses were performed for all pathways annotated in the KEGG database and all terms described in the GO database.

In the CAC study, tumor samples were compared with the pool of control and DSS samples. Significant changes in pathways annotated in KEGG were assessed employing a one-tailed Wilcoxon Rank-Sum test. The false discovery rate was controlled with the Benjamini-Hochberg approach.

2.2.12 Gene set enrichment analysis based on genes

For pathway analyses, which were independent of gene length and gene conservation, publicly available tools were applied. These programs were used for genes, which were previously filtered by e.g. variant count on gene level or expression values.

Gene set enrichment analyses based on pathways annotated in the KEGG database were performed for differentially expressed genes with the online tool InnateDB [168], whereby the correction for multiple testing was performed with the method developed by Benjamini and Hochberg using the R function ‘p.adjust’. The gene set enrichment analyses based on gene ontology (GO) terms was performed with the program g:Profiler [169]. Only terms categorized as ‘Biological process’ or ‘Molecular function’ were subjected to this analysis. The intersection size (Q&T) between GO-term and input genes was set to two. The false discovery rate was again controlled with the Benjamini-Hochberg approach.

The online tool Pscan v1.4 [170] was applied to investigate whether promoters of differentially regulated genes ($p<0.001$ an fold change > 4) were enriched for any conserved transcription factor binding sites.

3 Results

Parts of the GC study have been published in Esser et al. (2017) [171]. Parts of the study to investigate the molecular background of murine inflammation-associated colorectal cancer have been submitted for publication [172]. This is the case for all chapters, including introduction, material and methods, results, and discussion.

3.1 Investigation of human gastric cancer

The complete workflow of the human GC study is summarized in Figure S 1.

3.1.1 Study patients

Samples from two female patients were retrieved from the archive of the Institute of Pathology at University Hospital Kiel. The first patient died from a moderately differentiated, highly microsatellite unstable, intestinal type GC of the antral mucosa (tumor stage pT3 pN0 (0/20) L0 V0 R0 G2) without histological evidence of an EBV-infection. The second patient was diagnosed with a moderately differentiated, microsatellite stable, intestinal type GC of the antral mucosa (tumor stage pT4 pN1 (2/21) L0 V0 R0 G2).

3.1.2 Sequencing and mapping results

The whole exome and whole genome of one MSI (first patient) as well as one MSS (second patient) gastric carcinoma and the matched healthy tissues were sequenced with the SOLiD™ 4 System from Applied Biosystems / Life Technology and the Illumina HiSeq 2500, respectively. In the WES (SOLiD), between 160 and 199 million reads with a length of 50 bp (forward read) and 35 bp (reverse read) were produced per sample. Between 64% and 74% could be aligned to the reference, of which 24% to 38% were marked as duplicates. Around 60% of the bases were located in enriched genomic sections resulting in a median coverage between 28x and 41x in the target regions (Table S 7, Figure 3-1 A). The output for the WGS (Illumina) contained between 2.8 and 6.3 billion reads with a read length of 100 bp. Over 90% could be mapped to the genome, of which between 15% and 72% were marked as duplicates. This resulted in a median coverage of 25x to 84x (Table S 8, Figure 3-1 B). Although the percentage of covered bases was lower in the non-tumor sample of the second patient (WGS data), all reported filtered variants were covered by at least three reads in both data sets of the second patient. It has to be noted that the coverage in the WES data of the MSI tumor (first patient) was lower than in the other data sets. This could have led to a detection lack of some variants, but had no influence on the reported variants and the described mutational patterns.

The sequencing results confirmed that neither a viral nor a bacterial infection was present in the tumor samples. Also a contamination could be excluded.

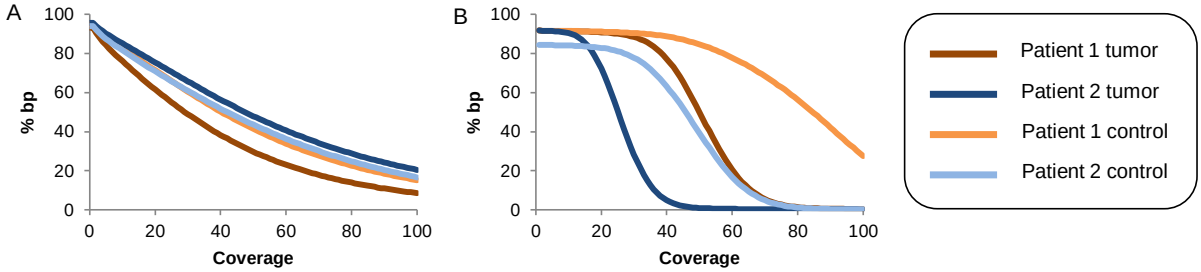


Figure 3-1 Coverage distribution of samples from patients with GC
(A) Results based on WES. (B) Results based on WGS.

3.1.3 Investigation of variants

The workflow to filter variants including the number of reads and variants after each step is summarized in Figure 3-2.

Sample preparation

Sequencing
WGS: Illumina HiSeq 2500, WES: SOLiD 4

Mapping

Remove duplicates

Variant calling
germline + somatic

Variant filter
quality filter, exclude germline variants, variant annotation

SNV pattern + pathway analysis

Sample	Patient 1 tumor (MSI)	Patient 2 tumor (MSS)	Patient 1 non-tumor	Patient 2 non-tumor
# reads WGS	2,818,634,692	6,269,201,959	3,117,880,513	3,165,759,194
# reads WES	160,307,884	195,373,036	179,749,208	198,844,362
% mapped reads WGS / WES	99.5 / 63.7	95.7 / 73.8	93.0 / 71.4	90,1 / 71.5
% on target WES	62	65	61	57
% duplicates WGS / WES	44.7 / 24.3	71.9 / 30.6	14.6 / 27.8	50.3 / 38.2
# SNVs WES (DiBayes / Samtools / GATK)	80,600 / 41,924 / 52,109	85,138 / 47,945 / 58,801	87,440 / 49,515 / 60,908	97,771 / 46,957 / 60,224
# SNVs WGS (Samtools / GATK)	3,432,126 / 3,244,363	3,274,555 / 3,046,801	2,546,586 / 3,216,067	3,420,179 / 3,257,008
# small InDels (Samtools WGS / DiBayes WES)	1,234,100 / 4,425	907,184 / 4,137	773,010 / 4,491	930,926 / 4,370
# large InDels # insertions / # deletions	75 / 191,076	0 / 122,651	unknown / 92,643	unknown / 106,420
# inversions	3,370	6,125	2,541	1,036
# interchromosomal translocations	29,068	72,444	1,513	12,162
# tandem duplications	1,652	1,199	1,548	1,488

	Patient 1 tumor	Patient 2 tumor
# somatic, exonic, novel, damaging or conserved, non-synonymous SNVs (damaging & conserved)	202 (52)	76 (7)
# somatic, exonic, novel small deletions / insertions (frameshift)	132 (113) /131 (126)	13 (9) / 23 (23)
# somatic large deletions / insertions	1,739 / 8	78 / 0
# somatic inversions	22	95
# somatic interchromosomal translocations	27	63
# somatic tandem duplications	0	0

Figure 3-2 Workflow of WES and WGS including subsequent data analyses
The figure displays the number of variants and reads per sample after each step in the analyses of the GC study.

3.1.3.1 SNV and small InDel calling

Several variant caller were used to detect SNVs and small InDels, which were filtered by allele fractions afterwards. This methodology was applied to identify the most robust call sets. Depending on the sequencing data set and the applied SNV caller between 41,924 and 97,771

germline or somatic SNVs were detected in the WES and around three million SNVs in the WGS mappings (Figure 3-2). The numbers of raw somatic SNVs are shown in Table S 9. A comprehensive comparison of the variant callers (supplementary chapter 6.2.1.2) demonstrated a low number of SNVs exclusively detected with GATK. Thus, the results of this caller were only considered for pathway analyses to enable a comparison with samples from the 1000 Genomes project (chapter 3.1.4). The union of all variants called by at least one caller (except GATK) in the WES or WGS data, which had a supporting mutant allele fraction larger 5% in the WES and WGS data, were defined as cross platform variants. All variants reported by one of the applied variant callers without verification with the second sequencing technology were named single technology variants. Out of the cross platform variants, 202 and 76 filtered SNVs were somatic (i.e. present only in the respective tumor sample) in the MSI tumor of the first patient and the MSS tumor of the second patient, respectively. Based on the mutant allele fractions of somatic SNVs called in the WGS data, the tumor cellularity was estimated to be around 30% for the MSI and 48% for the MSS tumor sample (Figure S 8).

In the WGS data, 1,234,100 germline or somatic InDels (324,205 insertions / 909,895 deletions) were detected in the MSI tumor sample and 773,010 germline or somatic InDels (292,938 / 480,072) in the control sample of the first patient. In the WES data of the first patient, 4,425 and 4,491 germline or somatic small InDels were identified in the MSI tumor and control sample, respectively. This resulted in 239 cross platform InDels, which were somatic, frameshift, and novel. In the WGS data of the second patient, 907,184 germline or somatic small InDels (346,265 / 560,919) were called in the MSS tumor sample and 930,926 (363,052 / 567,874) in the matching control. In the WES data of the second patient, 4,137 germline or somatic small InDels were found in the MSS tumor sample, while 4,370 germline or somatic small InDels were detected in the corresponding control (Figure 3-2). Out of this, 32 cross platform small InDels were somatic, frameshift, and novel in the MSS tumor of the second patient.

3.1.3.2 Mutational landscape of somatic SNVs in GC samples

The relative contribution of the six SNV classes (C>A, C>G, C>T, T>A, T>C, T>G) including the flanking bases were investigated for all somatic variants of the tumor samples called in the WGS data and for all SNVs of the corresponding controls (Figure 3-3). The transition/transversion ratio was similar between all WGS samples (2.2 in all samples). In contrast, the ratio was decreased in the exonic regions of the WES data in both tumor samples (2.2) compared to the controls [2.7 (patient 1) and 2.5 (patient 2)]. The relative contribution of each SNV class including the context distribution was almost identical between the control samples. In contrast, an overrepresentation of C>T base substitution, especially in the context of GpCpG and ApCpG trinucleotides, was observed in the MSI tumor sample of the first patient.

In the MSS tumor of the second patient, a prevalence of T>G in context of a five prime cytosine was detected, while a decrease of T>C, especially in context of a five prime thymine, was observed. Besides the different SNV type distribution, the MSI tumor of the first patient was characterized by a significantly higher total amount of somatic SNVs than the MSI tumor of the second patient ($p < 0.0001$).

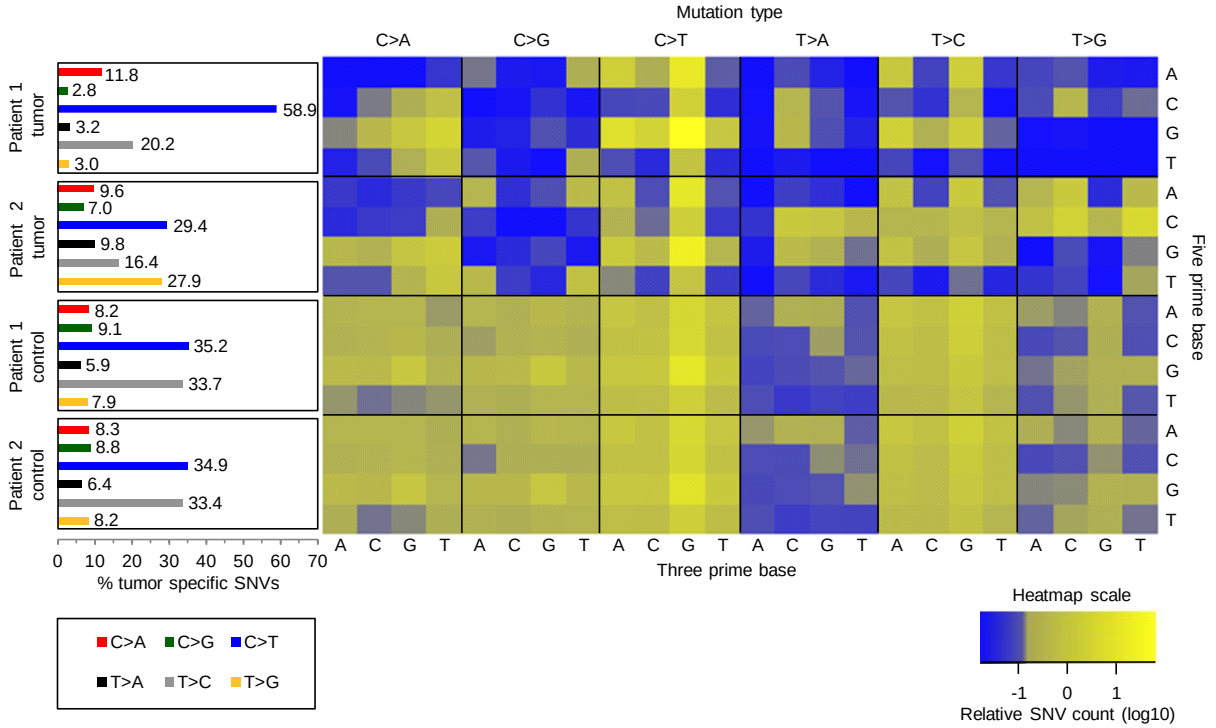


Figure 3-3 Comparison of SNV patterns including flanking bases
The left part of the figure shows the SNV type distribution for each sample. The heatmap on the right side illustrates the relative abundance of each SNV including the flanking bases. Exclusively somatic SNVs were considered for the tumor samples, while all variants were used for the controls.

The somatic SNV spectrums observed in the tumor samples have been compared with molecular signatures reported in the COSMIC database to identify factors with impact on the development of GC. While COSMIC's signatures 1 (49%), 6 (13%), and 15 (38%) contributed to the mutational pattern of the MSI tumor, signature 1 (46%) and 17 (35%) as well as unknown factors influenced the SNV distribution of the MSS tumor (Figure 3-4). The underlying biological processes included aging and spontaneous deamination (signature 1, mainly causing C>T), defective DNA mismatch repair in MSI tumors (signature 6, C>T and C>A), and DNA mismatch repair (signature 15, C>T), while causes of signature 17 (T>G) were unknown. Signature 15 is often observed in cases with a high amount of InDels, which was also the case in the investigated MSI tumor sample (Figure 3-2). However, compared to the observed SNV spectrums, the assembled signatures harbored a residual sum of squares (RSS) of 0.308 for the MSI tumor sample and 0.0149 for the MSS tumor sample. The RSS measures the differences between the values predicted by the estimation model and the read data. This illustrates that a high fraction could not be explained by signatures described in the COSMIC database. The highest deviations between assembled and observed SNV distribution were

observed for T>C variants in an ATA, ATG, GTA, and GTG context as well as for C>T variants in the first patient (MSI tumor), while differences were manifold in the second patient (MSS tumor) (Figure 3-4). The special characteristics of the GC samples in the current study were also reflected in clearly distinct SNV type distributions between the investigated tumor samples and WES studies from TCGA (Figure S 9).

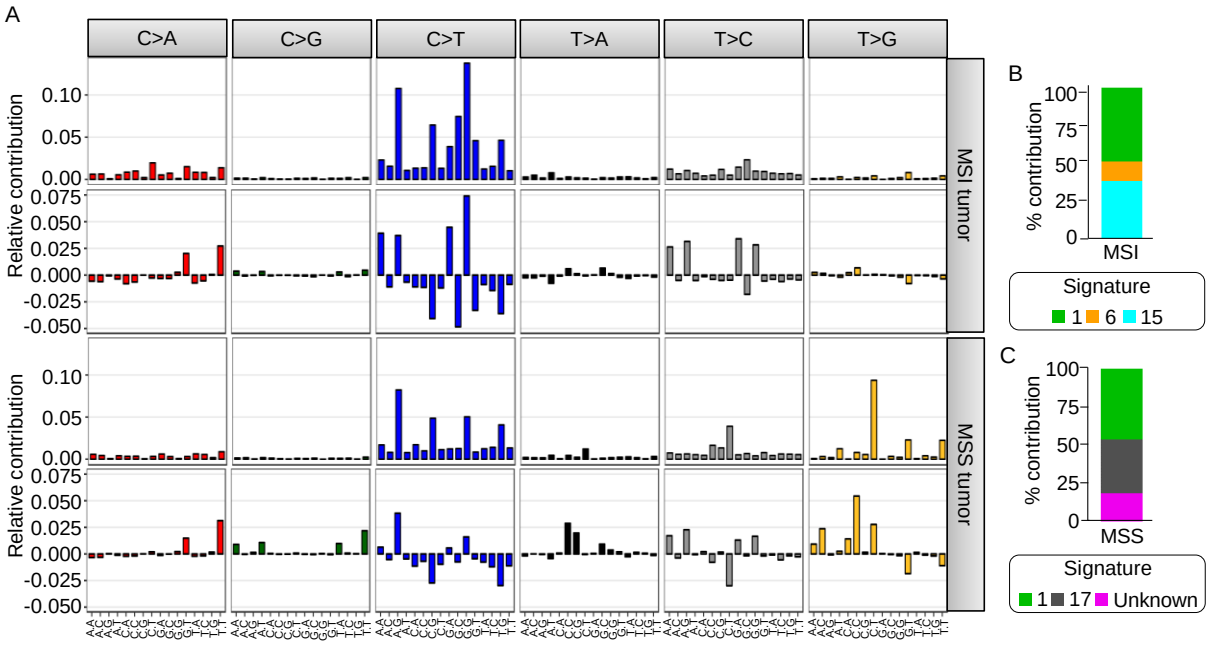


Figure 3-4 Reconstructed signatures of SNV spectrums
(A) Decomposed signatures were computed with non-negative matrix factorization based on signatures described in the COSMIC database. The upper row of each sample block shows the SNV type distribution of the reconstructed signature for the MSI and MSS tumor sample, respectively. The lower rows indicate the differences between the assembled and the observed somatic SNV spectrums, respectively. COSMIC signatures contributing with more than one percent to the reconstructed SNV spectrums are shown in (B) and (C).

The observed SNV spectrums could not completely be reconstructed with known cancer-associated SNV signatures. Thus, somatic signatures of the investigated GC samples were estimated with the statistical method non-negative matrix factorization (NMF) to identify common underlying patterns. The first signature was dominated by C>T base substitutions, especially in the context of ApCpG and GpCpG, while the second signature harbored mainly C>T variants in an ApCpG context and T>G SNVs in a CpTpT and CpTpC context (Figure S 10 A). The MSS tumor was mainly influenced by the first signature, while the second signature contributed to a large proportion to the SNV pattern observed in the MSI tumor (Figure S 10 B). Also in this analysis, it was clearly visible that MSI and MSS harbored distinct somatic SNV patterns.

A *kataegis* effect is defined as a local genomic region with hypermutation [153]. In the course of my studies, DNA sections with a high somatic SNV density, potentially representing *kataegis* events [153], were visualized by rainfall plots, in which the distances between each somatic SNV and the previous somatic SNV were plotted (Figure 3-5). Intragenic SNV clusters have neither been detected with this method nor with the more sensitive sliding window approach described in section 2.2.7.4.

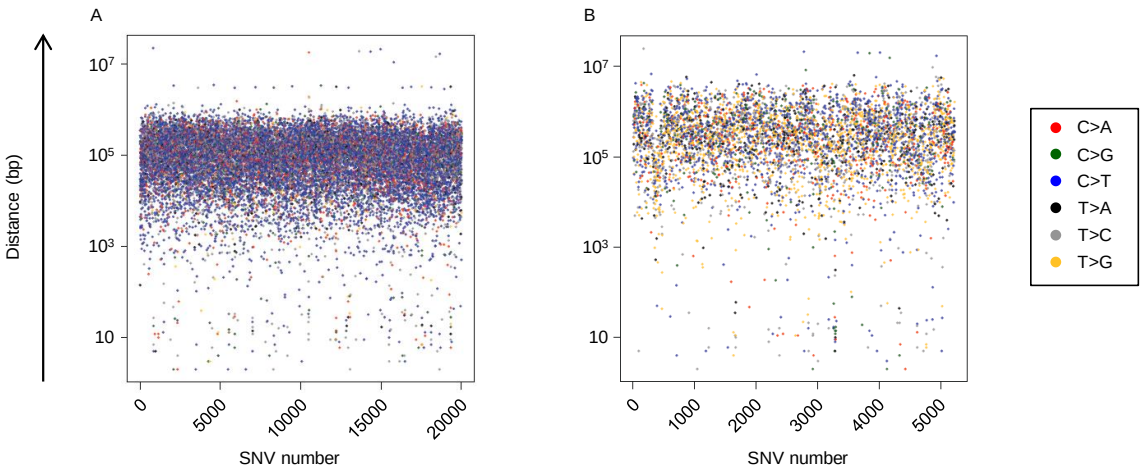


Figure 3-5 SNV density plots of somatic variants in tumor samples
In the rainfall plots, the SNVs ordered by genomic position are shown on the x-axis. The distance between each SNV and the previous one is plotted on the y-axis. The colors of the dots describe the SNV type. The figures are based on the somatic SNVs of the tumor sample of (A) patient 1 (MSI) and (B) patient 2 (MSS).

3.1.3.3 Comparison between observed and known cancer-associated SNVs

The cross platform approach was exploited to minimize the systematic technological bias and thereby to increase the reliability of the variant positions. 52 cross platform novel base substitutions in the MSI tumor (patient 1) and seven cross platform novel variants in the MSS tumor (patient 2) were (i) classified as stopgain SNV or (ii) predicted as damaging as well as at a conserved position or in a conserved gene. Out of these alterations, 41 (MSI tumor) and six (MSS tumor) SNVs did neither exist in ExAc nor in ESP (Table S 10 and Table S 11, respectively). Further 87 (MSI tumor) and 58 (MSS tumor) SNVs were either predicted as damaging or at a conserved position (Table S 12 and Table S 13, respectively). Of particular note is that four genes affected by a somatic SNV in the MSI tumor (*AFF3*, *DROSHA*, *JAK2*, and *PIK3CA*) and one gene affected by a somatic SNV in the MSS tumor (*GATA3*) were also listed in the cancer gene census list of the COSMIC database and were not mentioned in the exome sequencing databases ExAc or ESP. The somatic SNV, which was found in *PIK3CA* in the MSI tumor, was reported to be associated with GC in the COSMIC database. *PIK3CA* is an oncogene [173], which is frequently mutated in gastric adenocarcinomas [174].

In the MSI tumor of the first patient, 77 (74 frameshift) cross platform exonic novel small InDels were supported by the variant allele fractions of the WES and WGS data. This variant category comprised two (one frameshift) InDels in the MSS tumor of the second patient. Additionally, 186 / 34 somatic InDels (165 / 31 frameshift) were detected by one variant caller with at least two supporting reads in the WGS data but covered by less than three reads in the WES results in the MSI / MSS tumor sample. All InDels, which were not reported in the databases ESP or ExAC, are shown in the supplementary Table S 14 and Table S 15, respectively. InDels were also excluded from the supplementary lists, if the exact same InDel type was called in one of the six adjacent base pair positions in the matching controls. Out of the genes affected by a somatic small InDel in the MSS tumor, eight were also listed in the cancer gene census table

of the COSMIC database including *BRAF*, *ZFHX3* (both associated with GC), *ARID1A*, *GNAS*, *HOXC13*, *MAPK1*, *PREX2*, and *SPECC1*. In the MSS tumor, the genes *GNAS* and *HOXD13* affected by single technology InDels, respectively, were listed in the cancer gene census list of the COSMIC database.

Due to the function described in the literature and because of being part of COSMIC's cancer gene census list, *BRAF* was an especially interesting candidate. Noticeable, a germline SNV predicted as damaging was in just one sample out of the 1000 Genomes project located in *BRAF*. Further 38 samples harbored a germline frameshift InDel in *BRAF*. In all 38 samples, the InDel was located at the same position in the last exon. In contrast, in the MSI tumor sample of patient 1, a cross platform somatic frameshift InDel was found in exon 10 (18 exons total).

To detect predisposing somatic or germline alleles, cancer-associated variants annotated in the databases OMIM, HGMD, and GWAS were compared with all somatic and germline variants called in the investigated tumor samples. In total, 301 / 296 known cancer-associated SNVs (somatic or germline) were detected in the tumor samples of patient 1 (MSI) / patient 2 (MSS) including SNVs in *BRCA1* (MSI), *DCC* (MSI), *FLCN* (MSI), *LOX* (MSS), *RASSF1* (MSS), and *TERT* (MSS) (all germline). SNVs affecting exonic regions of the following genes were associated with GC (Figure 3-6): *MUC1* (germline, MSI), *NOD2* (germline, MSI+MSS), *CCL22* (germline, MSS+MSI), *VCAN* (germline, MSI), *PLCE1* (germline, MSS), and *TP53* (somatic, MSS). Out of the mentioned variants, 142 (MSI) / 141 (MSS) (overlap = 101) were found in the GWAS catalog, 82 / 72 (overlap = 61) in the database OMIM and 135 / 137 (overlap = 89) in HGMD. Nearly all SNVs were germline in the investigated tumor samples and thus considered as potentially predisposing, but not necessarily causative. One exceptional somatic mutation in *TP53* was noted (rs28934574, ExAc_Aggregated_Populations frequency = 1.647E-04) in the MSS tumor sample. This cross platform SNV was located in a cancer hotspot region [175] and has also been identified in earlier studies. As example, it was detected in a 10 year-old proband with an extensive family history of malignant tumors with an unusual prevalence of GC on the paternal side [176]. Besides rs28934574, two further cross platform SNVs were detected in cancer hotspot regions in *GRIN2* (T100 and A50, germline SNVs in MSS and MSI tumor samples) [175].

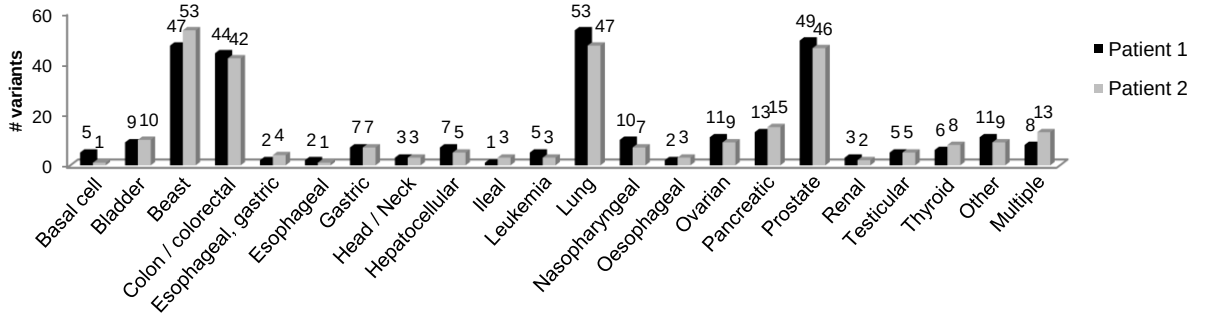


Figure 3-6 Number of known cancer-associated SNVs detected in the two GC samples
All known cancer-associated SNVs can potentially contribute to GC, even if an association is not known so far. Thus, all known cancer-associated SNVs were compared with somatic variants detected in the current study.

3.1.3.4 Somatic SNVs and small InDels without known GC association

Besides the detection of somatic SNVs and somatic InDels in genes recurrently mutated in tumor samples (described in section 3.1.3.3), cross platform somatic SNVs predicted as damaging at a conserved position might be involved in the development of GC. To further filter clinically relevant exonic small variants, the simplified indicator ECS was introduced (chapter 2.2.7.3) and validated with previous cancer-related scores (supplementary chapter 6.2.1.4). Based on a low value of the ECS and functional aspects, the most promising SNVs were in the genes *MSH4* (MSI), *PRDM2* (MSI), *GATA3* (MSS) (Table 3-1). Genes, which harbored a low ESC score and which were affected by an InDel, included *TP53I3* (MSI) as well as *ZIC1* and *IL17RD* (both MSS) (Table 3-1).

In addition, in patient 1 (MSI), four SNVs were found in non-exonic splice sites of the genes *C3*, *IDE*, *STAMBP*, and *NPSR1*. Interestingly three of them interact with the *CASR*, which was also affected by a non-synonymous SNV. The third gene (*IDE*) was via one interconnecting gene associated with *CASR*. In patient 2 (MSS), SNVs in splice sites of the genes *ARID5A* and *DMD* were detected.

Gene symbol	Official full name	Description
GATA3	GATA binding protein 3	• Involved in breast cancer progression and metastasis [177] • Somatic copy-number alterations found in *GATA4* and *GATA6* in GC [178]
IL17RD	Interleukin 17 receptor D	• Tumor suppressor [179] • Downregulated in breast, prostate, thyroid, and ovarian cancer [179]
MSH4	mutS homolog 4	• Plays a role in meiotic and mitotic DNA double strand break repair and DNA damage responses in human cells [180] • Mutations leading to dysfunctional hMSH4 may be involved in oncogenesis [181]
PRDM2	PR domain containing 2, with ZNF domain	• Tumor suppressor [182] • Decreased of *PRDM2* in GC [183] • Frameshift mutations may play an important role to GC with MSI [184].
TP53I3	Tumor protein p53 inducible protein 3	• Involved in p53-mediated cell death [185] • SNP at the downstream microsatellite sequence might be associated with differential susceptibility to cancer [186].
ZIC1	Zic family member 1	• Downregulated in GC tissues and cell lines [187] • Tumor suppressor in GC [187] • Potential therapeutic target for GC [188]

Table 3-1 Description of functional interesting genes harboring a somatic SNV or InDel
All listed genes were affected by a somatic SNV with predicted effect on the protein function or a somatic InDel and, based on the function, potentially associated with the development of GC. All genes were characterized by a low ECS value.

The cross platform non-synonymous somatic SNVs in the following genes were independently validated by pyrosequencing on a Pyromark Q24 device (QIAGEN) as third validation step: *DROSHA*, *MSH4*, *RERE*, *ROS1*, *TACC2*, and *TYRO3*. These high reliability SNVs are shown in the supplementary results [Table S 16 (MSS) and Table S 17 (MSI)]. In contrast, the validation of eight (MSI tumor) and seven (MSS tumor) SNVs, which were exclusively called in the WES results of both tumor samples but not supported by the WGS data, failed (Table S 18 and Table S 19).

3.1.3.5 Detection of large structural variants and large InDels

To find somatic large insertions, all unmapped reads in good quality (≥ 80 bases with QS ≥ 20) were de novo assembled for both tumor samples. In the first patient (MSI tumor), 14,639 contigs were formed, of which 367 had at least one putative insert position. The quality filter passed 75 contigs with on average 1.8 possible insert positions and a maximum length of 81 nucleotides. Eight of these variants with an insert length between 5 and 31 bp were somatic. One somatic insert position was located in the intronic region of the UCSC gene *uc021thc.2*, which encodes for parts of antibodies. The same insertion position was also close to the transcription start of the gene *X69637* (848 bp distance). Another insertion position was within the intronic region of the gene *MGAM*, which encodes the maltase-glucoamylase protein. The protein plays a role in the digestion of starch. All other putative insertion positions were located in intergenic regions. All putative somatic insertion positions with distance smaller 100 KB to the closest gene are reported in the supplementary Table S 20. In patient 2 (MSS), all unmapped high quality reads were assembled to 252 contigs. For six of these contigs, at least one putative insert site existed, but none of these passed the quality filter.

In both tumor samples, more germline and somatic inversions were observed than in the matching controls. In the MSI tumor sample, 3,370 germline and somatic were detected, while in the MSS tumor, 6,125 germline or somatic variants were called. Out of these variants, 71 (48 non-overlapping) and 201 (173 non-overlapping) somatic inversions passed the quality filter in MSI and MSS tumor, respectively. In contrast, 2,541 and 1,036 germline and somatic variants were detected in the matching control samples of the first and second patient, respectively, of which 33 (21 non-overlapping) / 47 (35 non-overlapping) met the quality criteria. In the first patient (MSI tumor), this resulted in 22 (17 non-overlapping) somatic inversions affecting the following eight genes: *AHCYCL2*, *CCDC88C, CCDC102B, DCK* (intronic), *FAM40B*, *LOC440434, LOC642236*, and *SPINK14* (Table S 21). In the MSS tumor, 95 (89 non-overlapping) somatic inversions in 30 genes were identified, of which eight were at least partially located in exonic regions of the following genes: *WARS2*, *NASP*, *AKR1A1*, *RIMS1*, *EPHA5*, *GRID1*, *KCNMA1*, and *SPINK14* (Table S 22). Thus, in both patients a somatic inversion affected the gene *SPINK14*.

Also an increased number of interchromosomal germline and somatic translocations was found in the tumor samples. In total, 29,068 (60 filtered in 37 regions) germline and somatic interchromosomal translocations were detected in the MSI tumor sample and 1,513 (26 / 24) germline interchromosomal translocations in the matching control of the first patient. Out of these, 27 breakpoint pairs in 18 regions with the following 14 affected genes were somatic: *ABCB1*, *ACOT13*, *ANKRD30BL*, *BAZ2A, C10orf54*, *CACNA2D1*, *CDH23*, *CENPH*, *CTNNA3*, *DCAF6*, *PDE4D*, *PHACTR2*, *SENP5*, and *SF3A* (Table S 23). The MSS tumor sample of the second patient harbored 72,444 (137 / 74) germline and somatic interchromosomal

translocations, while in the corresponding non-tumor sample, 12,162 (61 / 45) variants were called. 63 breakpoint-pairs in 42 regions with the following 14 affected genes were somatic (Table S 24): *CDC27*, *CMYA5*, *CNTNAP2*, *GUSBP1*, *IMMP2L*, *LOC642236*, *MBOAT1*, *MCC*, *MLL3*, *NDUFA13*, *PCMTD1*, *SLC6A16, SRRM1*, and *TNC*. Several breakpoints indicating interchromosomal translocations were located in the gene *MLL3*.

Next, deletions larger than five base pairs were investigated. The following numbers of germline and somatic variants were detected (all (filtered)): 191,076 (12,173) in the MSI tumor and 92,643 (8,892) in the matching control of the first patient, 122,651 (14,224) in the MSS tumor and 106,420 (8,096) in the corresponding non-tumor sample of the second patient. Out of 1,739 somatic deletions in the MSI tumor, 343 were located in intragenic regions (Table S 25). Interestingly, in the MSI tumor sample, over 97% of the somatic deletions were identified on chromosome five, six or seven. The mean length of the somatic variants was 123 bp (Figure 3-7). In the MSS tumor sample of the second patient, 78 deletions with an average length of 62 bp in 30 gene loci were somatic (Figure 3-7, Table S 26). All above stated intragenic variants were located in intronic regions.

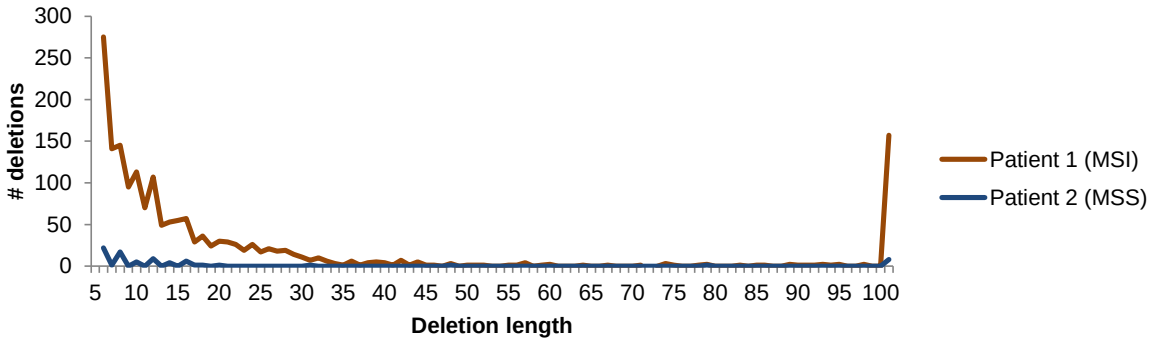
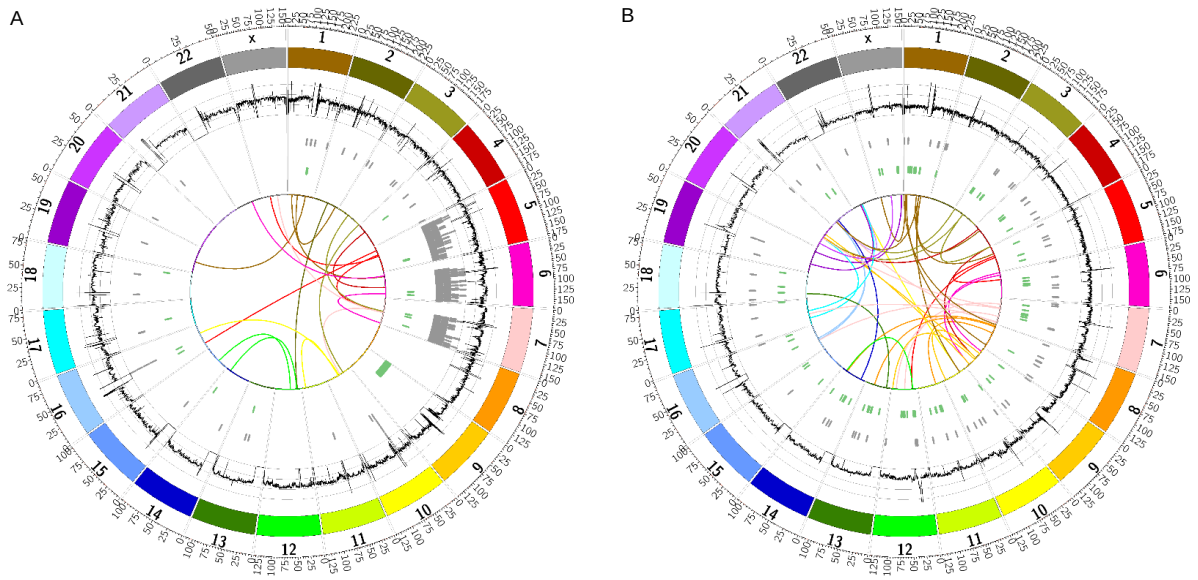


Figure 3-7 Length distribution of somatic large deletions
All deletions with length ≥ 100 are shown with deletion length = 100.

1,652 (72 filtered / 4 non-overlapping filtered) germline and somatic tandem duplications were detected in the MSI tumor, while 1,548 (53 / 2) germline tandem duplications were identified in the matching control of the first patient. In the second patient, 1,199 (69 / 1) germline and somatic tandem duplications were called in the MSS tumor and 1,488 (74 / 3) germline tandem duplications in the corresponding non-tumor sample. All filtered somatic tandem duplications were intergenic.

All quality filtered, somatic inversions, interchromosomal translocations, tandem duplications, and large deletions are shown in Figure 3-8. Due to their function described in the literature the most interesting genes affected by an exonic somatic structural variant were *DCK* (inversion, MSI), *MCC* (translocation, MSS), *MLL3* (translocation, MSS), *PARK2* (deletion, MSI), and *PDE4D* (translocation and deletion, MSI) (Table 3-2).



Patient 1 (MSI tumor) Patient 2 (MSS tumor)

Figure 3-8 Circos plots based on large structural variants
The plots are based on all quality filtered somatic large structural variants of the (A) MSI and (B) MSS tumor samples. The following information were arranged in order from outer to inner rings: genomic position, coverage (black), large deletions (grey), inversions (green), and interchromosomal translocations (center). The coverage histograms are based on a sliding window with window size 150,000 and step size 75,000. The maximum axis limit of the coverage histogram was set to 100.

Gene symbol	Official full name	Description
DCK	Deoxycytidine kinase	• Inactivation of DCK is involved in acquisition of gemcitabine resistance in pancreatic, gastric, colon, and bile duct cancer [189]
MCC	Mutated in colorectal cancers	• Putative colorectal tumor suppressor [190]
MLL3 (*KMT2C*)	Lysine (K)-specific methyltransferase 2C	• Tumor suppressor [191] • Mutations in chromatin remodeling genes (*ARID1A, MLL3,* and *MLL*) occur in 47% of GC [174]
PARK2	Parkin RBR E3 ubiquitin protein ligase	• Recurrent deletions in GC [178] • Tumor suppressor [192]
PDE4D	Phosphodiesterase 4D, cAMP-specific	• Often deleted in GC [178] • Homozygous deletion found in 3.56% of primary solid tumors [193] • Depletion of endogenous PDE4D causes cell death and growth inhibition in multiple types of cancer cells, including GC [193]

Table 3-2 Description of functional interesting genes harboring a somatic large structural variant

3.1.4 Pathways and functional terms potentially associated with gastric cancer

After the analysis on variant and gene level, genomic alterations were investigated on process level. To gain a broader insight into processes that might be associated with GC, a protein-protein interaction network was created for each patient using all genes affected by cross platform somatic exonic SNVs, which were (i) stopgain mutations, (ii) predicted as damaging or (iii) non-synonymous and at a conserved position or in a conserved gene (for definition see sections 2.2.7.1 and 2.2.7.3). Two subnetworks were formed for the MSI tumor of the first patient (Figure 3-9): (i) The larger one was enriched for genes involved in ‘Response to an external stimulus’ ($p = 0.0018$), ‘Cell communication’ ($p = 0.024$), and ‘Positive regulation of metabolic processes’ ($p = 0.038$) (Figure S 13 A). (ii) The smaller subnetwork harbored

genes associated with ‘Chromatin organization’ (p = 0.00026) and ‘Chromatin modification’ (p = 0.00009), ‘Positive regulation of biosynthetic processes’ (p = 0.016), and ‘Gene expression’ (p = 0.012) as well as ‘Cell motility’ (p = 0.0033), and ‘Cell migration’ (p = 0.0022) (Figure S 13 B). No larger network was detected for the MSS tumor of the second patient.

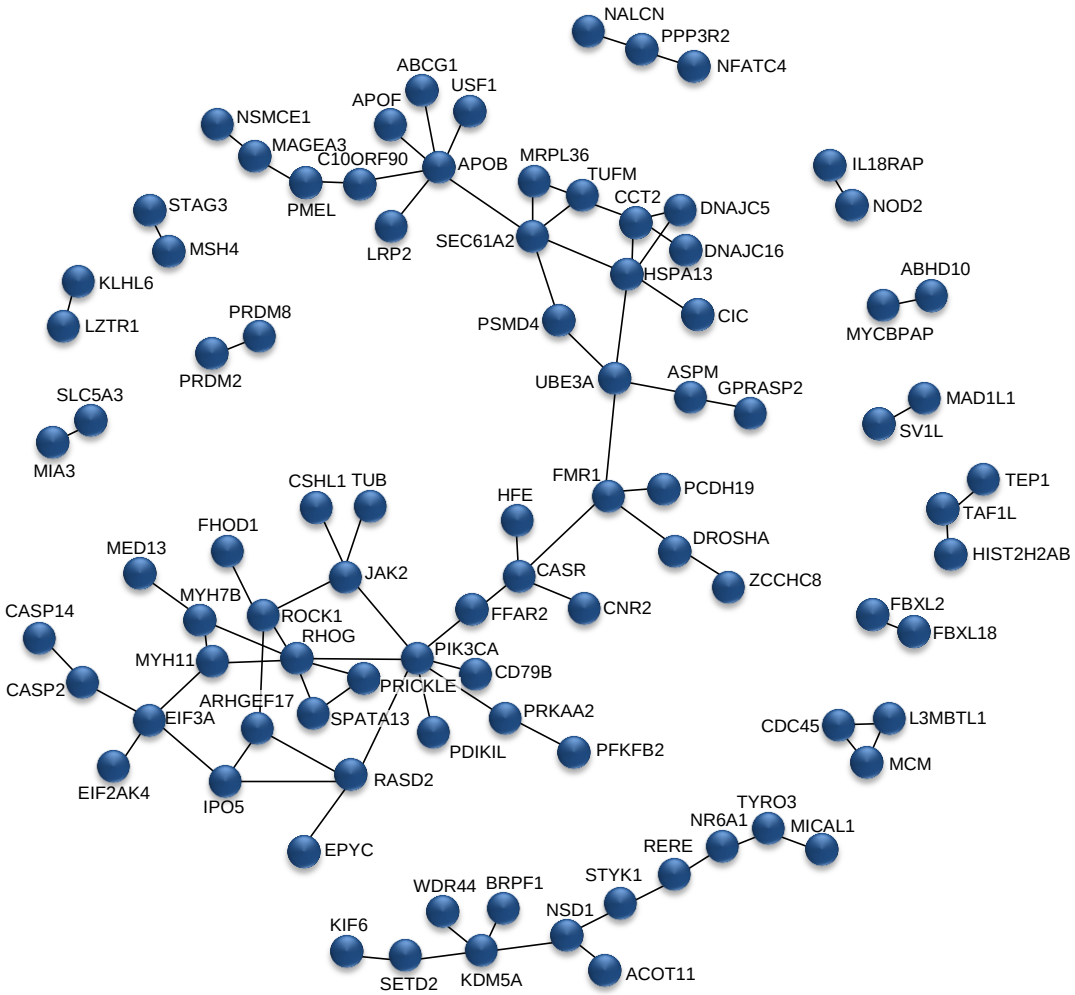


Figure 3-9 Protein-protein interaction network based on genes harboring at least one SNV with predicted effect on the protein function in the MSI tumor sample
The protein-protein interaction network is based on all proteins encoded by genes with at least one somatic SNV, which was non-synonymous as well as at a conserved position or in a conserved gene, a stopgain mutation or predicted as damaging on the protein function. Proteins without connection to any other affected protein are not displayed. For a better clarity, all protein connection types are displayed with a black line.

In a more relaxed approach, a protein-protein interaction network based on all genes affected by at least one cross platform non-synonymous SNV was created. For patient 1 (MSI), one cluster around the receptors *NR4A1*, *RARA,* and *RORC* as well as the tumor suppressor *TP53* was formed (Figure S 14). In the second patient 2 (MSS), one large cluster was observed besides some smaller subnetworks. This larger subnetwork was enriched for genes involved in the KEGG pathways 'Cell cycle' (p = 0.011), 'JAK-STAT signaling pathway' (p = 0.027), and 'B cell receptor signaling pathway' (p = 0.032). The subnetwork was further divided into two parts. In the upper cluster, the genes were mainly involved in the regulation of 'Cell division'

(GO:0051302, p = 0.000062), while the phosphoinositide-3-kinase gene *PKI3CA* had a central position in the second subnetwork (Figure S 15).

Based on the hypothesis that (rare) germline and somatic mutations together would contribute to malignant transformation, processes affected by a high load of germline and somatic variants were investigated. In this approach, the number of called variants in the tumor samples were corrected for the overall germline spectrum of such variants in all 1092 individuals of the 1000 Genomes project. The analyses were based on two different categories: (i) number of genes affected by at least one SNV with predicted effect on the protein function and (ii) number of SNVs predicted as damaging on the protein function. In the 1000 Genomes Project, GATK was applied for variant detection. To guarantee the comparability, exclusively variants called with GATK in the tumor samples were used for the comparison between the tumor samples and the results of the 1000 Genomes Project.

In the MSI tumor of the first patient, the SNV / gene count in 40 / 39 pathways was at least 1.5 times higher than the mean value of all individuals of the 1000 Genomes Project (Figure S 16). With the same method 18 / 22 KEGG terms were found in the MSS tumor of the second patient (Figure S 17). In both tumor samples, several cancer and metabolic pathways were affected. Further functional interesting candidates included the 'p53 signaling pathway', the 'ErbB signaling pathway', the terms 'Cell cycle' and 'Gastric acid secretion' (MSI, patient 1) as well as the 'mTor signaling pathway' (MSS, patient 2). The SNV / gene count of the investigated tumor samples was for all pathways lower than the maximum value of the individuals from the 1000 Genomes Project.

The same strategy as used for the KEGG pathway analysis was applied to GO terms. In 33 ontology groups, the MSI tumor sample of the first patient harbored more SNVs with predicted influence on the protein function than all individuals sequenced in the 1000 Genomes Project (Figure S 18 A). Moreover, 19 terms existed with more genes affected by an SNV with damaging prediction in the tumor sample than the maximum number in the 1000 Genomes Project (Figure S 18 B). In MSS tumor sample of the second patient, 30 GO terms on SNV level and 13 GO terms on gene level were found, which were more often affected in the tumor sample than in all individuals out of the 1000 Genomes Project (Figure S 19). GO terms consisting of exactly the same genes were merged and GO terms harboring only one gene excluded. Nine GO terms were shared between the first (MSI tumor) and the second (MSS tumor) patient including the functional interesting terms ‘Nuclear migration along microfilament’, ‘Negative regulation of transposition’, ‘Negative regulation of viral reproduction’, and ‘DNA cytosine deamination’ (Figure S 18, Figure S 19).

3.2 Investigation of murine inflammation-associated colorectal cancer

3.2.1 Phenotype caused by AOM/DSS treatment

A lower body weight and a higher DAI were observed in AOM/DSS-treated mice compared to DSS-treated and control mice. The severity of induced colorectal inflammation was dependent on the DSS concentration and treatment duration. The strongest effect (body weight loss, DAI increase, tumor induction) was noticed in the first set (highest DSS dose) (Figure 3-10). Over the course of the experiment the biggest DAI differences (including body weight, rectal bleeding, and reduced stool consistency) were detected shortly after the first DSS administration (rectal bleeding, reduced stool consistency) and at the end of the experiment (elevated rectal bleeding) for all AOM/DSS- and DSS-treated mice.

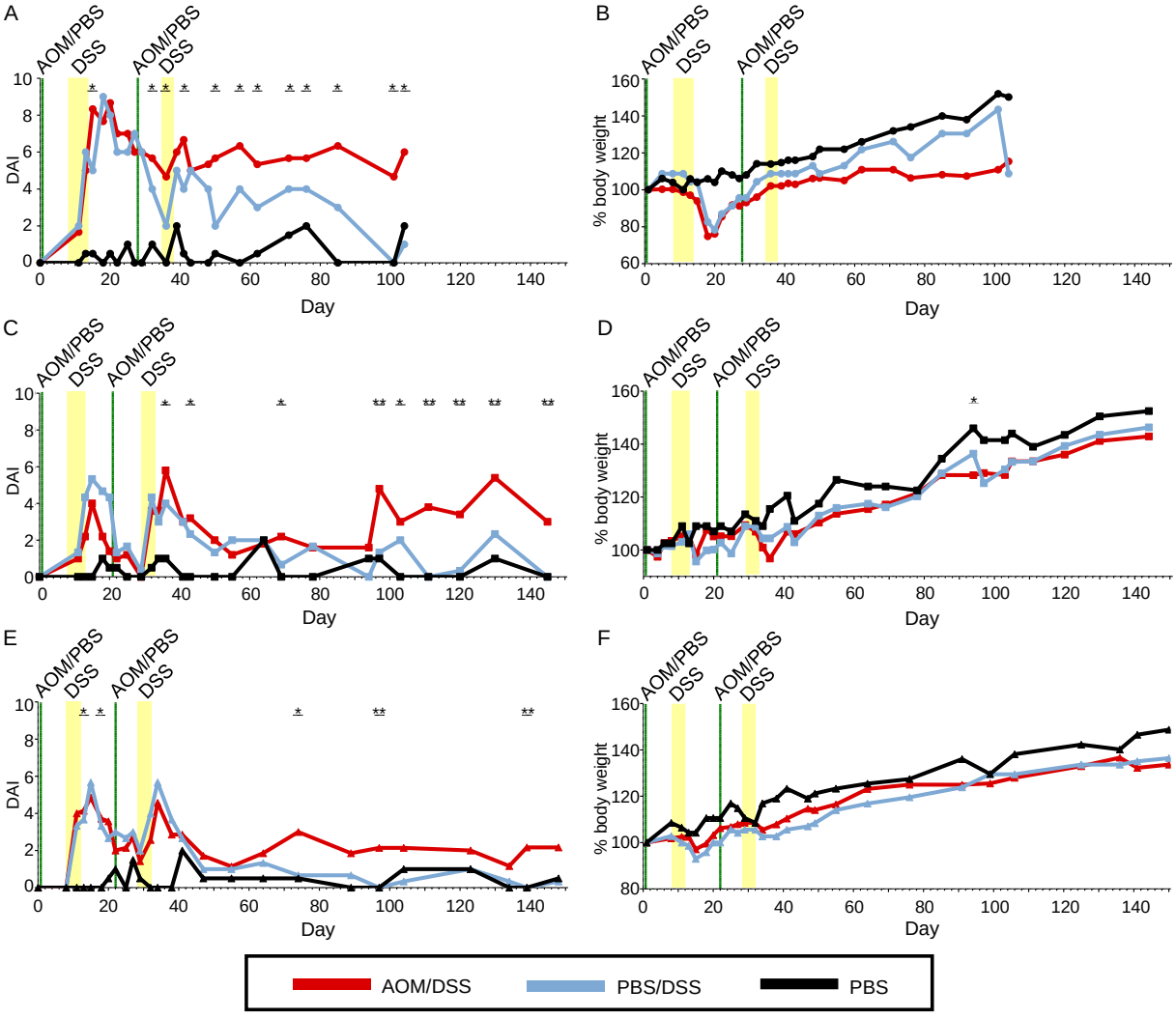


Figure 3-10 Disease activity index and body weight distributions during the course of the experiment
The vertical bars indicate days with AOM or PBS injection, respectively. The AOM/DSS- and DSS-treated mice received DSS in the drinking water on days marked with yellow background. The asterisks label significant differences between the AOM/DSS-treated mice and the union of DSS-treated and control mice without correction for multiple testing. Due to the low power, DSS-treated mice and controls were merged as one group. (A) DAI, treatment set 1 (high DSS dose). (B) Body weight, treatment set 1 (high DSS dose). (C) DAI, treatment set 2 (medium DSS dose). (D) Body weight, treatment set 2 (medium DSS dose). (E) DAI, treatment set 3 (low DSS dose). (F) Body weight, treatment set 3 (low DSS dose).

The number of tumors and the average tumor sizes shown in Figure 3-11 and observed by colonoscopy (Figure S 20) were larger in AOM/DSS-treated mice from the first set (high DSS dose) than in the second (medium DSS dose) and third set (low DSS dose) (Figure 3-12 A). In the mice C11R, C15WO, and C16R of the third set only dysplasia were noticed. The tumor burden was higher in the distal parts of the colon. No significant differences in colon length existed between the treatment groups (Figure 3-12 B).

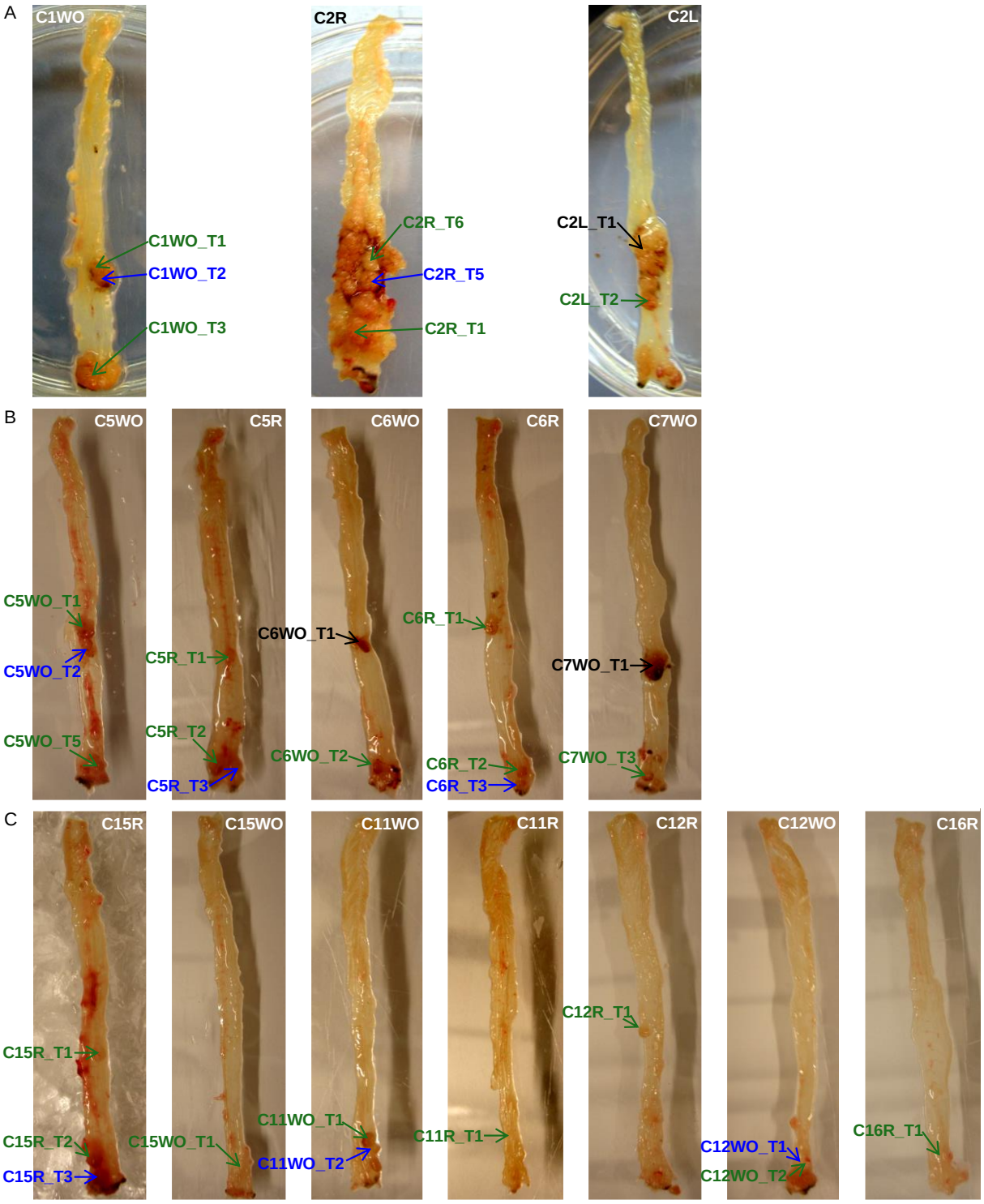


Figure 3-11 Longitudinally opened colons from AOM/DSS-treated mice
Tumor samples used for WES are marked in green and those for transcriptome sequencing in blue. For all samples labeled in black, the whole exome and the transcriptome were sequenced. (A) AOM/DSS treatment set 1 (high DSS dose). (B) AOM/DSS treatment set 2 (medium DSS dose). (C) AOM/DSS treatment set 3 (low DSS dose).

The AOM/DSS-treated mice C1WO (first set) and C15R (third set) were affected by an intestinal prolapse. Therefore, organ removal from the animal C15R was performed 19 days before the official end of the experiment.

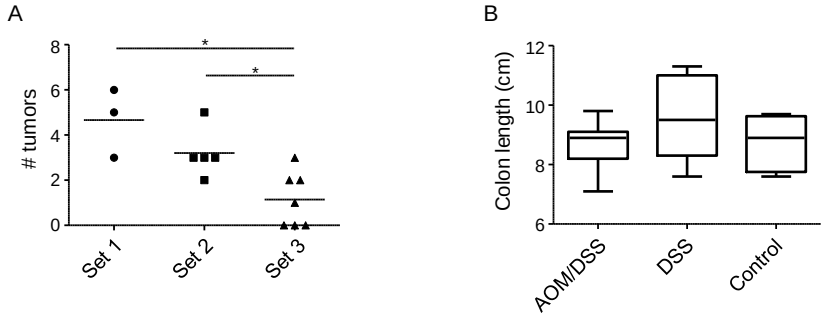


Figure 3-12 Tumor counts and colon length distribution
(A) Comparison of the number of tumors between the three different DSS dose sets of AOM/DSS-treated mice. (B) Colon length distribution of the three different treatment groups.

The performed histological analyses (Figure S 21) demonstrated that in the untreated animals, crypts were regularly distributed throughout the mucosa, formed straight tubular structures, and were arranged parallel to one another. An elongation and branching of the crypts was observed in the colon of DSS-treated mice. In contrast, in mice treated with AOM/DSS, the tumor development was associated with disruption of the crypt architecture, which was completely lost in the first treatment set (high DSS dose). Neoplasia with mucosal dysplasia and crypt atrophy was characterized by loss of columnar orientation, elongation, branching, infolding and crypt abscesses (Figure S 21), and elevated proliferation of neoplastic cells as visualized by Ki67 staining (not shown). Furthermore, the number of mucosal and submucosal infiltrating immune cells increased with rising DSS concentration and treatment duration. In addition, slight enlargement of hyperchromasia was visible in the AOM/DSS-treated mice.

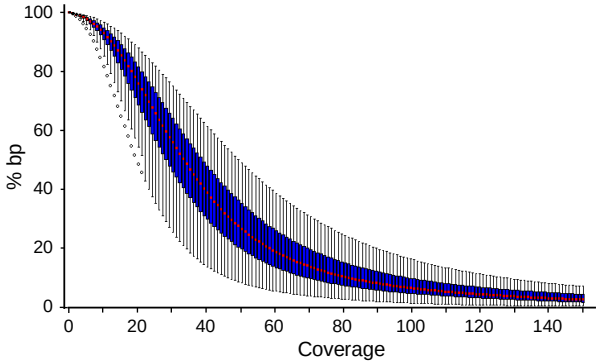
3.2.2 Exome sequencing

3.2.2.1 Sequencing results

WES was performed on 52 samples, of which 24 were from tumors, 15 from the corresponding proximal colon part, seven from DSS-treated mice, and six from untreated controls. For nine mice, two tumor samples each were sequenced, while for six AOM/DSS-treated mice from the third set (low DSS), only one tumor was investigated. Between 82 and 146 million reads were obtained per sample using the Illumina HiSeq2500 system. After applying the quality filter described in section 2.2.3.2, between 75 and 134 million reads remained with an average base quality of 36.3 and an average length of 99.1 nt. Between 73 and 133 million reads (~98.8%) of these reads could be mapped to the reference genome with a median mapping quality of 51. The average insert size of the proper read pairs ranged from 104 to 172 bp. Out of the mapped sequences, ~15% had a shared starting point. Around 28% of the total mapped bases

were located in the enriched target regions, which resulted in a mean coverage of 47x (Figure 3-13). No differences were observed between the samples groups (Table S 27).

A



B

Category	Mean
Coverage	46.9
# raw reads	114,299,866
# filtered	102,600,362
# mapped	101,352,157
% onTarget	28.5
% duplicates	14.7

Figure 3-13 Sequencing and mapping results
(A) Coverage distribution for all samples. For each coverage value, the median coverage is marked in red, the interquartile range in blue, and the 90% interval values in black. (B) Sequencing and mapping statistics.

Subsequently, the relative contributions of technical bias and biological background to the distribution of called SNVs and genes with detected variants were investigated (Figure 3-14). Most of the detected base substitutions seemed to be robust against batch effects during library preparation and sequencing, but also the treatment had just a small influence on variant occurrences at specific positions. This indicates that neither AOM nor DSS caused SNVs at recurrent positions. Base substitutions as consequence of the treatment were more likely to be randomly distributed, while some of the SNVs were responsible for the development of tumors. A surprisingly high percentage of genes mutated by at least one alteration was explained by library batch and sequencing run. However, the strong technical bias had no effect on the reported analyses, because samples from all treatment groups were randomly distributed on all library batches and sequencing runs.

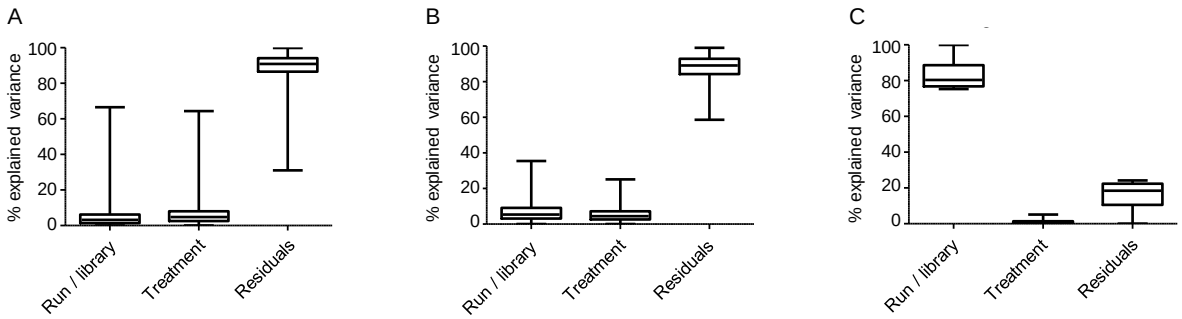


Figure 3-14 Influence of technical bias and treatment on distribution of SNV calls
Boxplots show distribution of the percentage of explained variance across all genes and all SNVs. Exclusively SNVs / affected genes existing in more than three samples were considered. Explained variance was based on (A) all called SNVs, (B) number of called SNVs per gene, and (C) genes affected by at least one called SNV.

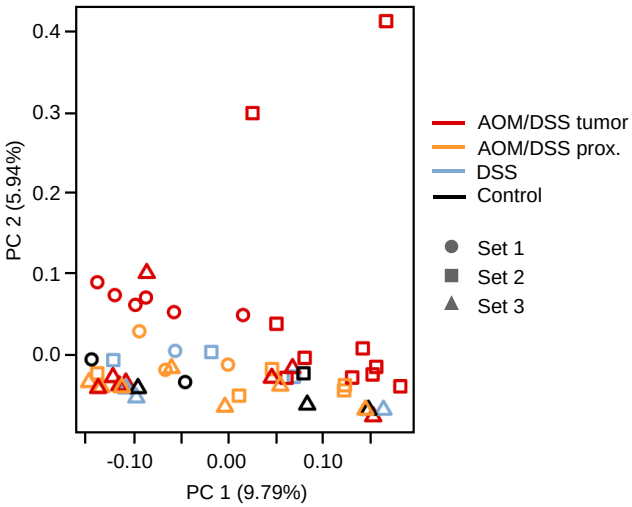
3.2.2.2 Test for sample differences and intertumoral diversity

To investigate whether general variant differences are detectable between the sample groups, a PCoA was performed using a Jaccard distance matrix [194] based on all exonic SNVs and

InDels (Figure 3-15 A). Tumor samples from the first (high DSS) and second (medium DSS) experiment were clearly distinguishable from the control samples (p = 3.186E-10 and p = 0.00132, respectively), while no differences were visible between the control samples and set 3 tumor (low DSS, p = 0.309) or post-inflamed samples (p = 0.855). The same tendencies were also observed in a PCoA only based on non-synonymous SNVs and InDels, while the separation was weaker on gene level (Figure S 22).

Next, it was investigated whether tumors from the same mouse shared more somatic variants than tumors from different mice (Figure S 23). Therefore, for nine AOM/DSS-treated mice, the whole exomes of two tumors were sequenced. The mutational distances based on Jaccard indices between tumors of the same mouse were compared with the median distance to all tumors. Interestingly, in the first and second treatment sets (high and medium DSS), the distances between tumors received from the same mouse were similar to the distances between tumors obtained from different mice. This indicates an independent occurrence of multiple tumor-initiating events in each mouse. In contrast, for mice treated with a low DSS dose (set 3), distances between mutational spectrums of tumors from the same mouse were smaller than those between tumors from different mice. This demonstrates that animal- or colon-specific variants superimpose tumor-specific alterations at an early tumor stage. These results were observed on three different data groups: (i) exonic variants, (ii) genes affected by a mutation, and (iii) number of mutations per gene (Figure 3-15 B, Figure S 24).

A



B

Set	Sample	Jaccard distance between	
		Tumors out of same mouse	All tumors
1	C1WO_T1	0.545	0.553
	C1WO_T3	0.545	0.545
	C2L_T1	0.570	0.550
	C2L_T2	0.570	0.554
	C2R_T1	0.570	0.567
	C2R_T6	0.570	0.570
2	C5R_T1	0.635	0.475
	C5R_T2	0.635	0.658
	C5WO_T1	0.634	0.518
	C5WO_T5	0.634	0.632
	C6R_T1	0.554	0.523
	C6R_T2	0.554	0.545
	C6WO_T1	0.442	0.470
	C6WO_T2	0.442	0.484
	C7WO_T1	0.456	0.508
	C7WO_T3	0.456	0.485
3	C15R_T1	0.444	0.454
	C15R_T2	0.444	0.518

Figure 3-15 Sample distances based on all exonic variants

(A) PCoA based on Jaccard distance matrix using all called exonic SNVs and small InDels. (B) Pairwise comparison of Jaccard distances between tumors from the same mouse (third column) and the median distances to all tumors (fourth column). Exclusively samples from mice with two sequenced tumors were considered. Black shaded cells indicate the larger distance within each row. Equal distances are displayed in grey. (A) and (B) The sets are based on the DSS treatment conditions: set 1 = high DSS dose, set 2 = medium DSS dose, set 3 = low DSS dose.

3.2.2.3 SNV patterns

3.2.2.3.1 SNV type distribution

Variants can arise spontaneously or due to environmental influences [195]. In particular, each mutagen and each tumor type is characterized by a specific variant signature, which can influence the formation of other alterations and the outcome of a disease [54, 112]. To reveal variant patterns associated with evolutionary mechanisms, the SNV spectrum of the mouse strain C57BL/6N (used in the experiments of the current study) was compared with the one of the mouse strain C57BL/6J (used as reference) (supplementary chapter 6.2.2.4). The somatic SNV distributions of the investigated tumor samples from AOM/DSS-treated mice were analyzed and compared with the SNV differences between the two mouse strains to gain insights into variant patterns, which are related to murine AOM/DSS-triggered colorectal cancer (see next paragraphs).

In total, between 35,000 and 115,000 germline or somatic SNVs were called per sample with at least one calling algorithm (Samtools, GATK or diBayes), whereof ~15% were detected with all three callers (Figure S 27). Although exonic regions were enriched during the library preparation for sequencing, over 50% of all SNVs were intergenic, ~30% intronic, and only less than 5% in known annotated exonic regions (Figure S 28). The number of called somatic exonic SNVs was significantly higher in the tumor samples compared to all other sample groups (Figure 3-16 A, p(tumor vs. AOM_DSS_proximal) = 0.0001, p(tumor vs. DSS) = 0.0075, p(tumor vs. control) = 0.003). Interestingly, the number of exonic somatic SNVs varied even significantly between all DSS dose sets of the tumor samples (Figure S 29 A+B p(set1 vs. set2) = 0.0462, p(set1 vs. set3) = 0.0059, p(set2 vs. set3) = 0.0217). In contrast, no significant differences were observed in the total SNV counts including the non-exonic SNVs (Figure S 30 A). In the comparison between the tumor sets, the total somatic SNV count was increased in the samples of the first set (high DSS dose), but no differences between set 2 and 3 (medium and low DSS dose) were observed.

In contrast to single SNVs, no significant differences (adjacent base substitutions) of the total amount of double SNVs were observed between the treatment groups. However, for eight substitution types, the comparison of the relative count between tumor and at least one other sample type resulted in a significant p-value (Figure S 36). Of special interest were the variants AC>CA (GT>TG), AG>GA (CT>TC), GA>AT (TC>AT), and TA>GC, because these alterations harbored the highest distance between tumors and controls, while the values of the post-inflamed samples were between these two.

Next, the relative contribution of the six SNV classes (C>A, C>G, C>T, T>A, T>C, T>G) was analyzed to decipher processes involved in the development of murine AOM/DSS-triggered

colorectal cancer (Figure 3-16 B). In the set of all somatic exonic SNVs, the relative amount of C>T base substitutions was significantly increased in the tumor samples compared to all other analyzed sample groups. This difference was especially high in tumor samples of the first and second set (high and medium DSS) (Figure S 29 C+D). In addition, a significant decrease of C>A substitutions was observed in the tumor samples for all pairwise comparisons. Thereby, the percentage of C>A SNVs decreased with increasing C>T count (Figure S 31). Moreover, somatic exonic T>C substitutions differed significantly between tumor and proximal AOM/DSS samples ($p = 0.0013$). In the set of all somatic SNVs including the non-exonic variants, an increased C>G count in the formerly inflamed tissue samples (AOM/DSS proximal and DSS) were found besides the differences in the C>T substitutions (Figure S 30 B-D, Table S 28).

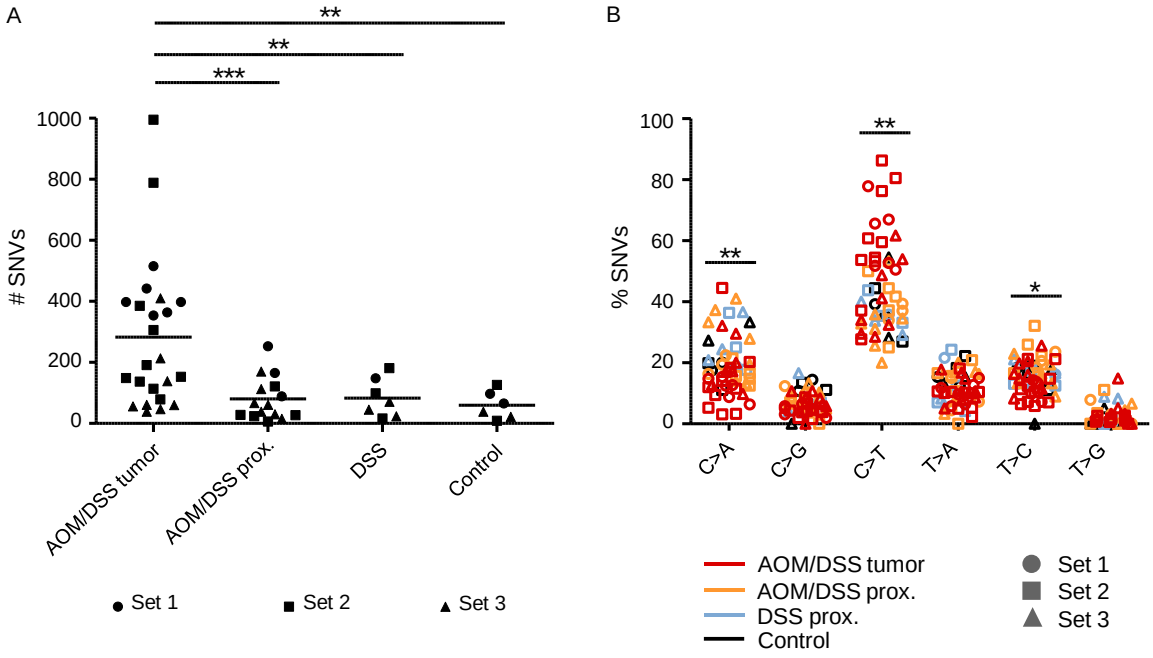


Figure 3-16 Basic SNV patterns of somatic exonic SNVs
(A) Number of somatic exonic SNVs. (B) SNV type distribution. Statistical significance was tested with a Kruskal-Wallis test. (A) + (B) The sets refer to the DSS treatment conditions.

Besides the basic SNV type distribution, the sequence context of the base substitutions was analyzed by considering the 5' and 3' flanking bases of all somatic SNVs. Thereby, the observed number of each trinucleotide was normalized according to the prevalence of each trinucleotide in the genome. In all non-tumor tissues of the first sample set, the C>T substitutions were overrepresented at CpG sites (Figure 3-17 A). Furthermore, C>A variants were dominant in a TCG trinucleotide combination. These variant spectrums reflect the SNV pattern caused by evolutionary processes. In contrast, the tumor samples harbored a more equal sequence context distribution for both SNV types. An unsupervised hierarchical clustering based on all trinucleotide counts confirmed the distinct signatures between tumor and control as well as post-inflamed tissue samples in the first treatment set (Figure 3-17 B). The same analyses were also performed with the samples from all three treatment sets

(Figure S 32 + Figure S 33). However, due to a strong technical bias, no clustering was detectable for the second and third set (low and medium DSS).

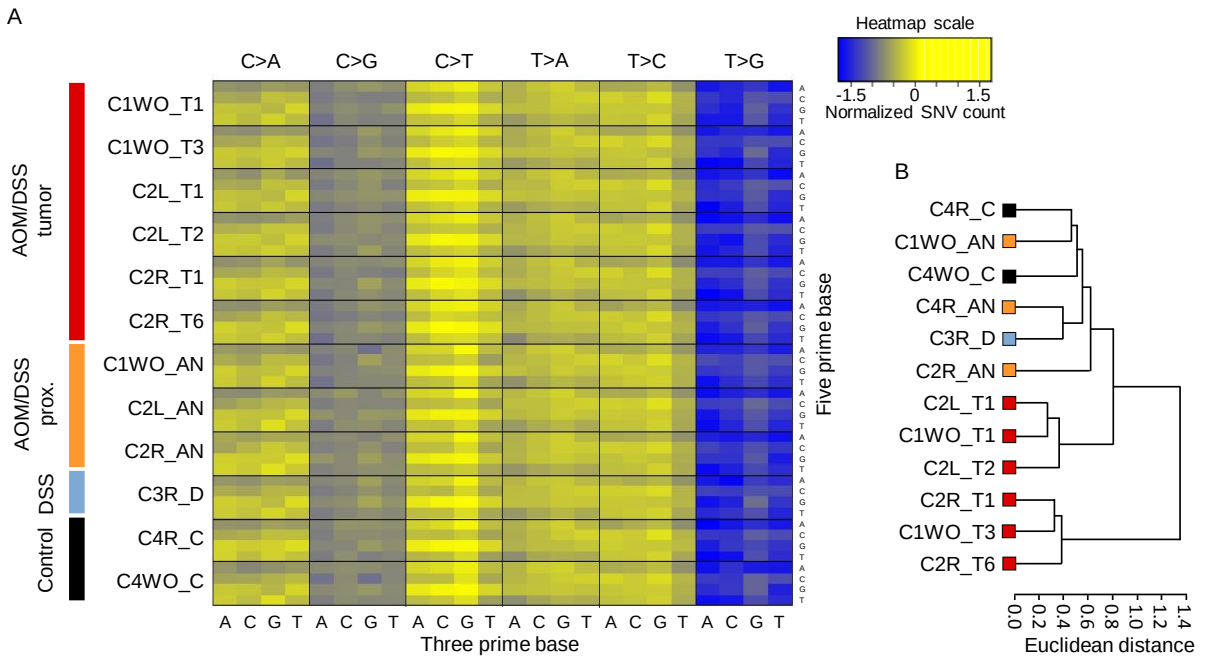


Figure 3-17 Distribution of somatic SNVs including SNV context
(A) Heatmap showing the log-transformed normalized counts of each SNV type including the base context. The 5' base is shown on the vertical axis and the 3' base on the horizontal axis. (B) Unsupervised, hierarchical clustering (Euclidean distance, ward method) of samples from the first treatment set (high DSS dose) based on SNVs including sequence context.

Genes, which are important for fitness or which are highly expressed, are often encoded on the Crick strand to avoid conflicts between replication and transcriptional machineries. In contrast, genes with a higher evolution rate are located on the Watson strand [196–198]. Also in my studies, SNVs (germline and somatic) were significantly more frequently located in genes on the forward (Watson) than on the reverse (Crick) DNA strand (p = 1.372E-07 for the group of all non-tumor samples and p = 3.56E-05 for tumor samples, respectively, Figure 3-18 A-D). In contrast, in the tumor samples, strand differences did neither exist for all somatic SNVs (Figure 3-18 E) nor for the subset of somatic C>T base substitutions, which represented the biggest somatic SNV type group in the tumor samples (Figure 3-18 F). This demonstrated a random strand distribution of somatic mutations.

Besides an unequal distribution of SNVs between Watson and Crick strand, a significant different mutation frequency was observed between coding and noncoding strand for all germline substitution types except for T>G variants, where just a tendency was visible (p = 0.0878) (Figure S 34). In contrast to this evolutionary driven SNV pattern, no significant strand preferences were observed for somatic SNVs in the tumor samples (Figure S 35).

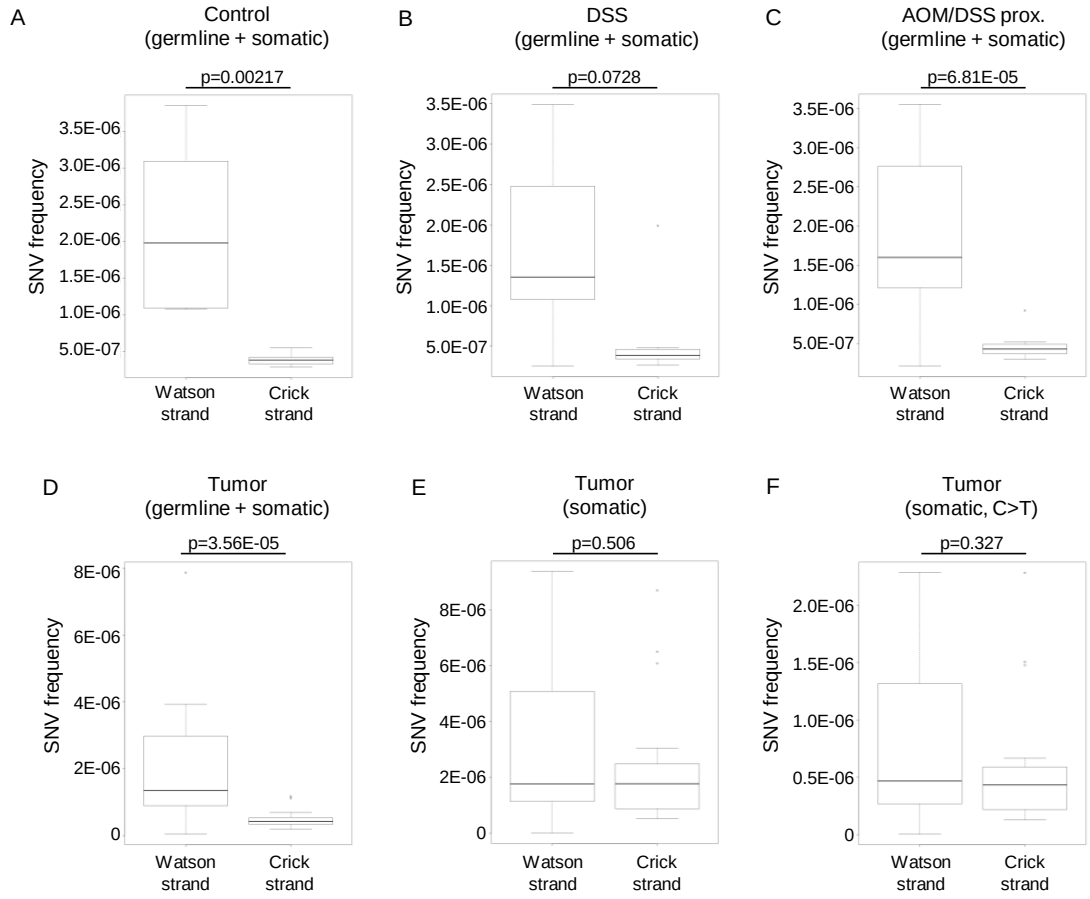


Figure 3-18 SNV strand bias test
Comparison between SNVs in genes on the Crick strand and SNVs in genes on the Watson strand. The SNV frequency was defined as [100,000 x (<# SNVs in genes on strand 'X'> / <# bases in genes on strand 'X'>] with 'X' = Crick or Watson. Comparisons were based on (A) all SNVs called in control samples, (B) all SNV called in samples from DSS-treated mice, (C) all SNVs called in proximal samples from AOM/DSS-treated mice, (D) all SNVs called in tumor samples, (E) somatic SNVs called in tumor samples, and (F) somatic C>T SNVs called in tumor samples.

3.2.2.3.2 Test for regional clustering of SNVs

Next, it was tested whether variants in AOM/DSS-induced tumors have a preference to cluster in specific regions, because some tumor types form regions with high SNV density, the so-called *kataegis* effect [153]. Occurrences of hypervariable sections (DNA regions with high number of variants) in the exonic regions were investigated with an adapted rainfall plot method, in which the distances between each somatic SNV and the adjacent somatic substitution were displayed (Figure 3-19). Here, small clusters in the tumor samples as well as in the controls were observed. In order to investigate these clusters in more detail, a sliding window approach with different window sizes and different thresholds for the initial SNV number was used. Using 50 SNVs within a 5000 bp range, two clusters were detected in the tumor sample C5R_T2 (Figure 3-19 B). These clusters were located close together on chromosome 7 [chr7:38189798-38194760 (51 SNVs) and chr7:38191088-38196990 (66 SNVs)]. The gene *1600014C10Rik*, which encodes for a small transmembrane protein with unknown function, was annotated in this region. In none of the other tumor samples, a hypervariable region was detected. Using a smaller window size (160 bp) and a reduced SNV

threshold (8) up to 16 hotspots with a mean length of 73.3 bp and an average of 9.7 SNVs could be detected in all samples. Thereby, the number of hypervariable regions and included SNVs was significantly increased in the tumor samples compared to the controls ($p = 0.029$ and $p = 0.032$, respectively). However, this result arose due to the higher total number of SNVs, as the percentage of SNVs included in hotspots was similar. No significant differences were observed between tumor and post-inflamed tissue samples (Figure S 37). However, it must be emphasized that only the exomes were sequenced and thus, hypervariable regions in intergenic regions, which make up a large parts of the *kataegis* events in other tumor genome analyses, could not be detected.

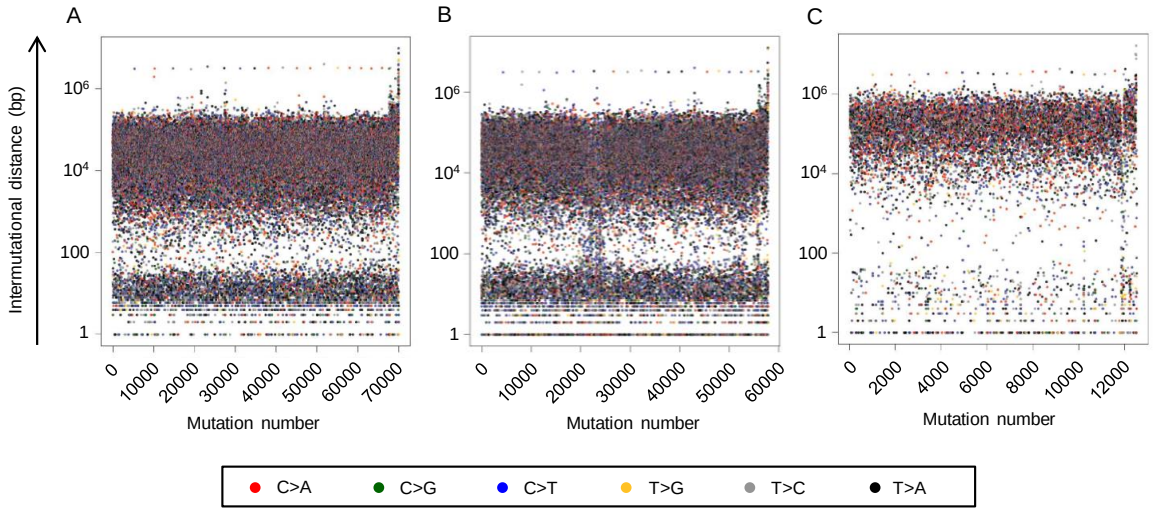


Figure 3-19 Rainfall plots
SNVs are ordered by genomic position on the x-axis. The distance between each SNV and the previous one is plotted on the y-axis. The color of a dot describes the SNV type. (A) Sample C2L_T2 (tumor, set 1), (B) Sample C5R_T2 (tumor, set 2), (C) Sample C20R_C (control, set 3).

In the next step, it was tested whether the number of exonic variants was associated with the expression level of the genes. Based on the FPKM values, all genes were divided into six classes. Thereby, the number of genes per class was similar between all samples (Figure S 38). Afterwards, the mean SNV counts normalized by gene lengths were calculated. Interestingly, in all sample groups, the mutation rate for germline and somatic SNVs was significant higher in lowly expressed genes (Figure 3-20 A) demonstrating an unequal SNV distribution during evolution. Based on a linear regression model, the distribution of variants to the FPKM classes did not differ significantly between the treatment groups. In contrast, no significant dependency between expression level and somatic SNV count were observed in the tumor samples (Figure 3-20 B). This was also the case for somatic C>T variants, which were e.g. induced by AOM (Figure 3-20 C).

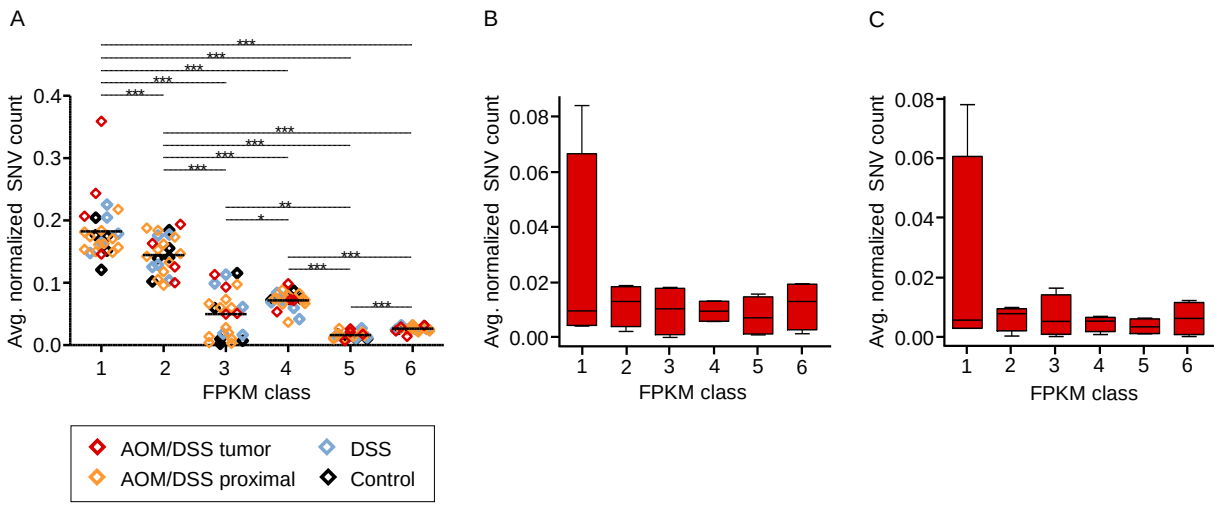


Figure 3-20 Connection between number of SNVs and expression level
The figures show the normalized number of SNVs for each FPKM class. The normalized SNV count was defined as [1000 x (SNV count / total exonic length)]. A higher FPKM class refers to a higher expression level. (A) Analysis based on germline and somatic SNVs of all treatment groups. (B) Analysis based on somatic SNVs called in tumor samples. (C) Analysis based on somatic C>T base substitutions called in tumor samples.

Taken together, SNVs were not enriched in specific genomic regions in the investigated tumor samples. Thus, AOM/DSS-induced colorectal cancer seems to be characterized by a random variant distribution.

3.2.2.3.3 Comparison between observed and known mutational signatures

The observed somatic SNV type distributions including the sequence context were compared with mutational signatures from different human cancer types described in the COSMIC database [122] (Figure 3-21). Interestingly, signatures 6 and 10, which are common in colorectal cancer, showed only a low contribution to the mutational patterns of the investigated tumor samples in the current study. Also signatures often observed in nearly all cancer types such as signatures 1, 2, 5, and 13 had only a small influence on the SNV distributions in AOM/DSS-triggered colorectal tumors, while signatures 3, 8, 12, 20, 22, and 30 had a high impact. Although the etiology of the signatures 8, 12, and 30 were not revealed yet, the signatures might point to molecular mechanisms underlying AOM/DSS-triggered colorectal cancer. As example, signature 20 is likely based on a defective DNA mismatch repair mechanisms, while signature 3 suggests a malfunctioned DNA double-strand break-repair by homologous recombination. Additionally, signature 20 is often related to small InDels in repeat regions and signature 3 indicates an increased number of large InDels at breakpoint junctions. However, WGS would be necessary to verify these variant types.

Based on the signatures described in the COSMIC database, it was not possible to reconstruct the complete observed somatic SNV distributions (Figure 3-22). The RSS values amounted between 0.00087 and 0.00216 (median = 0.00149). Major differences existed e.g. for C>T in the context of ACA as well as T>C in the context of CTG and TTG.

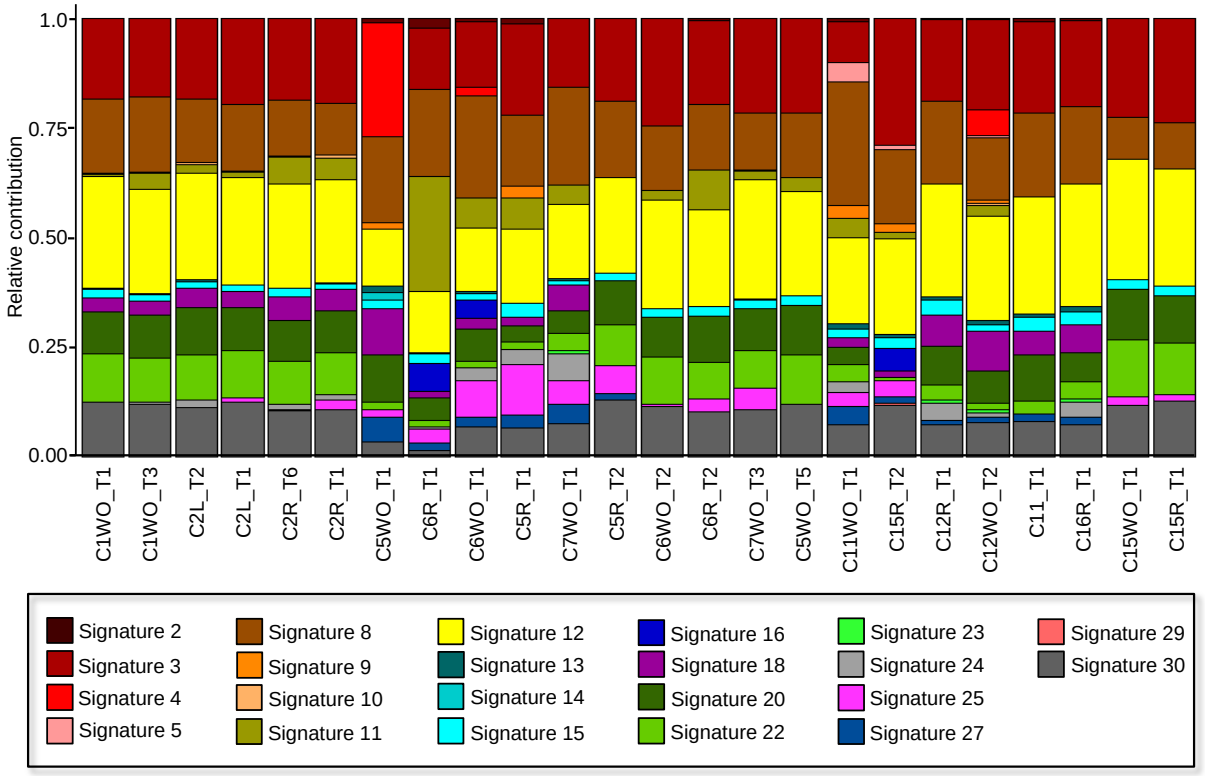


Figure 3-21 Contribution of known cancer-associated mutational signatures described in the COSMIC database to somatic SNV patterns observed in the tumor samples of the current study

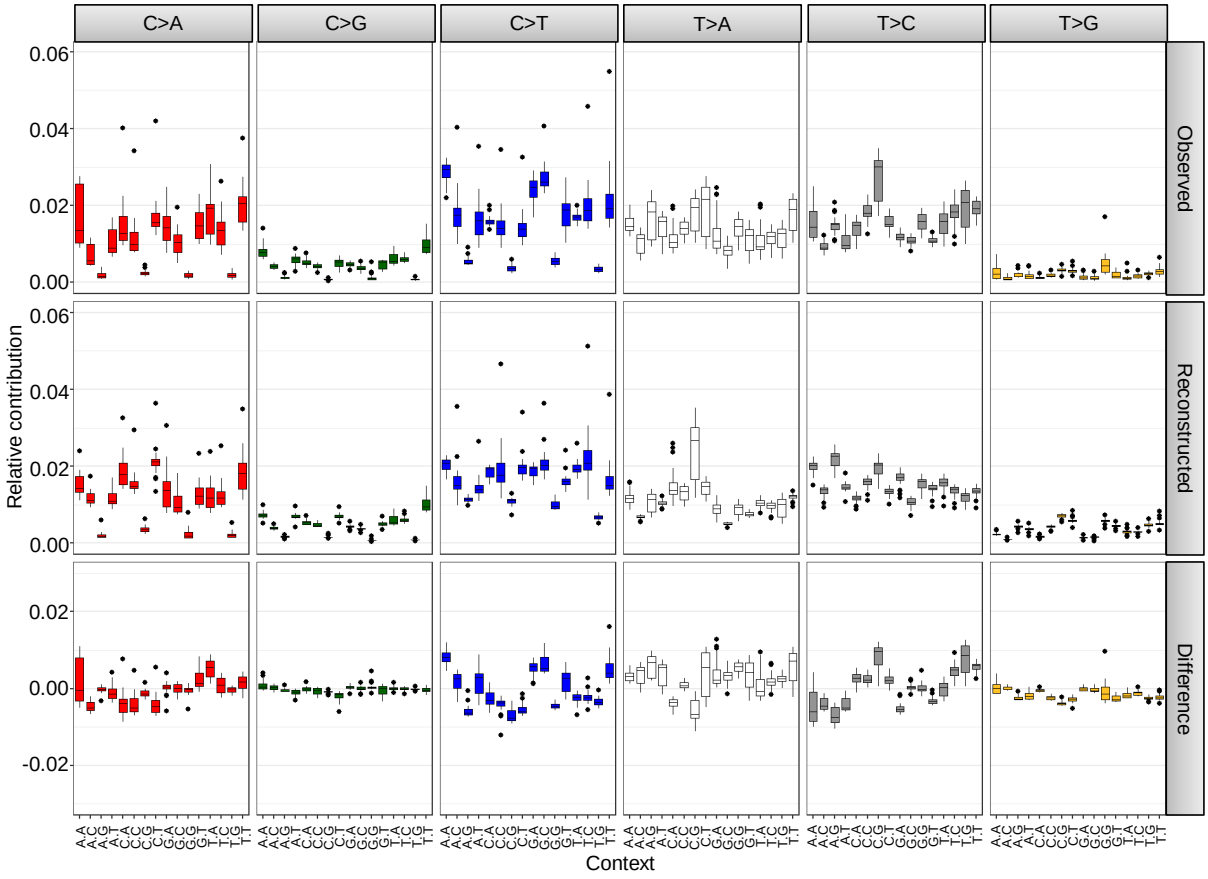


Figure 3-22 Comparison between observed and reconstructed SNV patterns
The observed SNV patterns were based on somatic SNVs of the investigated AOM/DSS-triggered colorectal tumor samples. The reconstructed version was computed using the cancer signatures described in the COSMIC database.

To enable the explanation of the complete observed mutational spectrums, novel signatures contributing to the somatic SNV patterns of the investigated tumor samples were computed. Six signatures were necessary to reduce the RSS value below 0.0001 (RSS values between 1.099E-5 and 5.397E-4 with median = 5.892E-5) explaining 99% of the variance (Figure S 39). Thus, the reconstructed SNV pattern based on the novel signatures represented the observed SNV distributions in a nearly perfect manner (Figure S 40).

The six novel signatures are described in Figure 3-23. Most of the signatures were characterized by a high amount of C>A, C>T, T>A, and T>C base substitutions, while C>G and T>G SNVs were rare. The causal factors for these signatures were completely unknown and further investigations are necessary to identify the underlying mutational processes of the identified signatures. Therefore, it might be that some underlying mechanisms were shared with those of the signatures suggested in the COSMIC database. As consequence, the novel signatures could not be merged with the already known ones, because the reported signatures should be independent of each other.

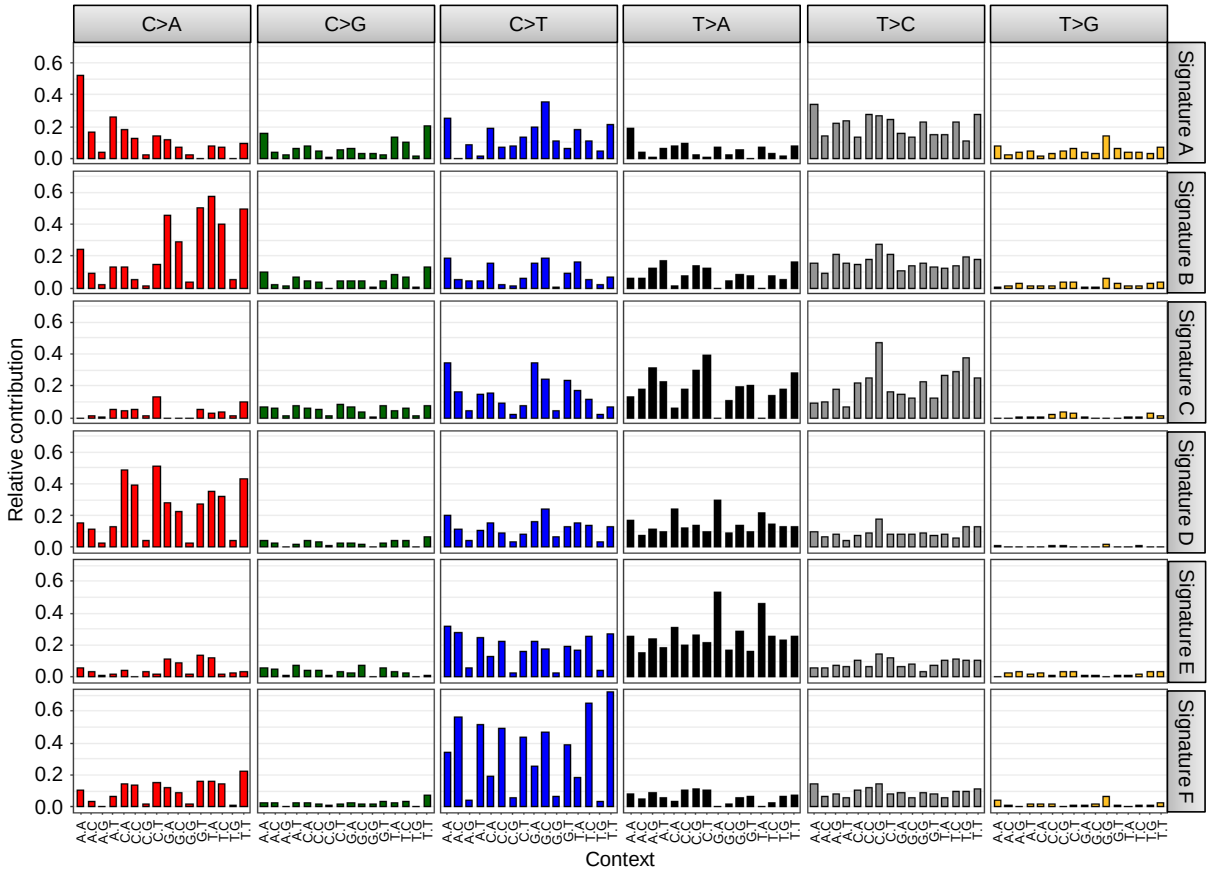


Figure 3-23 Description of novel contributing signatures
The plot describes the relative amount of each SNV type including base context for each of the six signatures underlying murine AOM/DSS-triggered colorectal cancer.

In all tumor samples, signatures D and E had the highest impact on the observed SNV patterns (Figure 3-24). Signatures B, C, and F affected mainly samples from the second treatment set, while signature A had a higher influence on samples from the first and third treatment set.

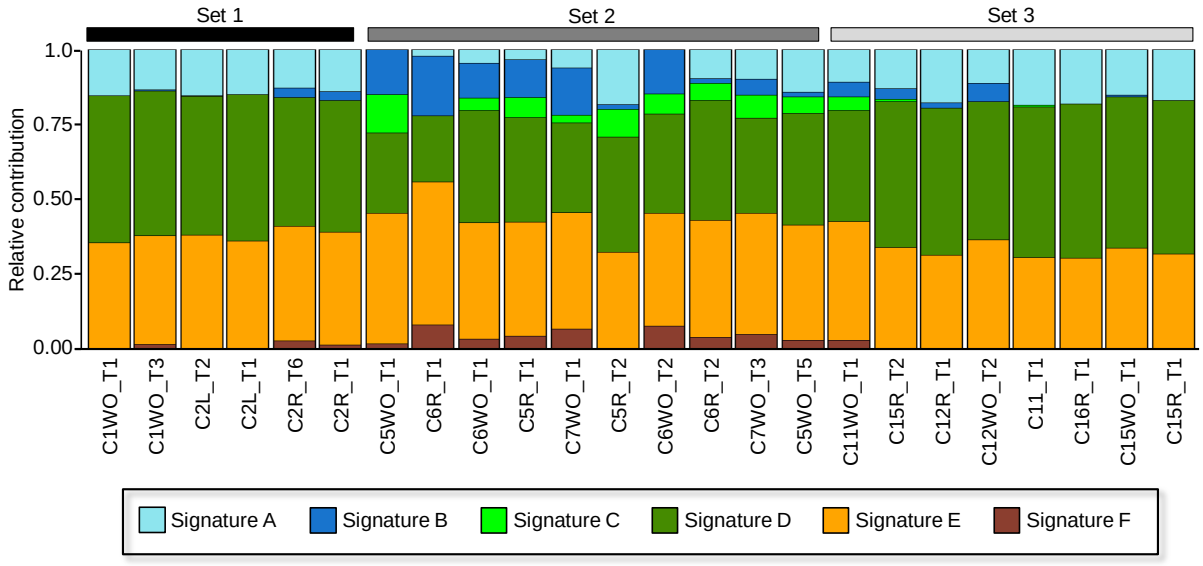


Figure 3-24 Contribution of novel signatures to the observed somatic SNV patterns
The figure shows the relative contribution of each signature to the observed somatic SNV patterns in the tumor samples of the current study.

The need of signatures in addition to those annotated in the COSMIC database demonstrates that the SNV pattern of murine AOM/DSS-triggered colorectal cancer differs from already investigated cancer types including human CRC.

3.2.2.4 Variants detected in AOM/DSS-associated colorectal cancer

3.2.2.4.1 SNVs and small InDels without cancer-associated annotation

In total, 9,063 exonic SNVs were called in at least one tumor sample. Out of the SNVs, which were not listed in dbSNP, 4,296 SNVs caused either a stop codon or were non-synonymous and located at a conserved position. Based on the 5% allele support threshold, 101 SNVs were shared by at least three tumor samples and did not exist in any of the controls. In a further quality control step, all SNVs were extracted, which were (i) at a position with a coverage of at least 5x, (ii) supported by at least two reads, and (iii) part of 20 or more percent of the mapped sequences at that position. This resulted in a final number of 53 mutations (four stopgain, 49 missense) including substitutions in the genes *Mapk10*, *Gnas*, and *Mcc (*Table S 29). In comparison to the non-tumor samples, the number of mutated tumor samples was especially high for the genes *Ctnnb1*, *Loc1000044193*, and *Sbf1*.

To test the functional relevance of all called somatic base substitutions in a more stringent approach, all exonic, non-synonymous SNV positions were converted to the human genome build hg19. This enabled to filter the variants additionally with damaging prediction tools like PolyPhen2 and Sift. For 2,288 variants at conserved positions, which did not exist in the database dbSNP, the program SIFT or the tool PolyPhen2 predicted an effect of the substitution on the protein function. Out of these SNVs, 57 did not exist in any of the untreated controls but were observed in at least three tumors, whereof 32 passed all final quality control

steps. All final variants were included in the results of the first method (Table S 29). Confirmed SNVs were for example located in the genes *Ctnnb1*, *Apobec2*, *Bcl6b*, and *Gfi1*.

33,888 small insertions and 148,197 small deletions (somatic or germline) were called in at least one tumor sample, whereof 1,077 were novel non-frameshift variants. From this set, 17 variants did not exist in any of the controls and were supported by at least 5% of the reads in three or more tumor samples. After the quality steps described in section 2.2.7.2, twelve somatic InDels potentially associated with inflammation-triggered colorectal cancer were identified (Table S 30). None of the affected genes were annotated in the COSMIC cancer gene census list.

The different filter strategies are summarized in Figure 3-25.

3.2.2.4.2 Known cancer-associated SNVs and small InDels

In a further step, it was investigated, which variants were already annotated in common disease databases. Therefore, 287,604 SNVs and 43,195 InDels detected in at least one tumor sample were successfully converted from the murine mm10 assembly to the human hg19 reference.

490 coding variants (470 SNVs and 20 InDels) were found in the COSMIC database, whereof 18 SNVs were predicted as cancer-promoting by the FATHMM algorithm. Five of these potential driver mutations were associated with tumor development in the large intestine (Table S 31). Out of the 20 identified InDels, nine existed in at least two tumor samples listed in the COSMIC database and were therefore most likely no artifacts (Table S 31). Further 39 noncoding SNVs were reported for at least two tumor samples in the COSMIC database. Noticeable, the ratio of small InDels associated with cancers of the large intestine was significantly higher in variants found in the investigated tumors than in all variants annotated in the COSMIC database (48.4% and 34.5%, respectively, $p = 0.0031$ based on one-tailed Fisher's exact test). Surprisingly, no differences were observed for SNVs (20.8% vs. 17.0%, $p = 0.295$).

In the database Clinvar, seven SNVs with relevance for cancer diseases were found: hepatoblastoma (n = 1), malignant melanoma (n = 5), tuberous sclerosis syndrome (n = 1). Further four SNVs were involved in the development of metabolism-associated diseases (Lucey-Driscoll syndrome (n = 1), deficiency of xanthine oxidase (n = 1), and deficiency of UDPglucose-hexose-1-phosphate uridyl transferase (autosomal recessive disorder of galactose metabolism) (n = 2)) (Table S 32). None of the InDels found in the investigated tumor samples were annotated in the Clinvar database.

None of the tested variants existed in the GWAS catalog.

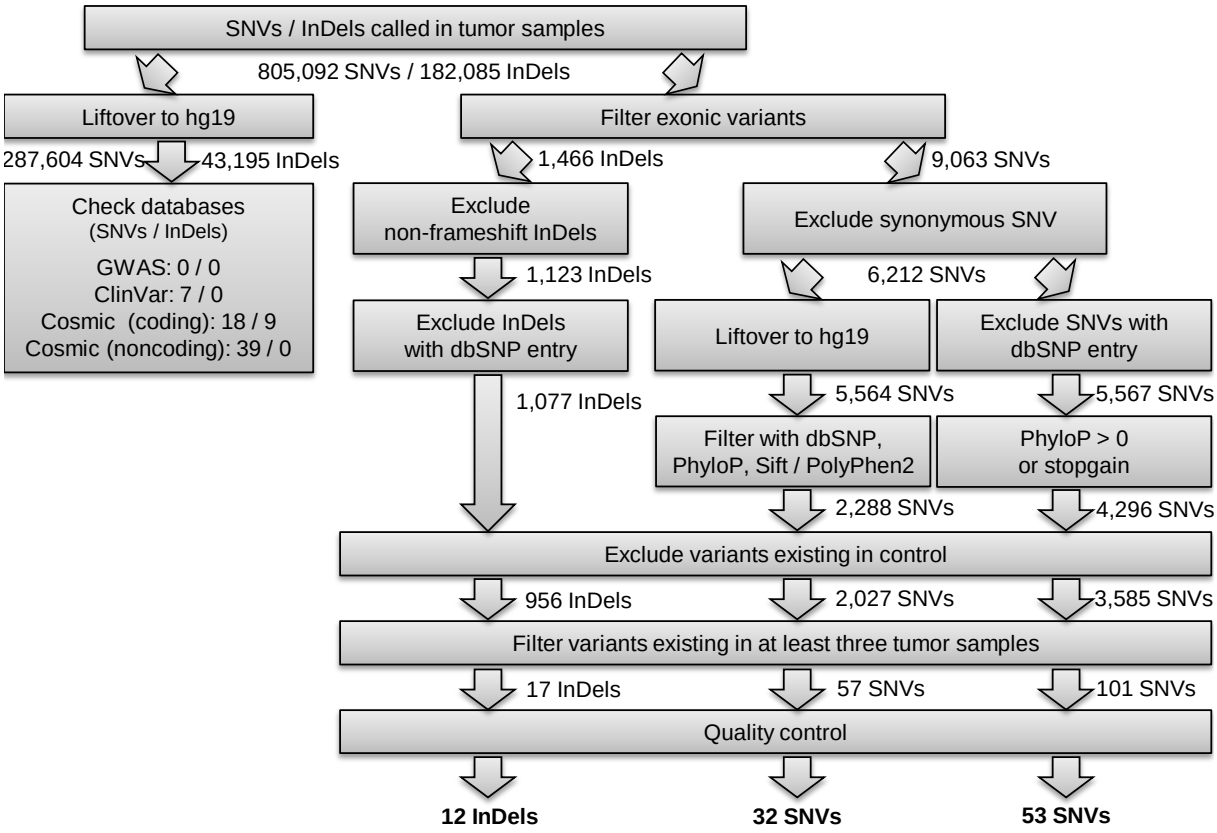


Figure 3-25 Variant filter workflow
Different strategies were applied to identify variants associated with AOM/DSS-trigged colorectal cancer. The left branch summarizes the results for known variants. The middle one displays the filter steps for novel InDels. For novel SNVs, two different approaches were used, which are shown on the right side.

3.2.2.5 Genes mutated in AOM/DSS-associated colorectal cancer

The likelihood to find a recurrent somatic variant was estimated to be low, because only a small sample number was investigated. Therefore, besides the investigation of variants, recurrence was addressed on gene level. Two approaches were applied to find genes potentially associated with AOM/DSS-triggered colorectal cancer. First, genes with a tumor-specific high variant density were identified. This included genes, which were affected by not more than one mutation (SNV or small InDel) in maximum one control sample and which harbored at least five variants in one tumor samples at the same time. Seven genes (*Triobp*, *Ran*, *LOC100044193* / *Gm20939*, *Foxo3*, *Eef2k*, *Clstn1*, and *1600014C10Rik*) fulfilled these criteria (Figure 3-26). The highest number of variants (18 SNVs) was observed in the gene *Foxo3* in sample C15R_T1. Interestingly, this mouse was affected by an intestinal prolapse and had to be excluded from the experiment 19 days before the regular finish date.

In a second strategy, genes were selected, which were, based on a one-tailed Fisher's exact test, significantly more often affected by a non-synonymous SNV or InDel in the tumors than in the union of control and DSS samples ($p \leq 0.05$). Genes having a mutation in more than two controls were excluded, because they were most likely hypervariable. Furthermore, all genes with a non-synonymous variant in at least three tumor samples and without any affected control

were included in the following analysis. This strategy resulted in a total number of 163 genes, which were called differentially mutated genes (Figure S 41). The subset with even four or more mutated tumor samples (31 genes) is shown in Figure 3-27. An especially high difference in the variant count between tumor and non-tumor samples was observed in the genes *Cenpf*, *Ctnnb1*, *Syne2*, and *Vmn2r51*.

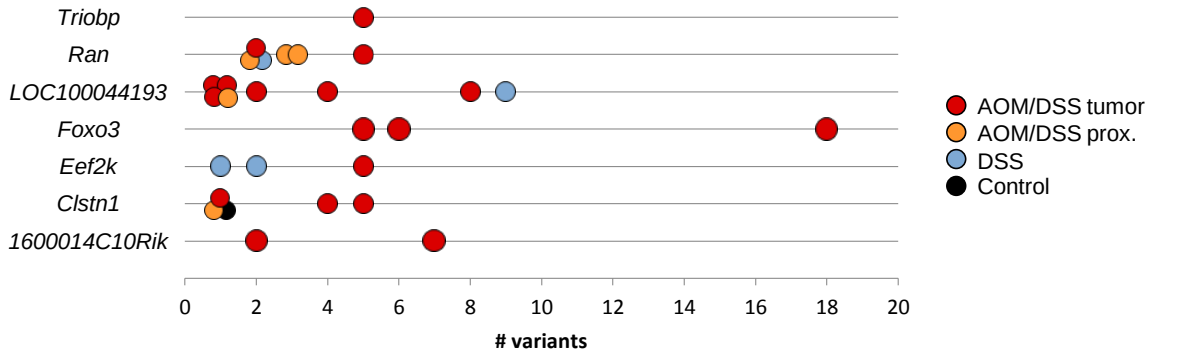


Figure 3-26 Genes with high variant density in tumor samples
The figure shows all genes with five or more variants (SNVs or InDels) in at least one tumor sample, maximum one SNV per control sample, and not more than one affected control sample. The plot illustrates for each sample the number of variants in the specific gene. Samples without mutation in the gene are not displayed.

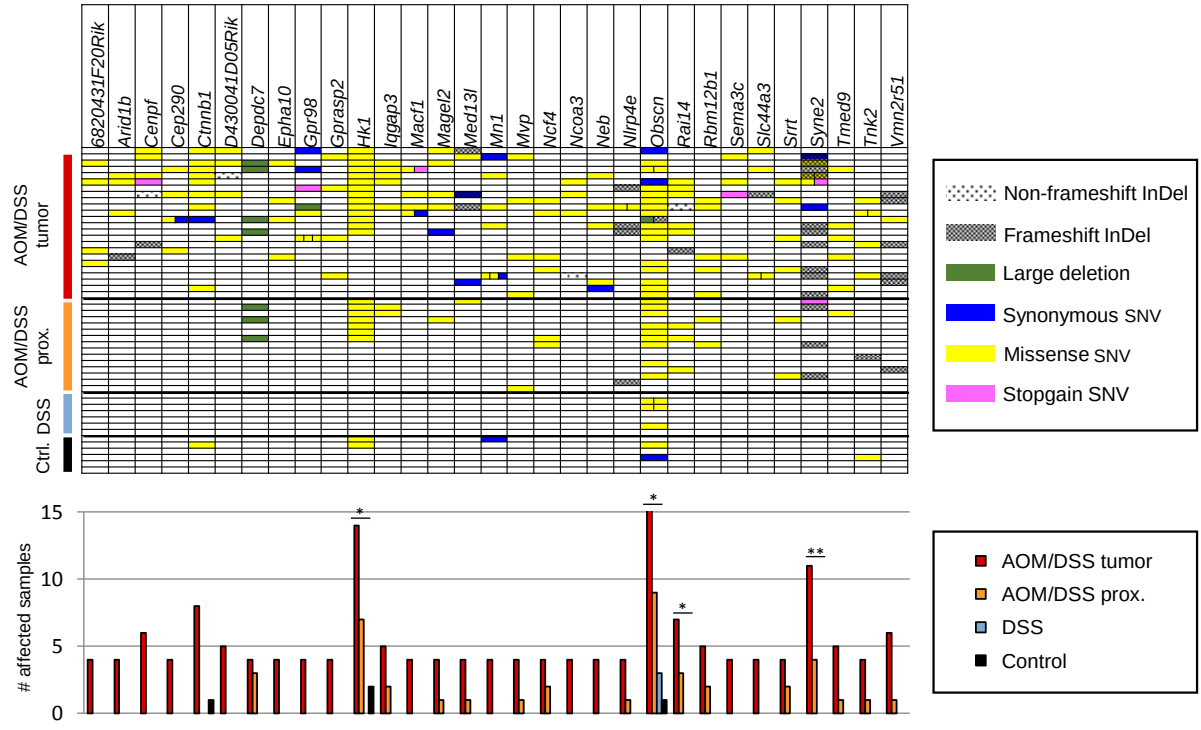


Figure 3-27 Oncoplot based on differentially mutated genes (subset)
The Oncoplot includes all genes, which were significantly more often affected by at least one InDel or non-synonymous SNV in the tumor than in the set of control and DSS samples. Furthermore, all genes having a non-synonymous variant in at least four tumors and in none of the controls are displayed. The colors of the boxes indicate the variant type. Each row provides information on one sample and each column on one gene. The lower part of the figure summarizes the Oncoplot and shows for each gene the number of samples affected by at least one InDel or non-synonymous SNV.

The 163 filtered genes, which were more often affected by a non-synonymous variant in the tumor than in the control samples, were used to create a protein-protein interaction network (Figure 3-28). All single nodes without connection to any other gene were not considered for further analyses. The connectedness of each node was investigated to infer important functional hubs of the network. A central position in this network was represented by *Ctnnb1* (β-catenin), which is known to be frequently mutated in CAC [49, 53] and had 14 connections to other differentially mutated genes. Other important hub genes were *Abl1* (nine connections), *Nos1* (eight), *Ncor2* (eight), and *Foxo3* (six). Based on the assumption that not just single genes but complete pathways were involved in the formation of tumors, proteins interacting with known cancer-associated genes were particularly strong candidates. Therefore, all genes, which were known to be involved in the development of tumors, were marked in the network shown in Figure 3-28. A gene was defined as known cancer-associated, if at least two samples with a mutation predicted as cancer driver / cancer-promoting by the tool FATHMM were listed in the COSMIC database. Based on this strategy genes such as *Chd8*, *Ncor2*, and *Myh7b* were potentially involved in the development of AOM/DSS-triggered colorectal cancer. Furthermore, known cancer-associated genes with evidence of an involvement in inflammatory processes were especially interesting genes. A salient gene within the WES data, directly linked to inflammatory signal transduction, was *Nfκbiζ*, which was characterized by the highest percentage of inflammatory, directly connected genes (Figure S 42).

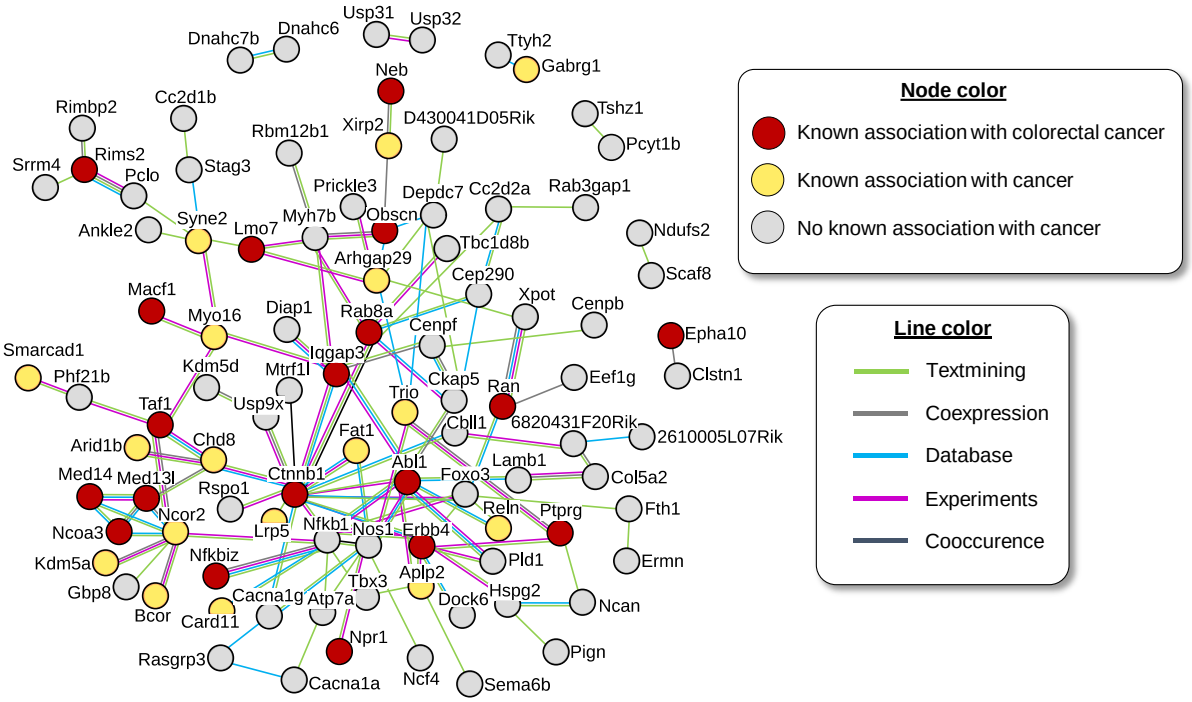


Figure 3-28 Protein-protein interaction network based on differentially mutated genes
The protein-protein interaction network is based on all proteins encoded by genes, which were more often affected in the tumors than in the set of DSS and control samples. The links are based on information from the STRING database. The line color indicates the source of the connection. Proteins with a known association to colorectal or colon cancer are marked in red. Proteins associated with another cancer type are displayed in yellow. All other proteins are shown in grey. Nodes without a connection to any other protein encoded by a differentially mutated gene were excluded from the figure. Nf-κb1 was added as one of the most important inflammatory genes, although it was not more often affected by a mutation in the tumors than in the pool of DSS and control samples.

A gene set enrichment analysis based on the 163 genes selected for the Oncoplot resulted in 133 significantly enriched GO terms (Table S 33). Surprisingly, no obviously cancer-associated, but several neurological, protein localization and protein modification processes were among the resulting terms. None of the KEGG pathways was significantly more often affected by the 163 genes than expected by chance.

3.2.2.6 Processes affected by non-synonymous variants in murine inflammation-triggered colorectal cancer

Pathways annotated in the KEGG database, which harbored more genes affected by a non-synonymous somatic variant in the tumor than in the pool of control and DSS samples, were investigated (Figure 3-29 A). In contrast to gene expression data, a direct comparison between sample types was necessary for each pathway to consider factors like conserved regions and gene length. 34 pathways were enriched for non-synonymous variants in the tumor samples including the three global and overview pathways ‘Metabolic pathways’ ($p = 0.0456$), ‘Chemical carcinogenesis’ ($p = 0.0488$), and ‘MicroRNAs in cancer’ ($p = 0.0372$). Among the specific KEGG pathways, it is noticeable that the p53 signaling pathway was enriched for genes having at least one non-synonymous variant ($p = 0.0456$), although no mutation in *Trp53* itself was detected in any of the tumor samples. Furthermore, an increased number of affected genes in the tumor samples was found in the PI3K-Akt signaling pathway ($p = 0.0456$) and several metabolic processes.

Although the proximal tissues of the AOM/DSS-treated mice might carry already some somatic mutations, it was decided to combine these samples with the DSS and control samples to further increase the statistical power. With this less stringent strategy, 42 additional potentially cancer-associated pathways were found, including the signaling pathways mTOR, Ras, and MAPK (Figure 3-29 B).

Instead of counting affected genes, the number of non-synonymous variants per pathway were investigated in a second approach. Nine processes harbored significantly more alterations in the tumor samples than in the non-tumor samples (controls, DSS, AOM/DSS proximal). While the enrichment of the pathways PI3K-Akt signaling and RNA degradation could be confirmed with this method, processes involved in cancer or metabolism harbored a similar number of variants between tumor and non-tumor samples (Figure 3-30).

Ctnnb1, which had a central role in the protein-protein interaction network shown above, is part of the Wnt signaling pathway. Although this pathway did not harbor significantly more mutated genes in the tumor than in the non-tumor samples ($p = 0.065$), subpathways of this process were investigated more detailed (Figure S 43). Interestingly, in the canoncial pathway and Wnt / Ca^{2+} pathway, tumor samples featured significantly more mutated genes in the tumor

than in the non-tumor samples ($p < 0.0001$ and $p = 0.0186$, respectively), while no differences were detected in the planar cell polarity pathway ($p = 0.491$).

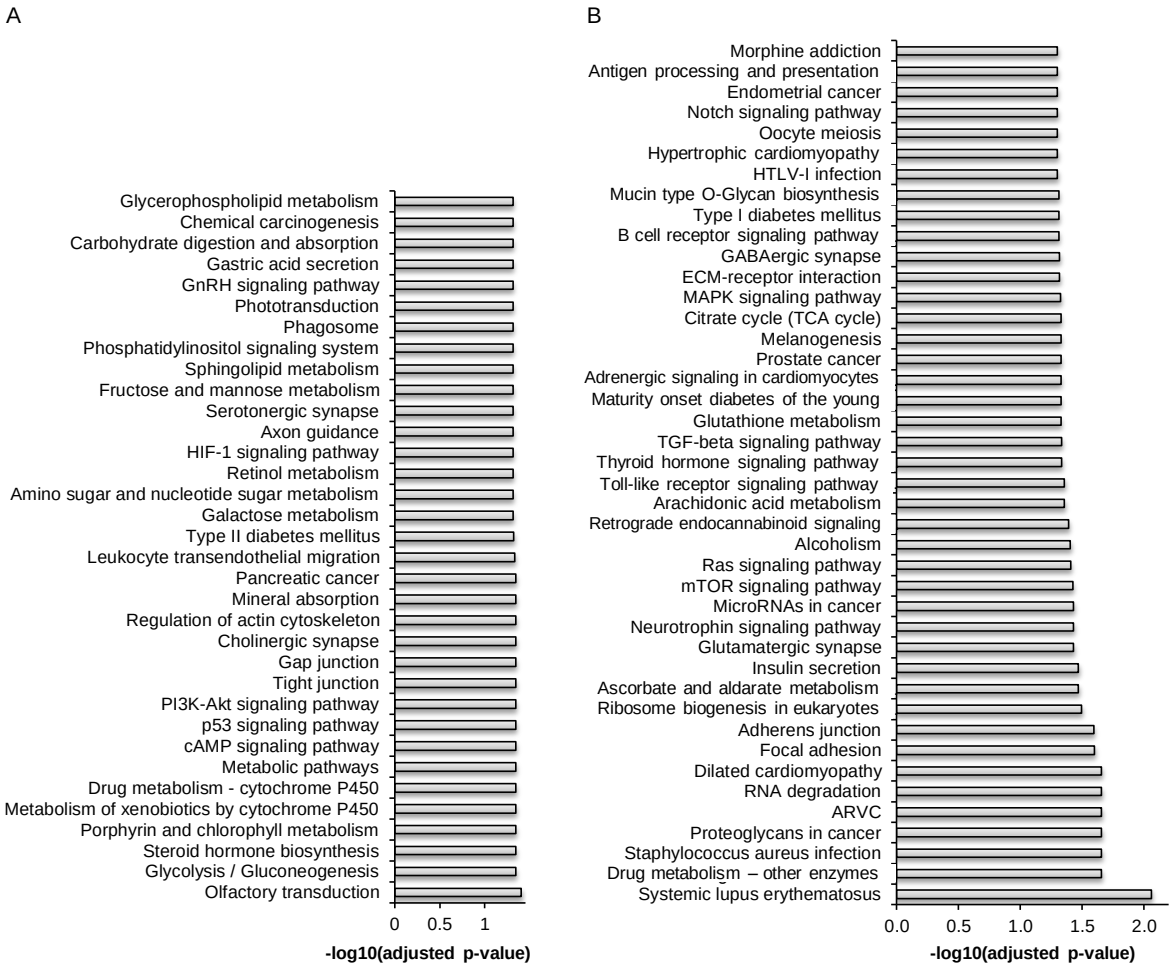


Figure 3-29 Enriched pathways based on genes affected by at least one non-synonymous variant
(A) Bar plot displays all pathways with significantly more genes affected by a non-synonymous SNV or InDel in the tumor samples than in the pool of control and DSS samples. (B) Barplot shows all pathways with significantly more genes affected by a non-synonymous SNV or InDel in the tumor samples than in the pool of control, DSS and AOM/DSS proximal samples. Pathways with significant p-value in the more stringent approach shown in figure part (A) are not displayed (B).

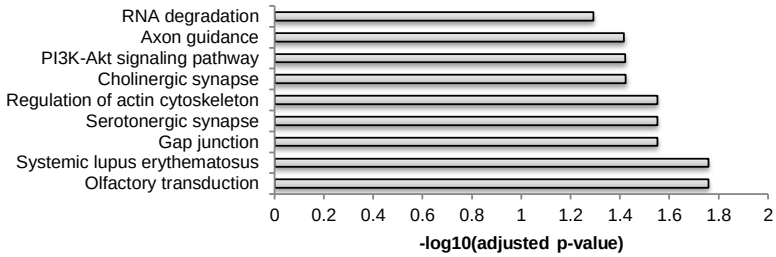


Figure 3-30 Enriched pathways on variant level
The barplot displays all pathways with significantly more non-synonymous variants in the tumor samples than in the pools of all non-tumor samples (controls, DSS, AOM/DSS proximal).

3.2.3 Transcriptome sequencing

3.2.3.1 Sequencing and mapping results

In total, 33 transcriptomes were sequenced using an rRNA depleted stranded RNA sequencing protocol on the Illumina HiSeq2500. Eleven samples were derived from tumor tissues, eleven

from the proximal non-tumor part of the colon from AOM/DSS-treated mice, five from mice exclusively treated with DSS and six from the proximal colon part from controls. On average 85 million reads were sequenced per sample. Between 57,491,068 and 194,958,544 sequences with a mean length of 99 nt and an average base quality between 35.6 and 37.3 passed the quality filter. A mean of 64 million reads per sample could be mapped to the murine reference genome mm10, of which ~87% harbored a unique alignment position (mapping quality = 50). Around 41% of the sequences shared the starting point with another read. These so-called duplicates were exclusively considered for quality control and, in contrast to the WES part, not excluded from further analyses. All samples showed a sense/antisense noise of less than 1.3%. On average 33% of the reads mapped in known exonic regions, whereby the fraction of ribosomal RNA was ~24%. Expression (FPKM value > 0.01) could be detected for ~59% of all known transcripts (Table S 34). A slight 5' bias was detected in all samples (Figure S 44). Interestingly, the interquartile range (IQR) of the FPKM values was higher in the tumor samples compared to the other treatment groups (Figure 3-31 A, Figure S 45 A) demonstrating the higher range of mRNA levels in cancer cells. This means that while many genes were downregulated due to mutations, the expression of tumor supporting genes were especially high. The IQR was particularly high in the tumor samples derived from mice treated with a high DSS dose, although possibly due to the low sample number, the comparison between different tumor sets (low, medium and high DSS dose) did not result in a significant p-value (Figure S 45 B). A technical bias causing this effect could be excluded, because tumor samples were processed at different points in time and the number of mapped reads did not differ from the control samples. Moreover, the influence of sequencing run and library batch on the variance of the FPKM values was low (4% and 7%, respectively), while ~35% of the variance could be explained by the treatment (Figure 3-31 B).

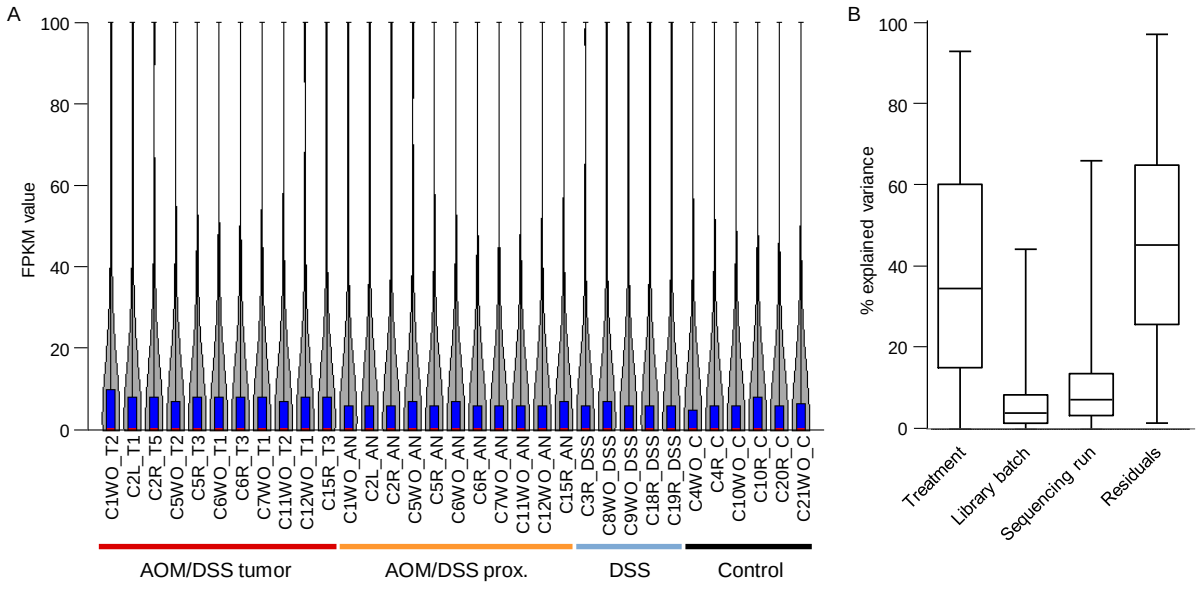


Figure 3-31 Expression IQRs and factors contributing to mRNA level distribution
(A) The Violin plot displays the distribution of the FPKM values. The IQR is marked in blue and the median in red. All FPKM values larger 100 were reduced to 100. (B) Influence of different factors on the gene expression level.

3.2.3.2 Sample distribution and outlier detection

Two tests were performed to detect outliers. First, the median K-dist based on all mapped sequences and the D-dist based on FPKM values were calculated. In this quality check, the sample C3R_D was identified as an outlier (Figure 3-32 A). This result was caused by an increased number of duplicates in the sample C3R_D produced by an amplification bias during the library preparation. Thus, the diversity of the library was lowered. This affected the expression levels of all genes in a similar amount but might led to a lack of detection of lowly expressed genes. Additionally, the high duplicate rate was supported by a high rRNA proportion in the sample C3R_D. The sample was used for further analyses, because the investigation of the transcriptome data was not influenced by the aberrant k-mer count.

In the second quality control step, an MDS plot was created to validate sample associations (Figure 3-32 B+C). The control sample C20R_C clustered together with the formerly inflamed tissue samples from DSS- and AOM/DSS-treated mice. Thus, the sample C20R_C was excluded as control sample from further analyses to avoid a falsification due to a potential real inflammation.

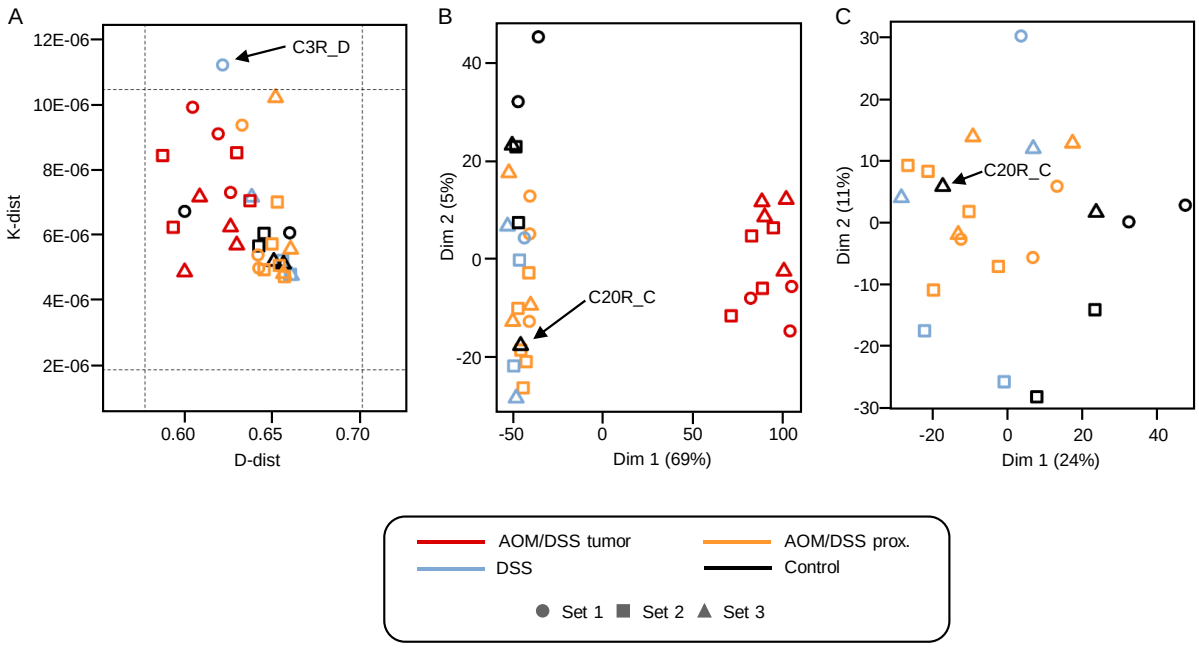


Figure 3-32 Sample distribution and outlier detection in RNA-Seq data
(A) The D-dist defined as median of all pairwise correlations of a given sample is shown on the x-axis, while the median k-mer distance (K-dist) based on all mapped reads is displayed on the y-axis. The dashed lines indicate the outlier thresholds calculated as Q1 minus 1.5 fold of the interquartile range and Q3 plus 1.5 fold of the interquartile range. The sample C3R_D was identified as an outlier. (B) MDS plot based on gene expression values of all samples. (C) MDS plot based on gene expression values of formerly inflamed and control tissue samples. The control sample C20R_C clustered together with the post-inflamed tissue samples and was therefore excluded from further analyses.

3.2.3.3 Genes and processes with modified transcriptional patterns in tumor samples

In the comparison between tumor and control samples, 5,663 significantly upregulated (1,078 with $p < 0.001$ and fold change > 4) and 5,829 significantly downregulated (1,467 with

p < 0.001 and fold change > 4) were detected. The 50 genes with lowest p-value are shown in Figure 3-33. A distinct separation between tumor and all other sample types was visible, but no clear clustering by DSS dose could be observed. Within the non-tumor samples, all control samples clustered together with two samples from mice treated with low DSS and were thus different from all other formerly inflamed and tumor tissue samples. As expected, the expression values of the tumor samples showed a higher deviation from the mean value per gene demonstrating a more extreme regulation.

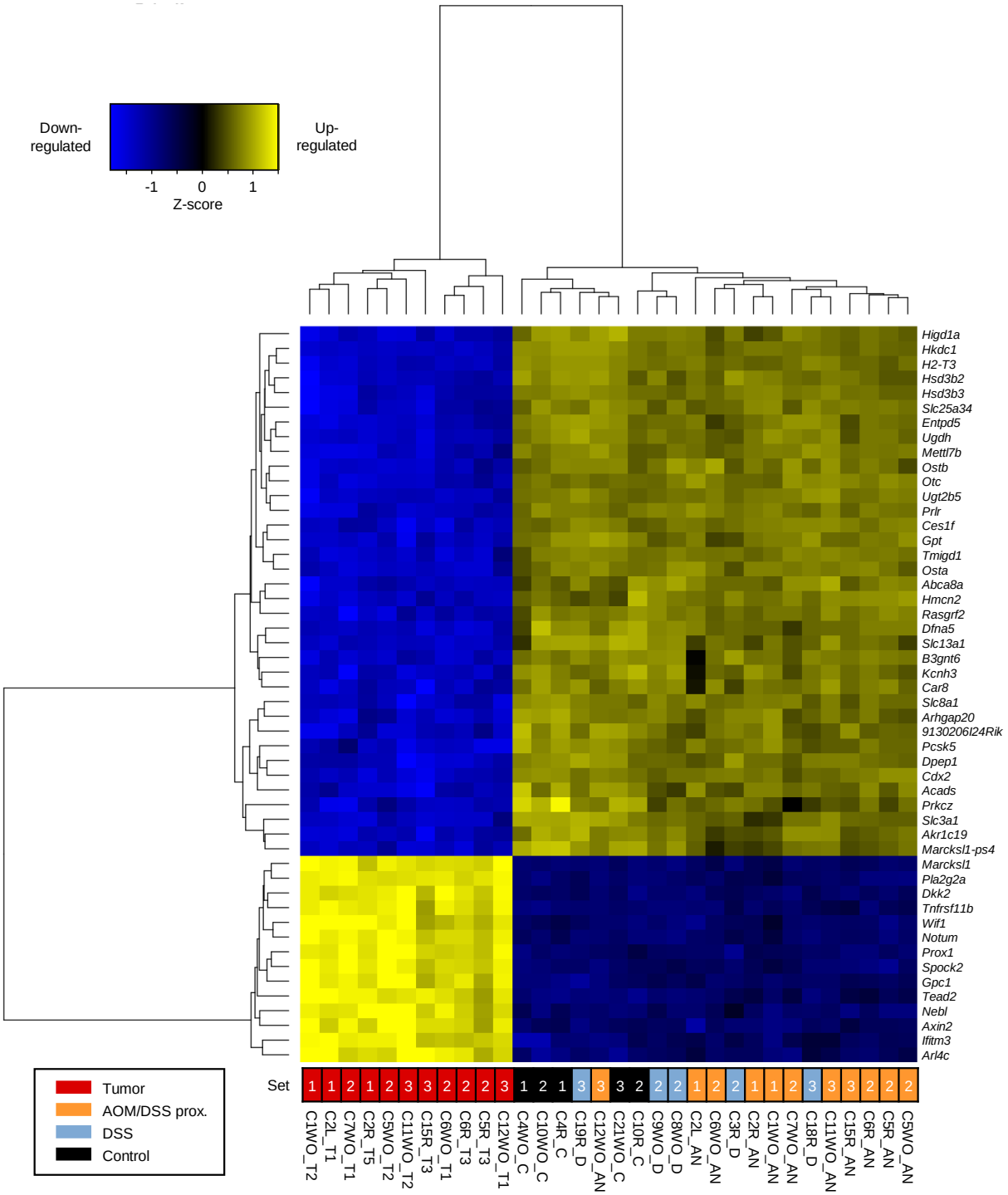


Figure 3-33 Genes differentially expressed between tumor and control samples
The heatmap visualizes the 50 differentially expressed genes with lowest p-value in the comparison between tumor and control samples. The color distribution indicates the z-score normalized expression values. The dendrograms are based on hierarchical clustering using Euclidean distance.

To identify up- or downregulated genes representing hubs in biological processes, a protein-protein interaction network based on all differentially expressed genes with p-value smaller 0.001 and a fold change larger four was created. This resulted in a so-called hairball without clear distinction between the elements. Therefore, the connectedness of each differentially expressed transcript within this network was interrogated. The 33 genes, which formed the overlap of the 50 genes with the highest number of direct connections and the 50 genes with the highest number of indirect connections (one intermediate node), were plotted as a subnetwork (Figure 3-34). The genes *Tnf*, *Acacb*, *Rac3*, and *Nos1* might play a central role in the development of AOM/DSS-associated colorectal cancer, as they were at highly branched positions within the complete network.

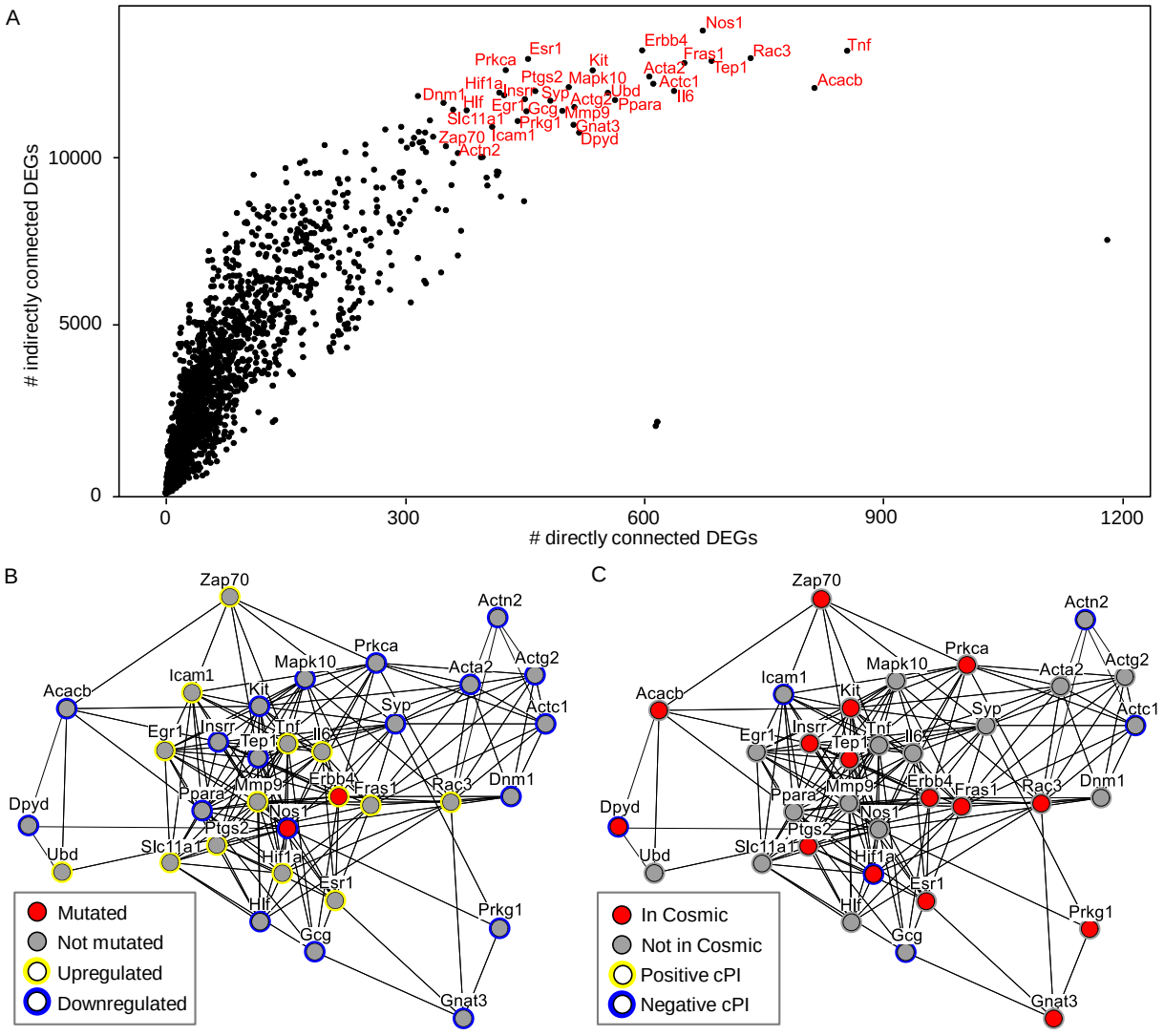


Figure 3-34 Protein-protein interaction network based on differentially expressed genes
(A) Number of direct and indirect connections of all genes, which were differentially expressed between tumor and control samples ($p < 0.001$ and fold change > 4). The 33 nodes with red label indicate genes, which were amongst the 50 genes with the highest number of direct connections to differentially expressed genes and amongst the 50 genes with the highest number of indirect connections to genes with altered mRNA level. (B) + (C) Protein-protein interaction network based on the 33 highly connected differentially expressed genes in (A). (B) The color coding is based on alterations observed in the current study. (C) The coloring is based on known cancer-associated alterations: The node filling indicates whether variants in the gene were annotated as cancer driver mutation in the COSMIC database. The node border shows the direction of the cPI score: A positive cPI score refers, like an upregulated gene, to a potential oncogene. Vice versa, a negative cPI score refers, like a downregulated gene, to a potential tumor suppressor.

3.2.3.4 Processes and transcription factor classes enriched for differentially expressed genes

In the next step, it was investigated whether subsets of differentially expressed genes with p-value < 0.001 and fold change > 4 between tumor and control samples were enriched for specific processes. Downregulated genes were more often than expected by chance part of processes involved in homeostasis and metabolism, such as 'Drug metabolism' (KEGG: p = 1.441E-06), 'Glucose homeostasis' (GO: p = 0.000148), 'Cellular homeostasis' (GO: p = 1.33E-05) as well as 'Regulation of cell communication' (GO: p = 0.000563) and 'Protein digestion and absorption' (KEGG: p = 0.000553). In contrast, upregulated genes were enriched in processes, which were mainly associated with cancer or immune response, such as the 'Wnt signaling pathway' (KEGG: p = 0.000642, GO: p = 0.0185), 'Pathways in cancer' (KEGG: p = 0.0109), 'Cell cycle' (GO: p = 3.67E-21), 'Cell migration' (GO: p = 2.36E-07), 'Stress-activated MAPK cascade' (GO: p = 1.16E-05), and the 'ERBB signaling pathway' (GO: p = 0.00149). Also the functional group 'Gene expression' harbored more differentially expressed genes than expected, which is in line with the higher interquartile range of expression values in the tumor samples shown in Figure S 45 and the higher number of identified nTars shown in chapter 3.2.4. Additional interesting processes enriched for upregulated genes included 'Cell-cell adhesion' and 'DNA repair'. In the GO term 'Cellular response to chemical stimulus' were more upregulated and more downregulated genes than expected. All KEGG terms enriched for differentially expressed genes are shown in Figure 3-35. All GO terms with the highest hierarchy level and an adjusted p-value smaller 0.001 are shown in the supplementary Table S 35 and Table S 36.

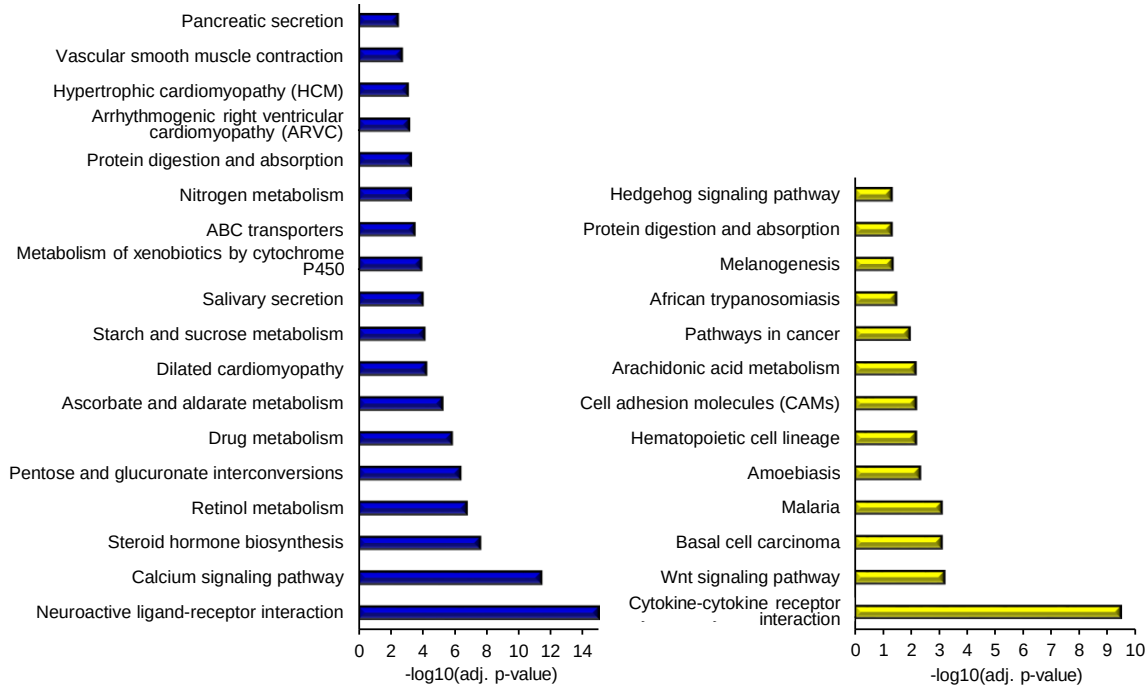


Figure 3-35 KEGG pathways enriched for differentially expressed genes
(A) Processes enriched for downregulated genes. (B) Processes enriched for upregulated genes.

Transcription factor prediction analysis showed that upregulated genes were enriched for 41 Transfac transcription factor classes (Table S 37) as well as 95 transcription factor classes annotated in the Jaspar database (Table S 38) including p53, Nfkb1, and Stat3. Downregulated genes were enriched for 40 Transfac transcription factor classes (Table S 39) as well as 67 transcription factors annotated in the Jaspar database (Table S 40) including Sp1, Nrsf, and Spz1.

3.2.3.5 Association between gene expression and specific tumor features

Besides the distinction between tumor and control samples, several genes and processes were differentially expressed between tumor samples. Thus, the effect of features such as DSS dose, intestinal prolapse, and intestinal tumor position were investigated next.

A comparison between tumor samples treated with different DSS doses were performed to identify genes and processes changing the activation level at specific tumor stages. In total, 388 genes divided into five clusters were significantly differentially expressed between the three tumor sets (Figure S 46, Table S 41). The number of five subclusters was deduced as the minimum amount of clusters with the highest amount of separation. While the first four clusters showed a clear upregulation in tumors from mice treated with a high DSS dose compared to those from animals exposed to a low DSS dose, the last cluster included genes characterized by a downregulation in animals, which obtained a high DSS dose. The first four clusters differed mainly in the expression level of tumors from mice fed with medium DSS dose. In detail, the first cluster contained genes, which were already overexpressed in all tumors from mice treated with medium DSS dose and were thus early activated during the tumor progress. These genes were enriched for adhesions processes such as ‘Focal adhesion’ (KEGG: p = 0.00013), ‘Regulation of the actin cytoskeleton’ (KEGG: p = 0.016), ‘Immune system processes’ (GO: p = 0.039) as well as cancer-associated processes such as ‘Cell migration’ (GO: p = 8.94E-06) and ‘Angiogenesis’ (GO: p = 0.047). The second cluster included genes, which were only slightly activated in the large tumors (C5WO_T2 and C7WO_T1) from mice treated with medium DSS dose. These genes, which were activated at a late stage during tumor progression, were more often than expected in the processes ‘MAPK signaling pathway’ (KEGG: p = 0.046) and ‘Regulation of actin cascades’ (KEGG: p = 0.030). The third cluster included genes, which were activated late during tumor progression, but before the genes out of the second cluster. These genes were upregulated in large (C5WO_T2 and C7WO_T1) but not in small (C6WO_T1, C6R_T3, C5R_T3) tumors from mice treated with medium DSS dose. These genes belonged to processes involved in the innate and adaptive immune system, such as ‘B cell receptor signaling pathway’ (KEGG: p = 0.00041), ‘Cytokine-cytokine receptor interaction’ (KEGG: p = 0.030), ‘Chemokine signaling pathway’ (KEGG: p = 0.0027), and ‘Natural killer cell mediated cytotoxicity’ (KEGG: p = 0.0096). The

fourth cluster contained early activated genes, which showed an even higher expression level in the large tumors from mice treated with medium DSS dose than in tumors from mice fed with high DSS dose. Also these genes were more often than expected included in processes involved in the immune system, such as ‘Antigen processing and presentation’ (KEGG: $p = 0.0066$), ‘Intestinal immune network for IgA production’ (KEGG: $p = 0.000099$), and ‘Regulation of acute inflammatory response’ (GO: $p = 0.014$). Genes of the fifth cluster were downregulated in tumors from animals treated with high DSS doses. Affected processes included the ‘P53 signaling pathway’ (KEGG: 0.00056), ‘Metabolic processes’ (GO: $p = 0.0079$), and ‘Regulation of the cell cycle process’ (GO: $p = 0.010$).

Besides the described differences in activation points in time of processes in tumor samples, several functional groups and pathways were potentially involved in the malignant transformation. This topic was addressed by a comparison of the expression levels between proximal colon samples from AOM/DSS treated mice and controls ($p < 0.001$). To exclude regulation changes due to chronic inflammation, all genes differentially expressed between samples from DSS treated animals and controls ($p < 0.05$) were excluded. The applied p-values thresholds of the two comparisons were set to different values to minimize artifacts due to e.g. power problems. The 115 genes potentially associated with the malignant transformation were assigned into five cluster (Figure 3-36, Table S 42), which was the lowest possible number of groups having a minimum cluster size of ten genes. The first cluster contained genes, which showed a slightly increased expression in samples from DSS treated mice compared to controls and an even higher mRNA level in the proximal colon samples from AOM/DSS treated animals. Interestingly, in the tumor samples, the gene regulation was similar to the controls demonstrating that these genes were exclusively active during malignant transformation. The first cluster included the tumor suppressor gene *KLF6*, which is known to be upregulated in IBD patients [199] but often lost or mutated in CAC tissues [200]. Also *RNF11*, which was suggested already suggested to be involved in the malignant transformation, was part of this group. However, it was also reported that *RNF11* is upregulated in several cancer types [201]. Moreover, the observed expression pattern of *CPNE2* contradicted previous studies, in which a positive correlation of the *CPNE2* mRNA level with tumor stage and distant metastasis was reported [202]. Genes of the first cluster were enriched in several processes mainly involved in protein metabolic processes ($p = 0.0432$) (Table S 42). The second expression cluster of genes potentially involved in malignant transformation contained genes, which showed an increasing activity from controls, to the proximal colon samples from AOM/DSS treated mice, up to tumors. These genes, which were activated at an early step during tumor progression, existed more often than expected in the KEGG pathways ‘P53 signaling pathway’ ($p = 0.0022$), ‘Wnt signaling pathway ($p = 0.046$)’, ‘Pyrimidine metabolism’ ($p = 0.031$), and ‘Glycerolipid metabolism’ ($p = 0.016$) as well as several functional

groups involved in the ‘Response to unfolded protein’ (p = 0.031) (Table S 42). Genes of the third and sixth regulation cluster showed an opposing behavior to the first one. These genes were especially downregulated during malignant transformation and thus might inhibit this process. The functions of genes included in this group were manifold and included tumor suppressors, such as *Rbm5* [203], as well as oncogenes, such as *Arhgef2* [204] and *Cdc14B* [205]. The fourth cluster contained genes showing decreasing mRNA level with increasing tumor progress and are thus potential tumor suppressors. While this is indeed true for *Flcn* [206], *Cdc28A* promotes tumorigenesis and cell proliferation in other cancer types [207]. The fifth regulation cluster showed a similar expression distribution like the second one, but the genes showed already a high expression in samples from AOM/DSS treated mice. These potential oncogenes were enriched for processes involved in ‘DNA binding’ (p = 0.0376) and ‘Transcription by RNA polymerase II’ (p = 0.0331) (Table S 42).

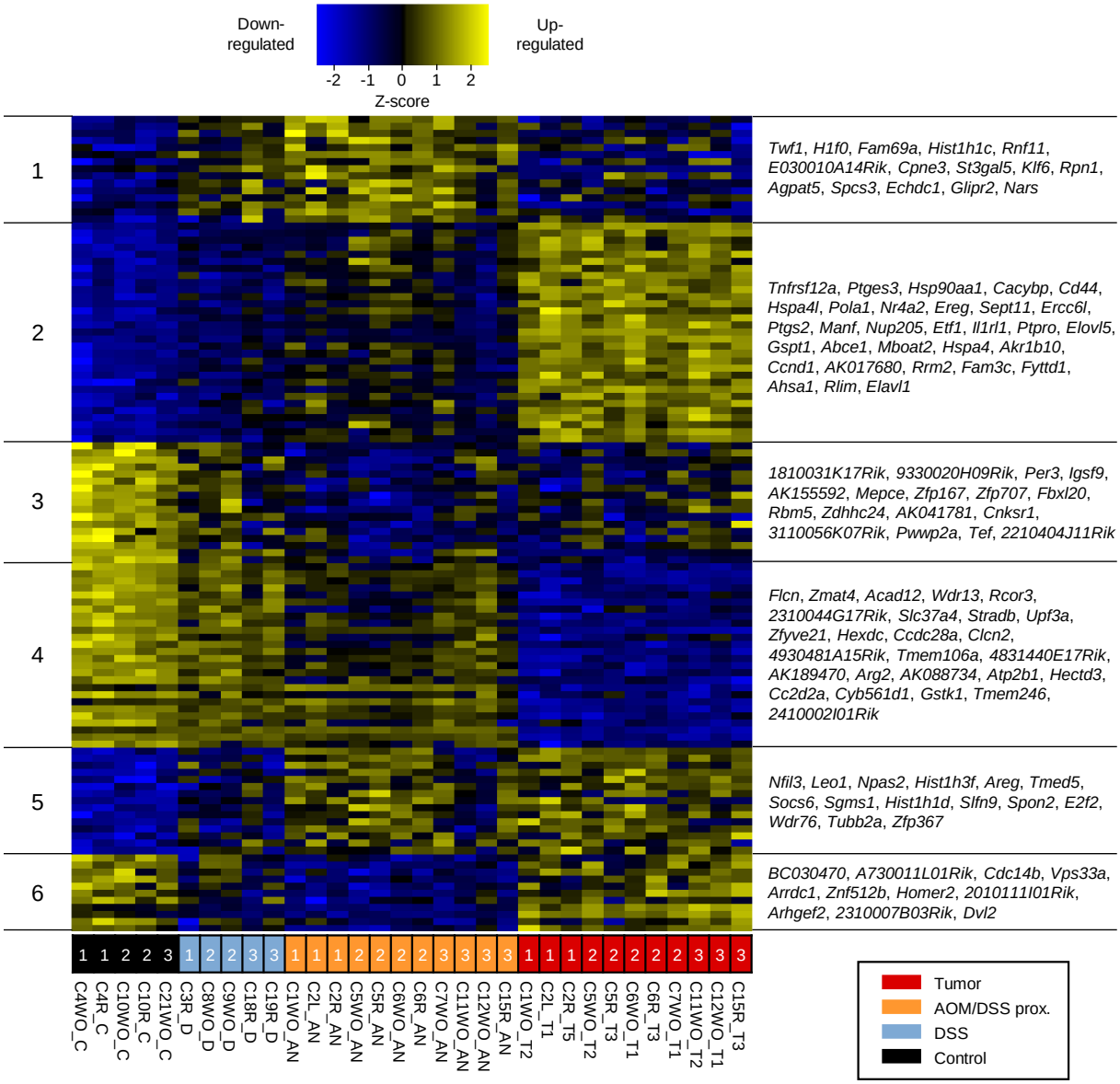


Figure 3-36 Genes potentially involved in malignant transformation
All genes shown in the figure were differentially expressed between controls and proximal colon samples from AOM/DSS treated mice (p<0.001) but not between controls and samples from animals exposed only to DSS (p>0.05). The grouping of the genes is based on a k-means clustering.

In addition to the tumor stage, the tumor localization in the colon associated with a specific expression pattern (Figure 3-37). The seven identified gene clusters demonstrated that not a smooth transition between the sections with tumor appearance existed, but a clear separation between tumors in the rectum, in the distal, and in the central part of the colon. As example, genes associated with the 'Cell cycle' (KEGG: p = 0.0060) or the 'Regulation of the actin cytoskeleton' (KEGG: p = 0.0061) were significantly higher expressed in tumors located in the distal part of the colon, while an increased expression of metabolic processes (KEGG: p('Drug metabolism') = 0.0022) was observed for tumors in the rectum. Pathways related to the immune system were characterized by the highest mRNA level in central located tumors followed by tumors in the rectum and the distal part of the colon. Genes with highest expression in the central colonic section were associated with the 'Wnt signaling pathway' (GO: p = 0.042), 'Chemical homeostasis' (GO: p = 0.027) as well as 'Transport' (GO: p = 0.037) and 'Regulation of the cell communication' (GO: p = 0.00022) (Table S 43, Table S 44).

Mice affected by AOM/DSS-triggered colorectal cancer can develop an intestinal prolapse, which is a very rare event in human CAC. The underlying molecular mechanisms for this difference are up to now mostly unknown. Therefore, tumors from mice with intestinal prolapse were compared with tumors from mice without intestinal prolapse resulting in 45 differentially expressed genes divided into three gene clusters, which were potentially associated with the development of an intestinal prolapse. The tumor sample C15R_T3 was retrieved from a mouse, which suffered from an intestinal prolapse for several days and was affected by a severe weight loss, while the tumor sample C1WO_T1 was from a mouse, which developed the intestinal prolapse shortly before the organ removal. This discrepancy was also reflected in the observed expression patterns (Figure S 47). The first out of three gene clusters contained genes, which were exclusively overexpressed in the tumor sample from the mice affected by an intestinal prolapse for a longer time. These genes were mostly involved in processes associated with 'Water homeostasis' (GO: p = 0.0063), 'Thickened skin' (GO: p = 0.0031), and metabolic processes, such as 'Pantothenate and coenzyme A biosynthesis' (KEGG: p = 0.0031) and 'Metabolism of xenobiotics by cytochrome P450' (KEGG: p = 0.0022). The second gene cluster contained genes with lower expression in both tumor samples retrieved from mice with intestinal prolapse. This group were mainly involved in the 'Cellular ion homeostasis' (GO: p = 0.042) and the 'Immune response' (GO: p = 0.032) with particular focus on the 'Innate immune response' (GO: p = 0.028). The third cluster was opposed to the second one and included genes, which were upregulated in tumor samples from mice with intestinal prolapse. This group was based on biosynthetic (GO: p = 0.047) and metabolic processes, such as the 'Cellular lipid metabolic process' (GO: p = 0.047) and the 'Organic acid metabolic process' (GO: p = 0.047).

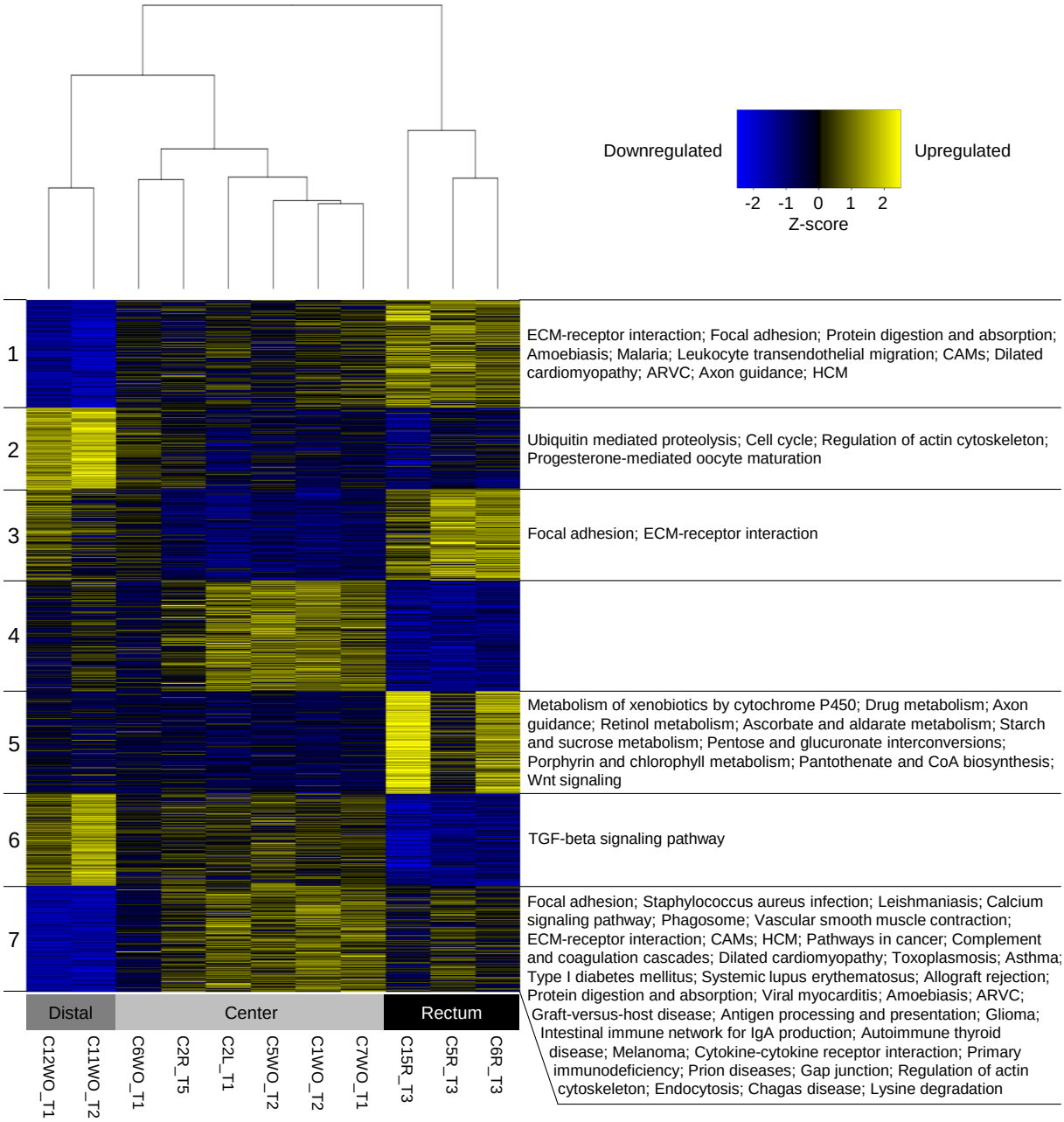


Figure 3-37 Processes associated with the colorectal position of the tumor
A k-means clustering was performed on all genes, which were differentially expressed between tumors located in rectal, distal, and central colorectal sections. KEGG pathways, which were significantly enriched for one of the resulting clusters, are shown on the right site of the plot. The following abbreviations were used in the figure: CAMs = Cell adhesion molecules, HCM = Hypertrophic cardiomyopathy, and ARVC = Arrythmogenic right ventricular cardiomyopathy.

3.2.3.6 Splicing patterns in AOM/DSS-induced tumors

It is known that aberrant alternative splicing is linked to cancer [208]. However, most publicly available programs for the investigation of splice variants use an algorithm to estimate the expression of different isoforms of a gene. This approach bears the risk of a high error rate due to overlapping sequence parts of transcript isoforms. Therefore, the following three splicing types were investigated separately in the current study: alternative exon usage, shift of splice site positions, and different intron retention rate. In total, 867 genes were affected by

at least one splicing event, which was significantly different between tumor and the union of formerly inflamed (proximal colon samples from AOM/DSS or DSS treated mice) and control tissue samples. The most abundant splicing type was a shift of the splice donor site position (612 genes) followed by a change of the splice acceptor site position (356 genes). Based on a binomial test, these two splicing types were significant increased compared to both general alternative splicing in mice [209] ($p = 1.648E-313$ and $p = 4.39E-18$, respectively) and the assumption of a random distribution ($p = 2.145E-31$ and $p = 1.973E-134$, respectively). Furthermore, introns and outer exons that were significantly more often retained in the tumor samples (71 and 45 affected genes, respectively) were frequently observed in the current study (Figure 3-38).

Genes harboring at least one significantly different splicing type were, amongst others, enriched in processes associated with metabolic processes (GO: $p = 4.92E-25$), cell adhesion (e.g. 'Tight junction' (KEGG: $p = 0.0241$)), 'Bacterial invasion of epithelial cells' (KEGG: $p = 0.0276$) as well as several signaling pathways including the 'NOD-like receptor signaling pathway' (KEGG: $p = 0.0241$) and the 'Wnt signaling pathway' (GO: $p = 0.0226$) (Figure 3-38, Table S 45). However, not all processes were affected by all types of alternative splicing.

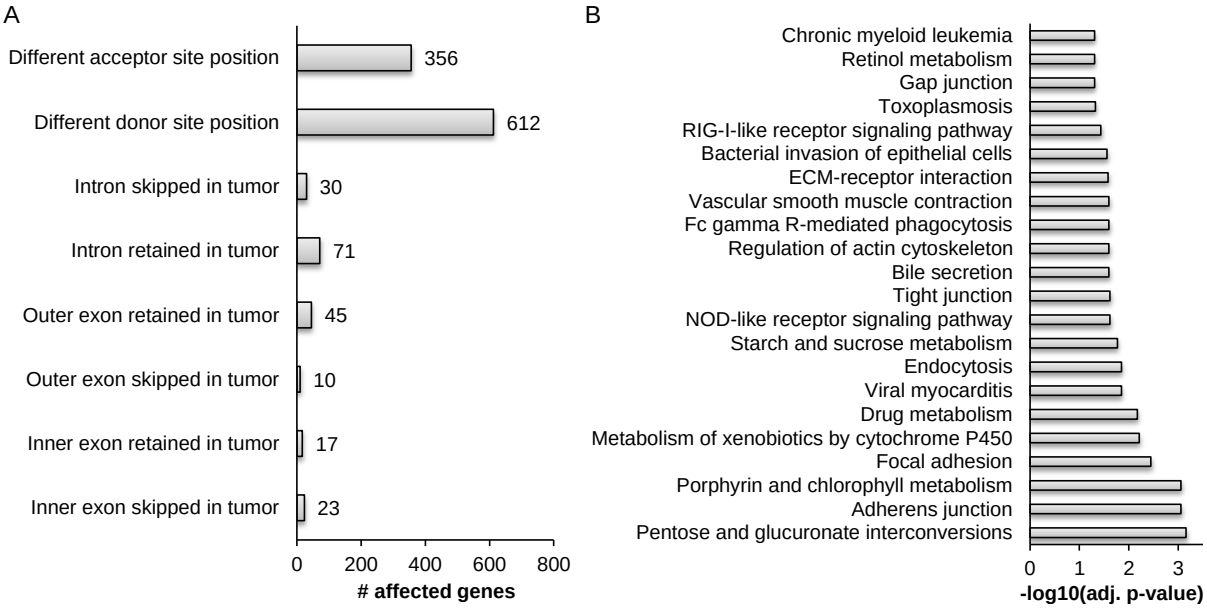


Figure 3-38 Overview of splice events
The shown results are based on all splice events, which were significantly different between tumor and all non-tumor samples. (A) Overview of the number of splice events for each investigated splicing type. (B) KEGG pathways, which were enriched for genes harboring at least one splice event.

To gain deeper insights into the effect of each splicing type, all splicing forms were separately subjected to an additional analysis. First, different exon usage was investigated between tumor and the union of control and formerly inflamed tissue samples. For 17 genes, at least one exon was more often included in the tumor than in the control samples. Affected genes were for example *Dgkd*, which might be involved in tumorigenesis [210], and *Gnas*, which is associated with tumor formation [211]. Genes with higher PSI value in the tumor samples were, amongst

others, enriched for the KEGG pathways ‘Tight junction’ (p = 0.04), ‘Regulation of actin cytoskeleton’ (p = 0.05), as well as for the GO-term ‘Biological adhesion’ (p = 0.027) including ‘Cell adhesion’ (p = 0.025) (Figure S 48 A+C). Vice versa, in 23 genes, at least one exon was skipped significantly more often in the tumor samples than in the controls, including genes encoding for tumor suppressors, such as *Ceacam1* and *Trp53inp1*, tumor-associated antigens, such as *Ctage5* and *Nbr1*, *Lrrfip*, which might function as activator of the canonical Wnt signaling pathway, and the gene *Ppp1r13b* encoding for a p53 interacting protein. Genes with significantly lower PSI in the tumor samples were enriched in the KEGG pathway ‘Focal adhesion’ (p = 0.0032) and the GO terms ‘Regulation of interleukin-10 secretion’ (p = 0.004), ‘Regulation of CD4-positive, alpha-beta T cell proliferation’ (p = 0.00067), and ‘Protein secretion’ (0.0469) (Figure S 48 B+D). However, most of the processes with aberrant PSI in tumor samples were affected by only two differentially spliced genes. To increase the power of the enrichment analyses, processes were identified, which harbored an unexpected high number of genes with at least one differentially spliced exon using an unadjusted p-value smaller 0.05. In this test, exclusively pathways affected by at least nine genes with differentially spliced exon were considered. As a result, genes with retained exon in the tumor samples were e.g. enriched for the KEGG pathways ‘MAPK signaling pathway’ (p = 0.017), ‘Regulation of actin cytoskeleton (p = 0.017) (Figure 3-39 A) as well as for the functional GO terms ‘Primary metabolic process’ (p = 5.58E-6), ‘Positive regulation of immune system process’ (p = 0.0269), ‘Wound healing (p = 0.0049), and ‘Cell cycle (p = 3.02E-04) (Table S 46). In contrast, the following processes were, amongst others, significantly enriched for genes with at least one exon, which was more often spliced in the tumor samples: The KEGG pathways ‘Wnt signaling pathway’ (p = 0.029), ‘B cell receptor pathway’ (p = 0.038), and ‘ECM-receptor inactivation’ (p = 0.021) (Figure 3-39 B) as well as the GO terms ‘Activation of innate immune response’ (p = 0.0426), ‘Cellular process’ (including ‘Cell division (p = 0.000578), ‘Cell cycle’ (p = 1.67E-11), ‘Cell death’ (p = 1.42E-06), and ‘Cytoskeleton organization’ (p = 1.32E-08)), and the subprocesses ‘P53 binding’ (p = 0.03), ‘IκB kinase / NF-κB signaling’ (p = 0.0017), ‘Telomere maintenance’ (p = 0.0215), ‘DNA repair’ (p = 0.0197), and ‘TOR signaling’ (p = 0.017) (Table S 47). Amongst processes, which were enriched for exon retention as well as for exon skipping in the tumor samples compared to the non-tumor samples, were the GO terms ‘Cytoskeleton organization’ [p = 1.62E-04 (exon retained in tumor) / p = 4.62E-09 (exon skipped in tumor)], ‘Apoptotic process’ (p = 0.0186 / p = 9.04E-06) and ‘Response to stress’ (p = 0.0059 / p = 1.62E-06).

Surprisingly, out of genes with different outer exon usage between tumor and control samples, significantly more genes retained the outer exon in the tumor samples (36 genes) than in the controls (eight genes). However, for genes with at least one exon retained more often in the

tumor samples than in the controls only the KEGG pathway ‘Purine metabolism’ (p = 0.0238) was significantly enriched.

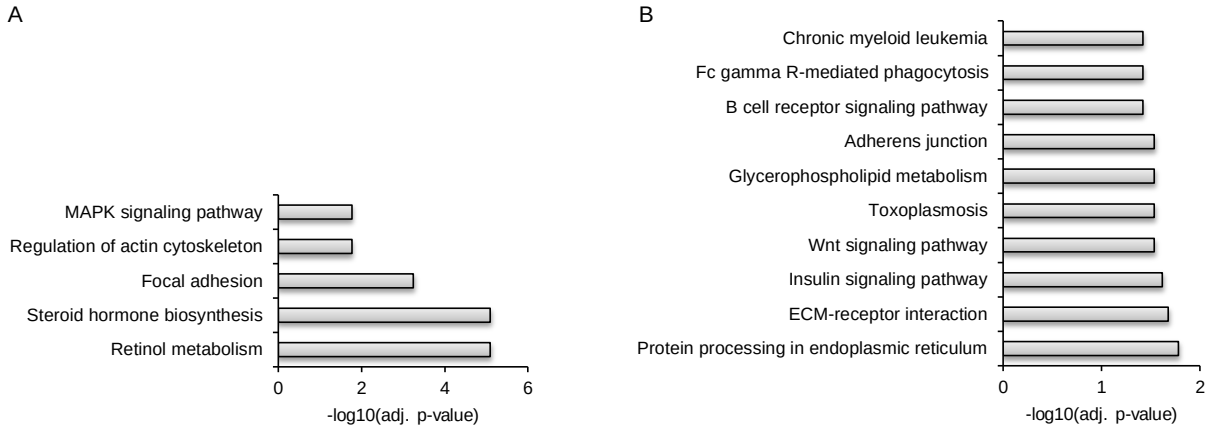


Figure 3-39 KEGG pathways enriched for genes with aberrant PSI between tumor and all non-tumor samples
(A) KEGG pathways enriched for genes with at least one exon, which was more often retained in the tumor compared to all non-tumor samples (gene selection based on unadjusted p-value < 0.05). (B) KEGG pathways enriched for genes with at least one exon, which was more often spliced in the tumor compared to all non-tumor samples (gene selection based on unadjusted p-value < 0.05).

In 356 genes, the position of at least one acceptor site differed significantly between tumor and control samples, although the same donor site was used. Thereby, the position of the acceptor site was known or novel, while the donor site was already annotated in the used reference. A shift of the acceptor site position affected, amongst others, the tumor suppressor genes *Ceacam1*, *Nbr1*, *Spint2*, and *Trit1*, the tumor metastasis related genes *Cd44* and *Cttn* as well as further genes related to cancer development or inflammation like *Dmbt1*, *Tbck*, *Dsc2*, *Lrrfip2*, *Maged2*, *Tpd52l2*, *Yap1*, *Samhd1*, and *Tax1bp1*. In addition, genes encoding for AXIN2, which plays an important role in the regulation of the stability of CTNNB1 in the Wnt-signaling pathway, and HUWE1, which ubiquitinates the p53 tumor suppressor, were affected. Amongst others, the genes were enriched for metabolic pathways, the KEGG process ‘NOD-like receptor signaling pathway’ (p = 0.003) as well as the GO terms ‘Regulation of CD4-positive, alpha-beta T cell proliferation’ (p = 0.03) and ‘Cellular response to chemical stimulus’ (p = 0.0077, Figure 3-40 A+C). In 612 genes, the position of the donor sites differed between the treatment sets, although the same acceptor site was used. An alternative donor site was observed in the tumor suppressor encoding genes *Casp2*, *Dcn*, *Ing*, *Mxi1*, *Rbbp8*, *Slc22a18*, *Trit1*, and *Mtus1* as well as genes associated with tumorigenesis, tumor progression, metastasis or other relations to cancer, such as *Pbrm1*, *Pcsk6*, *Rnf43*, *Vmp1*, *Cd151*, *Galnt12*, *Ing1*, *Mtus1*, *Mxi1*, *Rnf43*, *Steap4*, and *Tpd52l2*. Besides several metabolic processes, the genes were enriched for the KEGG pathways ‘ECM-receptor interaction’ (p = 0.0026) and ‘Regulation of actin cytoskeleton’ (p = 0.0058) as well as the GO terms ‘Response to drug’ (p = 0.006), ‘Regulation of cell death’ (p = 9.31E-5), ‘Positive regulation of gene expression’ (p = 0.0054), ‘Regulation of response to stress’ (p = 0.0081) and ‘Regulation of RNA splicing’ (p = 0.019) (Figure 3-40 B+D).

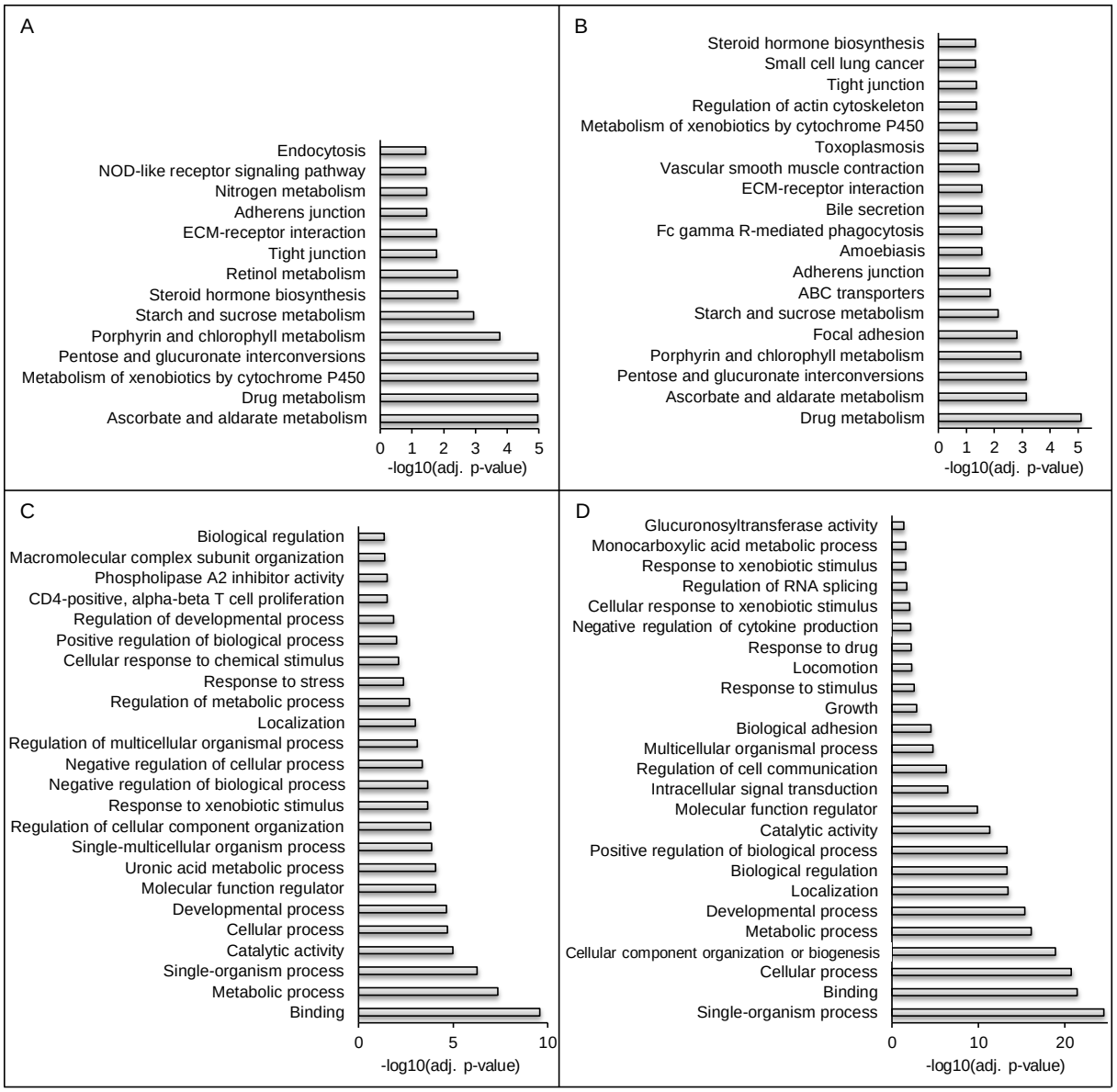


Figure 3-40 Processes enriched for genes with alternative splice site position
(A) KEGG pathways enriched for genes with alternative splice acceptor site position. (B) KEGG pathways enriched for genes with alternative splice donor site position. (C) GO terms enriched for genes with alternative splice acceptor site positions. (D) GO terms enriched for genes with alternative splice donor position. Only GO terms of the most general significant branch level are shown. Therefore, processes like 'Regulation of response to stress', which is a 'Biological regulation' as well as 'Response to stimulus', or 'Positive regulation of gene expression', which belongs to 'Metabolic process' as well as 'Biological regulation' are not displayed.

Besides the genes and processes with alternative splice site, the lengths of the splice site shifts were investigated. The majority of alternative positions were between 1000 and 9999 bp away from each other. This was the case for acceptor as well as donor sites (Figure 3-41). The alternative intron lengths seemed to be independent of open reading frames, because only between 31% and 34% of the distances between two alternative splice sites were divisible by three, which would also be expected by chance.

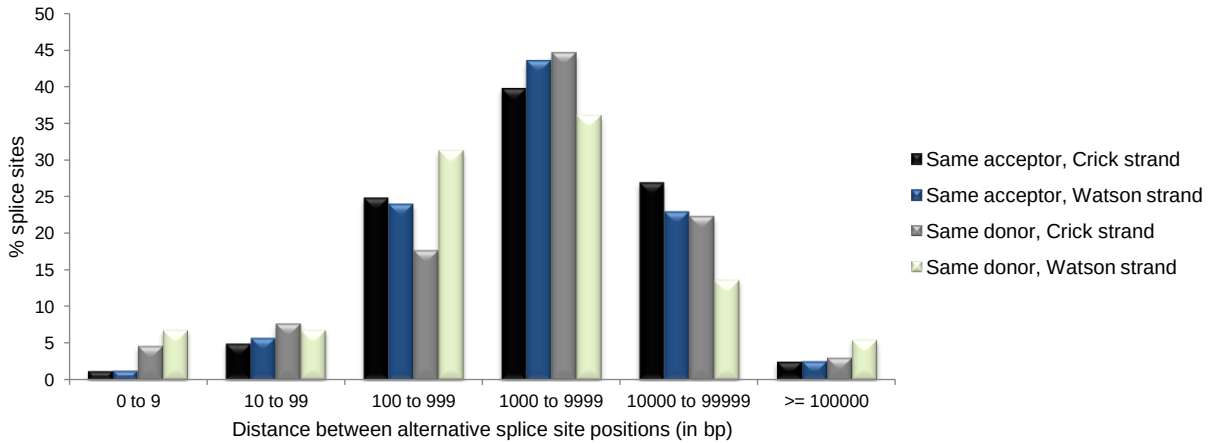


Figure 3-41 Distances between alternative splice sites

As last splicing type, the intron retention rate was investigated. In 71 genes, including the oncogene *Mecom* and the prostate cancer-associated gene *Rnasel*, at least one intron was more often retained in the tumor samples than in the controls. In 30 genes, including the oncogene *Kras*, at least one intron was more often spliced in the tumor than in the control samples. Intron read-through events were, amongst others, enriched in the processes 'Pathways of cancer' (KEGG: p = 0.041), 'RNA transport' (KEGG: p = 0.037) and 'Posttranscriptional regulation of gene expression' (GO: p = 0.05), while intron splicing occurred more often than expected in several metabolic pathways and in genes involved in 'Tight junction' functions (KEGG: p = 0.036) (Figure 3-42).

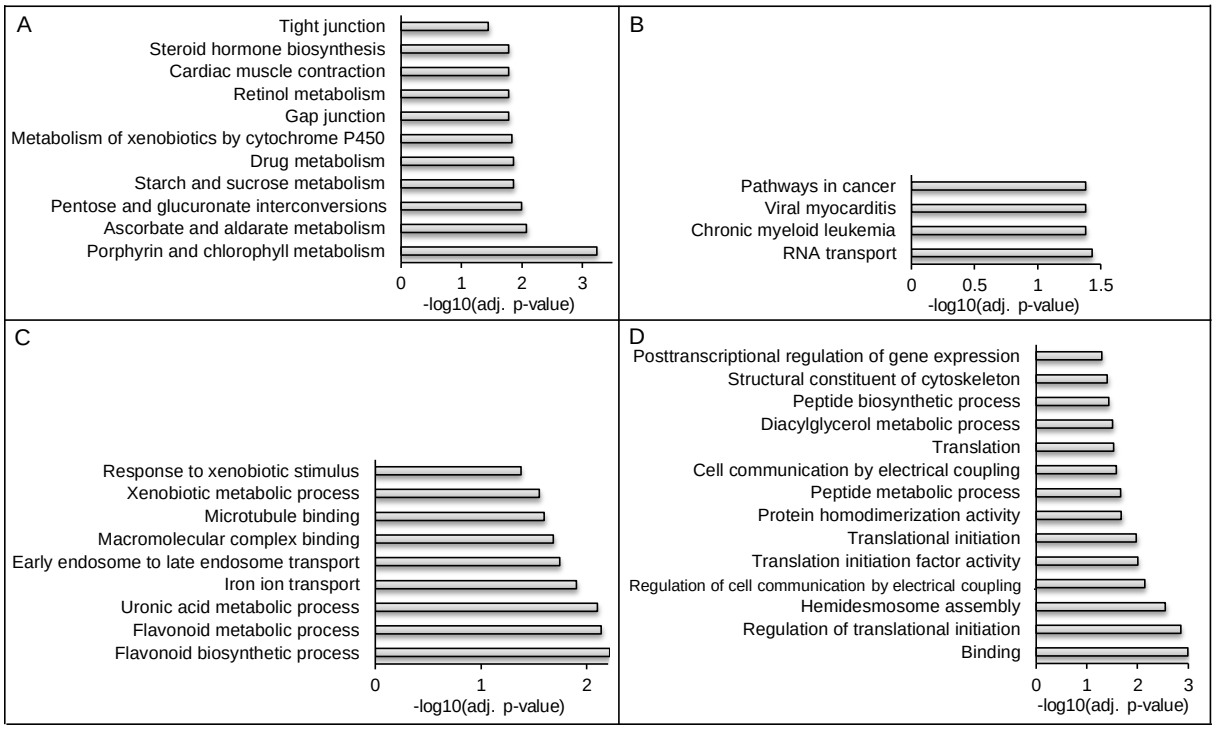


Figure 3-42 Processes enriched for genes with different intron retention rates in tumor samples
(A) KEGG pathways enriched for genes having at least one intron significantly more often spliced in the tumor than in the union of all non-tumor samples. (B) KEGG pathways enriched for genes having at least one intron significantly more often retained in the tumor than in the union of all non-tumor samples. (C) GO terms enriched for genes having at least one intron significantly more often spliced in the tumor than in the union of all non-tumor samples. (D) GO terms enriched for genes having at least one intron significantly more often retained in the tumor than in the union of all non-tumor samples. Only GO terms of the highest significant branch level are shown.

Interestingly, the GO term enrichment analysis of genes with altered intron retention rate resulted in a higher percentage of processes annotated in the first branch of the GO hierarchy tree than genes altered by another splicing type or differentially expressed genes. 60% of the processes enriched for genes with at least one intron exclusively spliced in the tumor samples and 82% of processes enriched for genes with at least one intron exclusively spliced in the controls were part of the first branch of the GO term tree, which harbors the most general processes. In contrast, a considerable lower percentage of processes enriched for alterations such as shift of the splice site or altered expression were located in the first branch of the GO term tree: shift in acceptor site position = 11%, shift in donor site position = 28%, downregulation = 36%, upregulation = 20%. Out of the processes enriched for genes with different exon usage, 50% were located in the first level branch (50% for exon skipping and 50% for exon retention in the tumor compared to control samples). This could indicate that genes with different intron retention rates were less specific assigned to a process. Thus, the regulation of this splicing type is less restrictive and directed.

3.2.4 Novel transcriptionally active regions in AOM/DSS-triggered colorectal cancer

At the time of my studies, no bioinformatic program existed, which were able to systematically investigate expressed transcripts, which were not annotated in the reference. To detect and characterize these so-called novel transcriptionally active regions (nTars) using RNA-Seq data and a genomic reference, I developed a novel tool during my studies. Besides giving information about the chromosomal coordinates of covered loci, this software reports connections via split reads or read pairs to annotated features or other nTars, correlations with expression values of annotated genes, the region in relation to annotated genes, neighboring genes incl. distances as well as several quality criteria such as coverage, number of nTar supporting reads with different start positions, mapping quality, and probabilities of mismappings. Further details are described in the methods section (chapter 2.2.9). As example, the program was applied to reveal nTars associated with AOM/DSS-triggered colorectal cancer.

In the AOM/DSS-associated colorectal cancer project of my studies, the number of covered regions, which were not annotated as exonic in the used reference, ranged from 702,926 to 902,991 per sample and were significantly higher in tumor samples compared to non-tumor samples (p(tumor vs. AOM/DSS proximal) = 0.00488, p(tumor vs. control) = 0.038)). In each sample, on average 7,708 nTars were characterized by a length of at least 50 bp and a mean coverage of 10x or higher. Although the highest amount of these nTars were located intronic (mean = 32.7%), exon-linked (mean = 26.8%) or intron-spanning (mean = 21.5%), these regions were not considered in further analyses to exclude a bias due to potential sequencing of pre-mRNA. However, differences between tumor and control samples in intronic regions

were covered by the splicing analyses described in chapter 3.2.3.6. Also in the filtered nTar subset, significantly more regions were detected in the tumor samples than in the proximal colon samples from AOM/DSS- and DSS-treated mice ($p = 1.134E-05$ and $p = 0.00550$, respectively) (Figure 3-43 A). In addition, a strong tendency was observed between tumor and control samples ($p = 0.0517$).

Despite the higher number of nTars in the tumor samples, the percentage of nTars having an ORF larger 50 amino acids and a start codon was significantly lower in the tumor samples than in the non-tumor samples (Figure 3-43 B). The same pattern was observed for nTars featuring at least one splice site (data not shown). In contrast, no significant differences were detected regarding the distribution of the region in relation to the applied gene annotation (Figure 3-43 C) and the percentage of nTars connected via split reads or read pairs to known genes (data not shown).

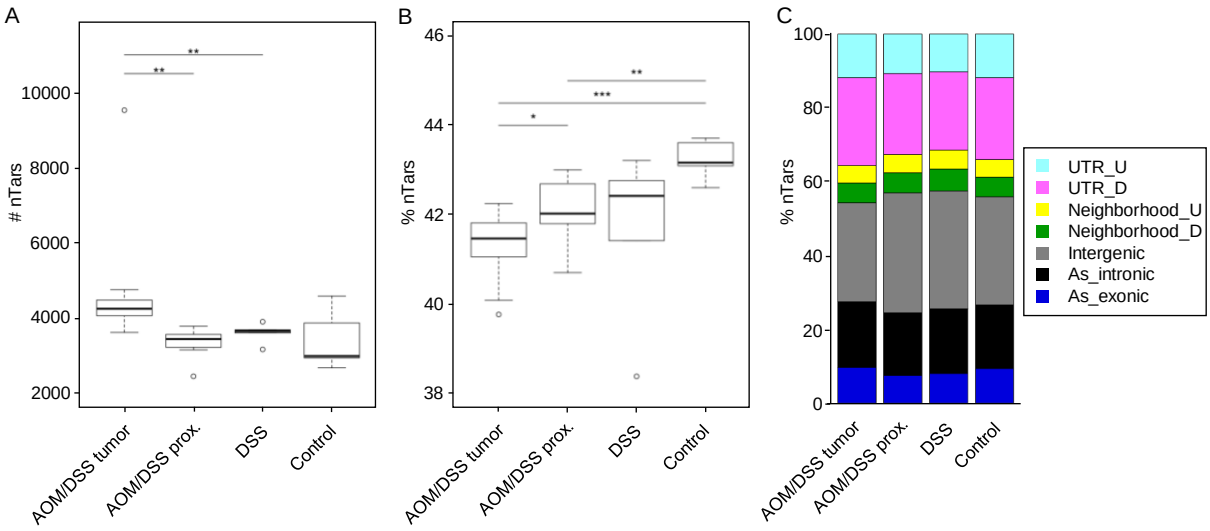


Figure 3-43 Description of detected nTars
(A) Number of detected nTars with a minimum length of 50 bp, a mean coverage larger 10x, and a location outside of an intronic region. (B) Percentage of nTars having an ORF larger 50 amino acids and a start codon sequence. (C) Region distributions of the detected filtered nTars (length > 50bp, minimum mean coverage > 10x, not intronic) in relation to the applied genome annotation.

The filtered nTar regions of all samples could be merged to 10,392 loci. An MDS-plot based on reads mapping to these regions clearly separated tumor from control and formerly inflamed tissue samples (Figure 3-44). 4,668 nTars were significantly differentially expressed between tumors and controls, of which the highest amount were located in intergenic regions (36.13%) followed by downstream (29.37%) and upstream UTR regions (11.02%). The lowest number of differentially expressed nTars were detected in the gene neighborhood (downstream = 7.67%, upstream = 5.50%). Even for the nTars, which were located antisense to an annotated exon (10.22%), several indicators suggested that a huge amount of these nTars were part of transcripts and not an artifact due to failed strandedness of some reads during the sequencing process: 94.07% harbored a reading frame larger 50 amino acids and 71.74% had additionally a start codon. Moreover, only 49.44% showed a significant positive

correlation with the corresponding exon on the sense strand, while ten nTars were even negatively correlated. The latter classification included the genes *Brsk1*, *Crym*, *Cyp3a13*, *Fam162a*, *Pes1* (two nTars), *Phgdh*, *Sac3d1*, *Sh2d7*, and *Sord*.

Besides differentially expressed nTars between tumor and controls, many nTars were also characterized by a different expression pattern between tumor and formerly inflamed samples as well as between formerly inflamed and control samples (Figure 3-44).

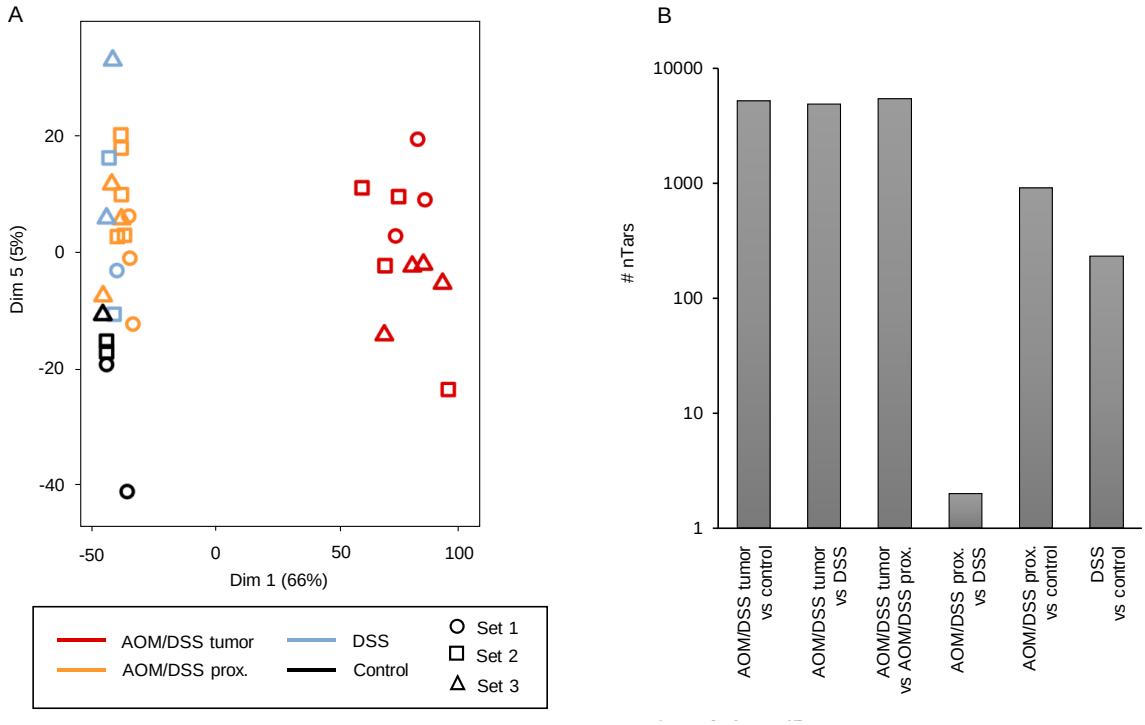


Figure 3-44 Comparison of nTars between sample groups
Both figures are based on all regions, which were not annotated as exon or intron, had a minimum covered length of 50 bp and harbored a mean coverage larger 10x. (A) MDS plot based on identified nTars. (B) Number of differentially expressed nTars based on pairwise comparisons between all treatment groups.

As example, the nTar shown in Figure 3-45 was differentially expressed between tumor and control samples (p = 4.65E-07). The nTar was 4,737 bp long, harbored an ORF of 174 amino acids and a splice site. This nTar was characterized by a significant negative correlation with the neighboring gene *Ucn2* (p = 0.0087, Spearman's rho = -0.48). *UCN2* is overexpressed in human CRC and was suggested to trigger inflammation, which contributes to CAC [212]. On the other hand, the nTar showed a positive correlation with the metabolic enzyme *Pfkfb4* (p = 3.27E-05, Spearman's rho = 0.69). *PFKFB4* was reported to be overexpressed in human aggressive metastatic tumors, while suppression of PFKFB4 reduces tumor growth and the formation of metastases [213]. However, metastasis and aggressive tumor growth are missing in AOM/DSS-induced tumors [50, 214], which might explain the lower expression of *Pfkfb4* in the tumor compared to the control samples. Interestingly, the nTar was also connected with *Pfkfb4* via split reads and read pairs in 30 out of 32 samples, indicating that the nTar might be an additional exon of one *Pfkfb4* isoform.

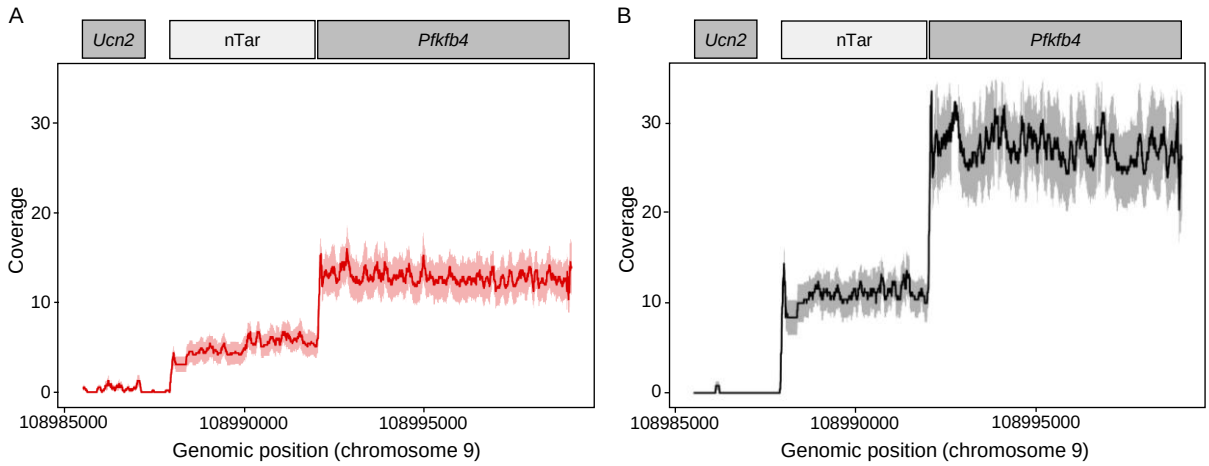


Figure 3-45 Example for one nTar
The figures illustrate the mean coverage including standard error deviation for (A) all tumor samples and (B) all control samples. The nTar position was located between the genes *Ucn2* and *Pfkfb4*. Both genes and the nTar were significantly differentially expressed between tumor and control samples. The nTar region was positively correlated with the *Pfkfb4* and negatively correlated with *Ucn2*.

Taken together, the developed tool enables to detect, filter, and characterize covered loci in genomic regions, which are not annotated yet. This is especially important to investigate species with an incomplete annotation, like non-model organisms, and to analyze phenotypes expressing genes, which are downregulated in already investigated control individuals. The described results demonstrate that even in the well-annotated murine system, a huge amount of novel potentially functional transcripts can be detected and associated with the development of AOM/DSS-triggered colorectal cancer.

3.2.5 Combination of data layers to reveal genes and processes potentially associated with AOM/DSS-induced cancer

The application of several data layers enables to gain a complete picture of a phenotype. Thus, WES and transcriptome sequencing results were combined to reveal novel insights into the molecular background of AOM/DSS-induced colorectal cancer. First, sample clustering was compared between WES and expression data to investigate whether general alteration differences were detectable between the sample groups. In a PCoA (Jaccard distance) based on all exonic SNVs and InDels (Figure 3-15 A), tumor samples from the first (high DSS) and second (medium DSS) sets were separated from the remaining samples. However, the third tumor set (low DSS), formerly inflamed, and control tissue samples were not distinguishable from each other. In contrast, a clear separation between all treatment groups was observable on expression level using MDS with Euclidean distance (Figure 3-32 B). Samples from the third set formed again a subcluster within the tumor samples. The distance measure methods, which were applied for the sample cluster analyses, changed due to different data types for WES (discrete) and transcriptome (continuous) results. These findings demonstrate that the step-wise progression of CAC is uniformly reflected on transcriptome level, although changes in the genome occur at an early stage.

In the next step, the expression level of genes, which were more often mutated in tumor than in control samples, were analyzed. Genes, which were unexpressed in all sample groups, were unlikely to contribute to the development of AOM/DSS-triggered colorectal cancer. For the following ten genes, no expression could be detected in at least ten samples: *4930467E23Rik*, *Epha10*, *Gabrg1*, *Magel2*, *Nlrp4e*, *Olfr143, Olfr1378*, *Pgr15l*, *Slc22a12*, and *Smok2a*. In contrast, genes, which were differentially mutated and differentially expressed between tumor and control samples at the same time, were especially likely to contribute to murine inflammation-associated colorectal cancer. Twelve genes were more often mutated in tumor than in the control samples and differentially expressed with a p-value lower 0.001 and a minimum fold change of four (Table 3-3).

	Gene	P-value	Adj. p-value	Log2(FC)
Downregulation	*Pld1*	4.532E-85	1.34E-82	-2.211
	Pclo	1.744E-20	2.23E-19	-2.731
	Nos1	3.584E-17	3.47E-16	-4.529
	Rimbp2	5.136E-17	4.92E-16	-3.059
	Kdm5a	1.198E-15	1.109E-14	-2.681
	Depdc7	1.405E-11	8.39E-11	-2.313
	Ncan	8.758E-05	2.38E-04	-2.512
Upregulation	*Odz4*	3.398E-81	8.641E-79	3.951
	Pcyt1b	1.241E-13	8.98E-13	2.278
	Neb	2.723E-08	1.16E-07	3.214
	Myh7b	1.401E-4	3.69E-04	3.021
	Erbb4	2.093E-4	5.35E-04	3.083

Table 3-3 Differentially expressed mutated genes
All genes listed in the table were more often mutated in tumor than in control samples and at the same time differentially expressed with $p < 0.001$ and fold change (FC) > 4.

In the next step, the protein-protein interaction network based on differentially mutated genes (Figure 3-28) was combined with the 250 differentially expressed genes with lowest p-value (Figure 3-46). The connections between the individual subnetworks were inferred by protein links annotated in the STRING database. 129 direct connections between genes of the WES and transcriptome subnetworks existed, while one gene (*Pld1*) was shared. Interestingly, more upregulated than downregulated genes were connected to the WES subnetwork ($p = 0.0159$). Also the average number of connections per node was significantly higher in upregulated than in downregulated genes ($p = 0.007183$). No difference regarding the number of connections was observed between up- and downregulated genes within the network of differentially expressed genes.

Genes involved in interactions between the two subnetworks were unequally distributed within each data set. In the WES results, *Ctnnb1* (13 connections) harbored clearly the highest number of connections followed by *Abl1* (11) and *Erbb4* (10). In the transcriptome data, hub genes connecting both subnetworks included *Nras* (12), *Ppara* (11), and *Mmp14* (7). Similar results could be found in a network combining mutated genes with all differentially expressed genes ($p<0.001$, fold change > 4). Thereby, in the network based on mutated genes, *Ctnnb1* (106 connections) harbored clearly the highest number of connections followed by *Nos1* (103),

Erbb4 (85), *Abl1* (84), and the additionally inserted not altered gene *Nfkb1* (79). In the transcriptome data, hub genes connecting both subnetworks included *Rac3* (19), *Tnf* (15), *Actc1* (14), *Acta2* (13), *Actg2* (12), and *Ppara* (11). These mentioned genes might play a central role in linking the different molecular layers.

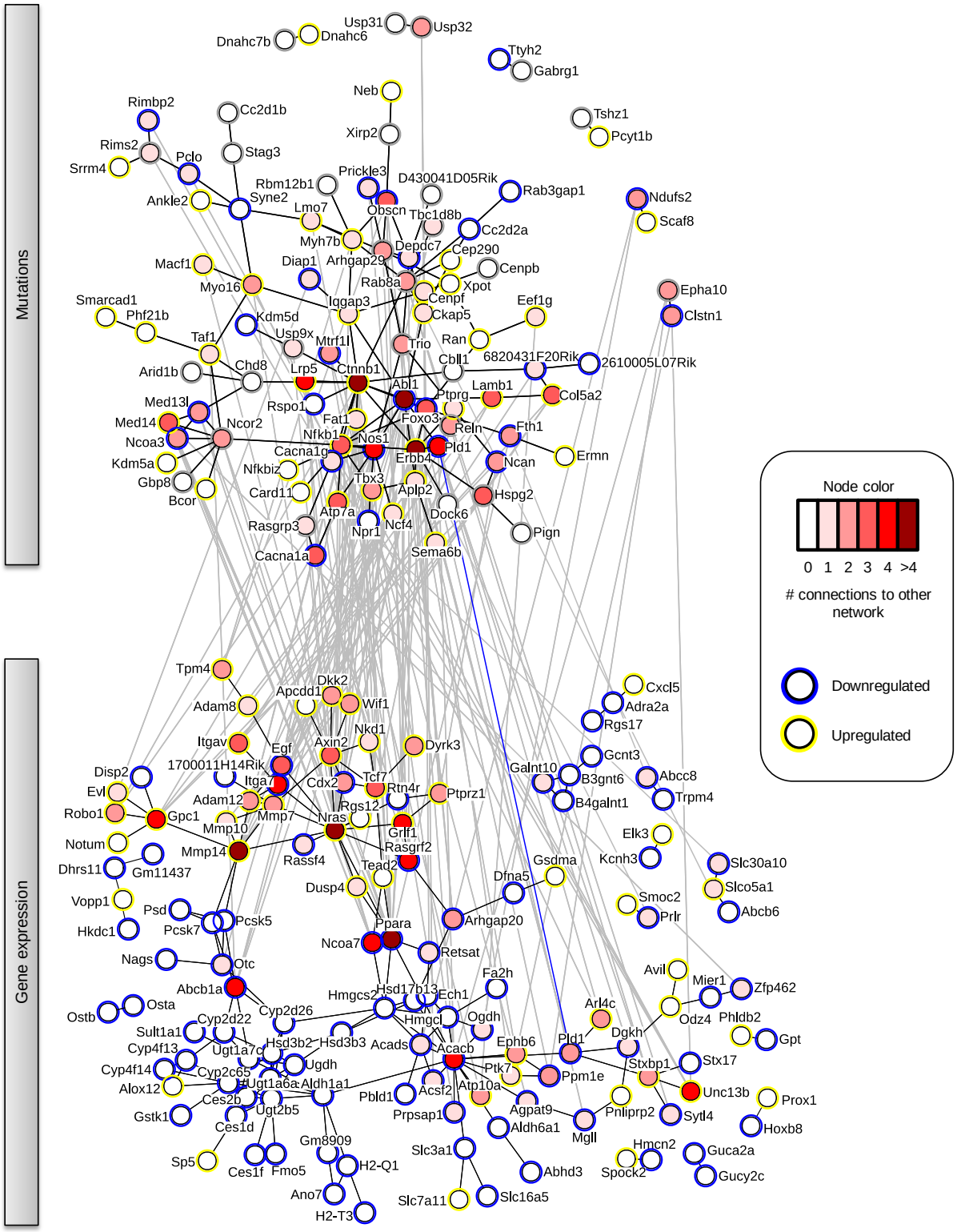


Figure 3-46 Protein-protein interaction network connecting WES and transcriptome results
The upper subnetwork displays the WES results as described in chapter 3.2.2.5, while the lower subnetwork is based on the 250 differentially expressed genes with lowest p-value. The grey lines indicate direct connections between two genes of different subnetworks. The blue line illustrates a gene existing in both subnetworks. The node color is based on the number of connections between the gene and the genes out of the other subnetwork. The color of the regulation direction is based on differentially expressed genes with p-value smaller 0.5.

As expansion of this protein-protein interaction network, the highest ranked genes affected by alternative splicing were added as an extra layer (Figure S 49). In the WES subnetwork, the three genes with the highest number of connections to the expression subnetwork shown in Figure 3-46 were also among the genes with the highest number of connections to the splicing subnetwork: *Abl1* (13), *Ctnnb1* (10), and *Erbb4* (10). Also in line with the previous network (Figure 3-46), *Ppara* and *Nras* out of the transcriptome subnetwork were highly connected with the splicing subnetwork (9 and 13 interactions, respectively). Other hubs connecting differentially expressed genes with the splicing subnetwork included *Tcf7 (5)*, *Mmp14* (5), and *Egf* (5). Thus, the highest number of connections to either WES or splicing harbored *Nras* (25), *Ppara* (20), and *Mmp14* (12). In the splicing subnetwork, *Kras* showed the highest number of connections to the WES (12) as well as to the transcriptome (14) subnetwork. Several hub genes detected in the described networks could be confirmed in a network combining mutated with all differentially spliced genes and all differentially expressed genes ($p<0.001$, fold change > 4). In this network, *Abl1* (65), *Ctnnb1* (57), *Nos1* (47), and *Erbb4* (47) harbored the highest number of connections from the WES subnetwork to the splicing subnetwork, while *Rac3* showed the highest number of interactions from the transcriptome to the WES as well as to the splicing subnetwork (19 and 74, respectively). Other hubs connecting differentially expressed genes with the splicing subnetwork included *Tnf* (63), *Acta2* (53), and *Actg2* (52). In the splicing subnetwork, *Mapk14*, *Kras*, and *App* showed the highest number of interactions with the WES or transcriptome subnetwork (153, 144, and 134, respectively). However, not all genes were equally involved in the connections between the data layers. As example, the gene *Gnas*, which was alternatively spliced in the tumor samples compared to the controls, harbored 50 connections to the expression subnetwork but none to the WES genes.

To summarize the results and to identify the most promising candidate genes and processes over all reported analyses, an integrated protein-protein interaction network was created. This comprised all differentially expressed genes ($p < 0.001$ and fold change > 4) and genes more often affected by a variant in the tumor samples than in the controls. In contrast to the comparison between the WES network and the transcriptome network, this protein-protein interaction network did not connect two independent data layers, but produced one network based on all affected genes (differentially expressed or differentially mutated) without considering the alteration type in the first step. Even in this plot, genes connected more often with genes of the same alteration type (mutated: $p = 5.159E-07$, downregulated: $p = 4.514E-13$, upregulated: $p = 2.269E-14$). Due to the high network density, two approaches were used to identify potentially relevant hubs: (i) In the first approach, the absolute cPI values of connected genes were summed up and plotted against the number of connected genes having at least two samples annotated in the COSMIC database with a mutation predicted as cancer driver (Figure 3-47). Amongst the top scored genes were *Ctnnb1*, *Tep1*, *Rac3*, *Abl1*,

Tnf, and *Erbb4*, which were all already known to be associated with cancer [215–221]. Genes such as *Fras1* and *Nos1* might be interesting novel candidates. The most relevant hub genes marked in Figure 3-47 A existed more often than expected in inflammatory and cancer-related pathways, such as ‘Colorectal cancer’ ($p = 0.0012$), ‘Wnt signaling pathway’ ($p = 0.0018$), and ‘mTor signaling pathway’ ($p = 0.0012$) (Figure 3-47 B). All genes potentially involved in AOM/DSS-associated colorectal cancer were highly connected among themselves (Figure 3-47 C).

In a second approach, each node of the overall protein-protein interaction network was first ranked according to its number of connected genes and subsequently to the number of connections to cancer-relevant genes as defined in the COSMIC database v79. Using a cutoff of 70 and 45 connections, respectively, the analysis resulted in 36 genes with high centrality. This was especially the case for *Tnf*, *Il6*, *Ctnnb1*, and *Fras1* (Figure S 50 A). The relatively higher number of connections of *Ctnnb1* within the investigated data set compared to the known cancer genes may point to a specific role of the gene in the tumorigenesis process in the AOM/DSS model. The highly connected genes within this network were enriched for several KEGG pathways related to inflammation and cancer, such as the ‘Wnt’ ($p = 0.027$), ‘MAPK’ ($p = 0.026$), ‘ErrbB’ ($p = 0.009$), ‘NOD-like receptor’ ($p = 0.00033$), and ‘Chemokine’ ($p = 0.00042$) signaling pathways as well as the ‘Colorectal cancer pathway’ ($p = 0.0045$) (Figure S 50 B). Interestingly, all filtered genes were again highly connected among themselves (Figure S 50 C), but formed two subclusters. Genes of the upper subnetwork were enriched for similar inflammatory and cancer-related processes like described for the enrichment analyses of all highly connected genes (Figure S 51). Out of the genes of the upper cluster, 95% were also detected with the first filter approach described above and were therefore even more likely to be involved in the development of AOM/DSS-triggered colorectal cancer. Genes of the lower subnetwork were enriched for pathways involved in metabolic and secretion processes (Figure S 51). Except *Gnat3* all genes of this second subcluster were exclusively detected with the filter approach based on the number of interactions and might therefore be less likely to contribute to AOM/DSS-associated colorectal cancer.

Next, differentially spliced genes were combined with differentially mutated and differentially expressed genes (Figure S 52 and Figure S 53) and both filter approaches described above repeated. The highest ranked genes based on the sum of cPI values and the number of connections to cancer-associated genes confirmed all genes, which were identified as most important in the protein-protein interaction network comprising only differentially mutated and differentially expressed genes. In addition, the genes *Kras*, *MAP14*, and *Rac1* were among the highest ranked genes and thus, potentially involved in the development of inflammation-triggered colorectal cancer.

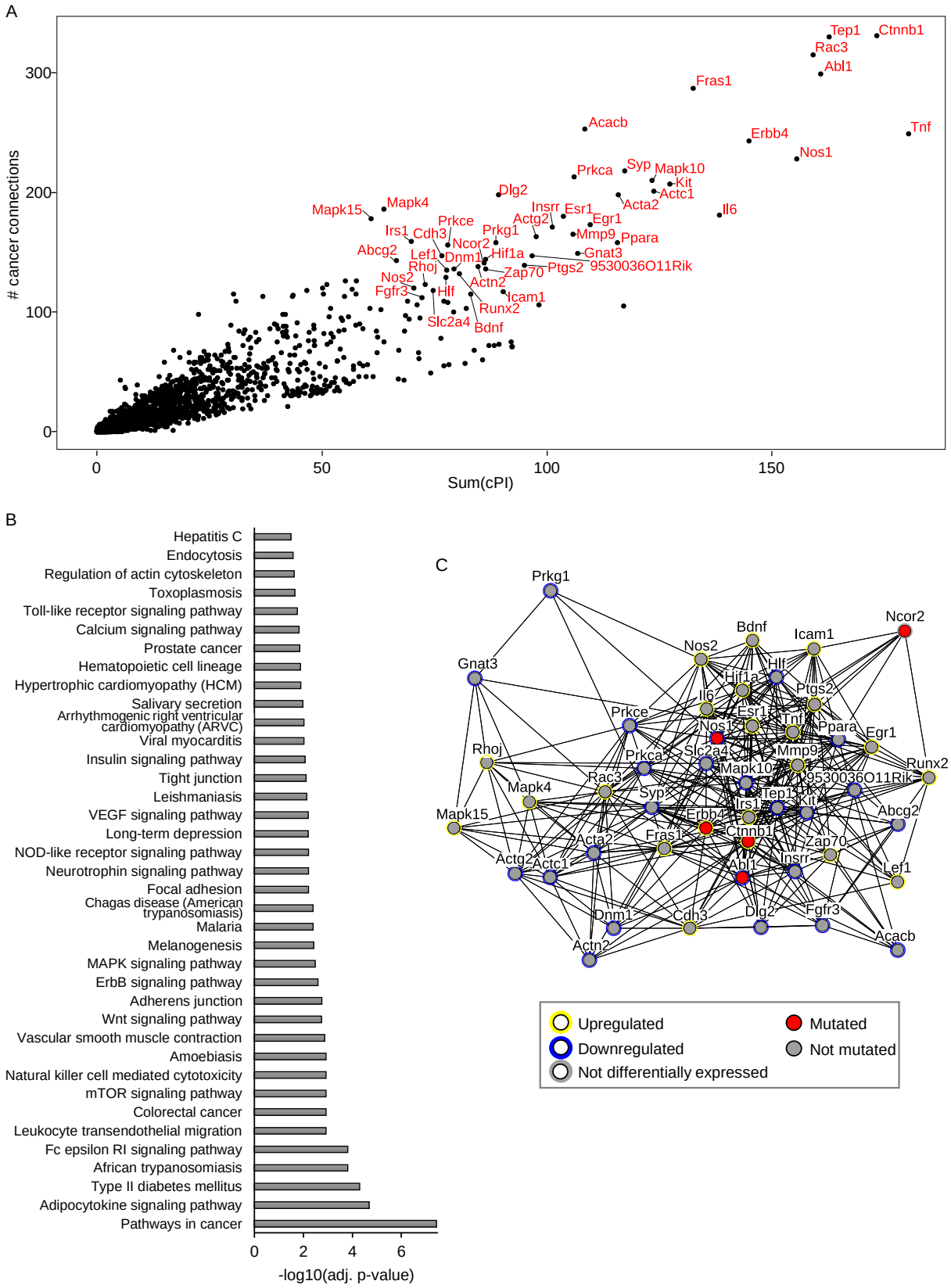


Figure 3-47 Genes with highest cPI sum and highest number of connected cancer genes in a protein-protein interaction network based on differentially mutated and differentially expressed genes

A protein-protein interaction network based on all genes, which were differentially mutated or differentially expressed between tumor and control samples, was created. (A) Genes with the highest sum of absolute cPI values of all interaction partners and the highest number of connections to cancer-related genes were filtered and labeled in red in the scatter plot. For this analysis, all interaction partners including those, which were not altered and therefore not part of the network, were considered. (B) KEGG pathways enriched for genes filtered in (A). Exclusively processes with at least two affected genes are shown. (C) Protein-protein interaction network based on highly connected genes labeled in red in (A). The colors of the nodes indicate the observed alterations.

4 Discussion

4.1 Gastric cancer

Until a few years ago, decision of treatment strategies for cancer patients were largely based on the anatomic tumor site. However, improvements of targeted therapies demonstrate that the response to treatments can even differ between tumors from the same anatomical origin [222]. Although this has led to major improvements in the treatments of lung and colon cancers, it is still in its infancies for the treatment of GC. Several NGS strategies have been exploited by e.g. the cancer consortia ICGC as well as TCGA to fill this gap and shed further light on the genetics of GC [3, 7]. Nevertheless, the molecular background of GC is still not completely understood. Thus, samples from sporadic gastric adenocarcinoma and matching controls were investigated using WGS and WES in the first part of my studies.

4.1.1 Pitfalls for application of NGS in the clinic

One aim of the study was to compare different sequencing tools, both from technological aspects (sequencing by synthesis vs. sequencing by ligation, whole genome vs. whole exome) as well as from the bioinformatic perspective in a small set of two GC samples. The results clearly demonstrate major pitfalls, which have to be overcome before such an in-depth sequencing analysis will become clinical reality. Starting from available clinical material, conducting such analyses will often be associated with problems of DNA quality (e.g. due to tumor necrosis), different sequencing depths, and problems of comparability between bioinformatical pipelines. As a relevant example, which has also been extensively been described by other groups in other cancer types [223, 224], it was here shown that the employed SNV callers led to various results and were characterized by particular strengths and weaknesses as demonstrated in the selected GC cases. The GC part of my studies also indicates a problematic recurrent false positive signals in both tumor samples derived from sequencing-by-ligation whole exome data, a technology that since the inception of the present study has nearly completely vanished from the market. This led to several failed validation attempts for SNVs exclusively detected in the WES data in the current study. Also other studies have reported a high false-positive rate in WES data [225]. However, in my studies, an even higher amount of SNVs detected in the WGS data could not be confirmed with the WES approach demonstrating the crucial importance of validating the variant calls with a second sequencing technology. Here, it was shown that - given all the limitations of the study - the combination of two sequencing data sets delivers reliable results even in a scenario where only relatively low coverage data is available. In a clinical setting, however, besides precision, also cost effectiveness and duration of analyses play major roles in the decision for the applied technology. Thus, ideally WGS as a single technology encompassing the entire variant

spectrum has to be applied. Besides actionable recurrent SNVs (e.g. within p53), this covers also structural variants such as large deletions, large insertions, and gene fusions. In reality, even the most sophisticated high-resolution WGS still has a way to go to reach that aim. Until then targeted analyses will clearly have their place in a molecular pathology setting.

4.1.2 Identified alterations consistent with previous large studies

All of the genes discussed in this section were among the highest ranked genes in my studies. This means that the results of the current study suggested a role in GC for these genes and that the variants were likely no sequencing artifacts or passenger alterations. Despite the small sample size, many results of this study were in line with other larger studies. As example, in the gene *PIK3CA* that encodes for an oncogenic Phosphoinositide-3-kinase, a somatic SNV was identified in the GC MSI tumor investigated during the course of my studies. Also in a comprehensive molecular characterization of 295 primary gastric adenocarcinomas, which belongs to the TCGA project, a high mutation frequency in the *PIK3CA* gene (42%) was reported for MSI tumors [7]. Additionally, a somatic SNP (rs28934574) was detected in the tumor suppressor gene *TP53* in the MSS tumor sample, which is also in line with the description from the TCGA that *TP53* is often altered in chromosomal unstable tumors [7]. Furthermore, a somatic large deletion was located in the cyclin gene *CCND3* in the MSI tumor sample from the first patient. *CCND3* is involved in the G_1-S phase progression and alterations were observed in previous studies on e.g. breast cancer [226]. High alteration rates found in the cell cycle mediators (e.g. *CCNE1, CCND1* and *CDK6*) in the TCGA project [7] were also confirmed in my studies. Moreover, in the MSI tumor, a somatic variant was detected in *JAK2*, which encodes for a receptor tyrosine kinase. Recurrent alterations of this gene were also reported in a previous GC study of the TCGA [7]. However, in contrast to the results in my studies, these were found in tumors, which were positive for Epstein-Barr virus.

ARID1A was affected by a somatic frameshift InDel in the MSI tumor of the first patient. ARID1A is involved in chromatin-remodeling, has an antiproliferative effect and was validated as novel tumor suppressor gene by Zang et al. [174], who carried out WES on tumor samples of 15 GC patients (11 intestinal, three diffuse and one mixed type according to Laurén) followed by a prevalence screening. An *ARID1A* gene mutation was present in nine out of 110 GC patients with tumors characterized by MSI and *PIK3CA* variants.

Several genes were affected by somatic structural variants in the MSI tumor sample investigated in my studies: *EPHB3* (somatic non-frameshift deletion), *GATA4* (somatic non-frameshift deletion), *PDE4D* (somatic interchromosomal translocation, somatic large deletion), and *PARK2* (somatic large deletion). Alterations in these genes are in line with results revealed by Dulak et al. [178] using high-density genomic profiling of 486 gastrointestinal adenocarcinomas of different anatomical origin.

Recently, WGS and comprehensive molecular profiling of 100 tumor-normal pairs of GC patients identified genetic and epigenetic alterations as well as mutational signatures, which were specific for each subtype [57]. Mutated genes reported in this study include, amongst others, *TP53*, *ARID1A*, *IRS2*, and *SUPT3H*, which were all also altered by somatic variants in my GC studies. In addition, Wang et al. suggested a high heterogeneity of GC [57], which was also observed in the results of the current study.

Even starting from fewer patients compared to other projects, further gene mutations potentially involved in the development of GC were revealed in the current study. Although an association with GC was not reported for the affected genes so far, these genes merit further attention. For example, the second patient (MSS tumor) harbored somatic SNVs in *GATA3*. The GATA family contains transcription factors that are important for the regulation of specification and differentiation of various tissue types through controlling gene transcription, proliferation, and cell survival [177]. Although *GATA3* has not been linked to GC yet, a connection between other GATA factors and GC was shown by focal amplifications of *GATA4* and *GATA6* in GC [178]. A potential role of GATA3 in GC is further supported by the finding that GATA3 is involved in breast cancer metastasis [177].

A multitude of other alterations potentially associated with GC were detected in the investigated samples: *RERE* (somatic SNV in MSI tumor) plays an important role in the induction of apoptosis and might contribute to the regulation of cell survival [227], while *PRDM2* (somatic SNV in MSI tumor) and *BRAF* (somatic frameshift InDel in MSI tumor) were already reported to be associated with other cancer types than GC [66, 228]. Moreover, non-synonymous alterations in highly conserved genes with tumor suppressor activity were, amongst others, identified in *PCGF2* (somatic SNV in MSS tumor) [229], *MLL3* (associated with GC, somatic interchromosomal translocation in MSS tumor) [174, 230], and *MCC* (somatic interchromosomal translocation in MSS tumor) [190].

Besides alterations on gene level, several molecular processes were predicted to be involved in the development of GC in my studies. As example, the process ‘Negative regulation of the viral reproduction’ harbored more genes affected by a variant than expected by chance. Viral replication is usually inhibited before any damage is caused. However, tumor cells were shown to have a higher susceptibility to viruses, which might be influenced by defects in the antiviral innate immune response [231]. Furthermore, the process ‘Negative regulation of transposition’ was more often affected by variants than expected in the investigated tumor samples. Also previous studies have demonstrated a link between transposition and cancer development [232]. Malignant processes in the cell can cause the activation of mobile elements, which can also trigger an elevated mutation and recombination frequency in the genome [233]. Both traits were also observed in the tumor samples of the current study. Additionally, an enrichment of variants was observed in the process ‘DNA cytosine deamination’, which might be linked to

the observed increase in the number of somatic C>T base substitutions in both investigated tumor samples. The dominance of this SNV type within the GC mutation spectrum was also reported in previous GC studies [54, 174] and contributed to the slightly increased transition/transversion rate in the investigated tumor samples compared to samples from the 1000 Genomes Project [148]. Furthermore, in both tumor types, an elevated relative number of somatic C>A SNVs was observed in the current study. This might be caused by reactive oxygen species (ROS), which are significantly increased in GC [234, 235]. In contrast to the MSI tumor, somatic T>G substitutions occurred with a higher frequency in the MSS tumor of the second patient than in the controls. These characteristics were also reported by Wang et al. [57] as well as the TCGA consortium [7] and could indicate that this tumor type is caused by another mechanism or results from an aberrant activity of the error-prone polymerase η [110].

4.1.3 Conclusion of the gastric cancer study

A comprehensive characterization of the molecular background of one microsatellite stable as well as one microsatellite unstable gastric carcinoma was performed using WES and WGS. The combination of two different NGS technologies as well as the applied multi-tool approach clearly increased both the number of detected variants and the reliability of the results. Moreover, the use of population-based resources enabled not only a confirmation of recent findings, but revealed also a multitude of novel somatic potentially damaging variants. Thereby, it was shown that MSS and MSI GCs harbored markedly different mutational patterns. Of uttermost importance, the large variety of distinct genetic alterations might influence therapy response and the patients' outcome. It is thus tempting to speculate that signatures identified by cancer genome sequencing could reflect a potential future biomarker for therapy stratification. In particular, the current study clearly demonstrated the value of individual WGS to depict the multidimensional aberrations in GC. Thereby, the suggested procedure can be used as benchmark setting in clinical applications with a low number of sample pairs. However, some challenges (e.g. functional annotation in the light of high genomic diversity, sample heterogeneity, and rapid turnaround times) must be solved before genome-driven therapy stratification will become clinical reality.

4.2 Molecular characterization of murine AOM/DSS-induced cancer

The AOM/DSS mouse model [53] is a widely used method to study inflammation-associated colorectal cancer. However, the knowledge about the molecular background of this mouse model was up to now very limited and the mutational landscape of AOM/DSS-induced colorectal cancer has not been investigated systematically on a genome-wide level yet. In the current study, WES and transcriptome sequencing of colonic tissues and tumors from

AOM/DSS-treated C57BL/6N mice was applied to gain novel insights into genetic signatures and potential mechanisms behind malignant transformation. Advantages of the applied AOM/DSS model primarily include both high reproducibility and potency. More so, this method includes a short latency time for the induction of inflammation-triggered cancer and resembles the multistep process observed in human CAC, in which normal crypts form foci of aberrant crypts followed by the development of adenocarcinoma due to proliferation and crypt fission [48]. In human UC-CAC, tumors are mainly located in the sigmoid colon / rectum (63-64%) followed by 20-29% in the transverse / descending colon, and 8-16% in the ascending colon / cecum [44, 236]. This distribution is similar to the observed tumor sites in the current study: 59% of the clearly visible tumors (no dysplasia) were located in the sigmoid colon / rectum, while 41% were in the transverse / descending colon. Of particular note, molecular signatures differ between distinct tumor locations and have been described for the first time in my studies (discussed in section 4.2.3).

Colitis severity can be determined by investigating stool consistency and bleeding [237]. These factors in combination with weight loss were included in the DAI and increased together with both tumor incidence and sizes with higher DSS doses in the current study. This confirms that the applied AOM/DSS model is, like human CAC [35], dependent on the strength of the inflammatory stimulus. This observation is supported by a reported synergic effect of AOM and DSS leading to elevated tumor formation and reduced latency period compared to models using only AOM or DSS treatment [48, 49]. In previous studies, it was demonstrated that higher DSS concentrations contribute to elevated expression of proinflammatory cytokines and a higher amount of crypt damage [238]. In turn, this inflammation can cause a tumor-prone microenvironment, influence the number of aberrant crypt foci, enhance the growth of dysplastic crypts, and promote the transformation to adenocarcinomas [48, 49, 68]. The connection between DSS dose and tumor development clearly demonstrates that the applied AOM/DSS model leads towards inflammation-triggered carcinogenesis. However, in contrast to human CAC, murine AOM/DSS-induced tumors rarely develop metastases or mucosal invasiveness [50]. These differences were also reflected on molecular level in the current study (discussed in section 4.2.5) and suggest that the AOM/DSS model might not completely reflect human CAC [61].

4.2.1 Molecular changes best visible on transcriptome level in tumors induced by high DSS doses

On variant level, only tumor samples from mice treated with a high or medium DSS dose were distinguishable from all other sample types, while a well-defined separation between control, formerly inflamed, and tumor tissue samples was visible on transcriptome level. This observation demonstrates that the molecular signatures of inflammation-triggered colorectal

tumors are distinct from those of formerly inflamed colorectal tissues. Although the step-wise progression of AOM/DSS-induced cancer is first uniformly reflected on transcriptome level, causative changes in the genome might occur initially as mutations. Reasons for the earlier visible effect on transcriptome than on variant level can be manifold. As response to changing environmental conditions, cells can adapt gene regulation as short-term reaction [239]. However, long-term changes in the expression patterns can affect the susceptibility to various diseases [240] and influence genomic alterations by e.g. downregulation of mismatch and recombinational break repair processes [239]. These mechanisms potentially support adaptive evolution by modifying the genome, so that it is potentially more suitable for the novel conditions [239]. Additionally, epigenetic changes, which also regulate gene transcription, can support localized alterations in the DNA sequence by deamination of 5′meC resulting in a C>T base substitution. These modifications lead to similar functional effects as caused by the applied epigenetic mechanisms and thus, manifest an epigenetic adjustment as a genetic alteration [241]. These examples demonstrate that short-term adaptation through transcriptomic changes can cause long-term adaptive mechanisms such as induction of DNA variants [239, 241]. In turn, genetic components can have an impact on transcript levels. Genetic variants contribute to 20-65% of the expression pattern in a cell and thus, some variants might substantially influence mRNA levels of a large proportion of genes [242]. A further reason for earlier visible overall changes on transcriptome level might be that AOM methylates guanine in an initial step [51]. These alterations may directly influence the expression level of e.g. genes involved in repair mechanisms, while base substitutions due to mispairing occur at a later point in time.

In samples from mice treated with a low DSS dose, the variant distributions of tumors received from the same animal were more similar than those of tumors received from different animals. In contrast, in the first and second treatment sets (high and medium DSS dose, respectively), the distances based on the mutational spectrums between tumors received from the same mouse were similar to the distances between tumors received from different mice. These results demonstrates for the first time that tumors induced by AOM/DSS grow independently of each other and thus, AOM/DSS-induced colorectal cancer is a multilocular process. In tumors from mice treated with low DSS dose, which harbored a lower number of variants than tumors of animals treated with a high DSS dose, this effect was superimposed by a cancer-independent variant distribution. The unbiased formation of base substitutions is supported by the analyses of the SNV patterns, which are discussed in the next chapter (chapter 4.2.2).

4.2.2 Mutational landscape of somatic SNVs in AOM/DSS-induced colorectal cancer characterized by random variant distribution

Mutational signatures of tumors are influenced by mutagen exposure, tumor type, and potential defects in DNA repair mechanisms [5, 54, 153]. Analyses of variant patterns can deliver information about underlying mutational processes and thus, reveal conclusions for cancer etiology as well as for potentially beneficial interventions to prevent or treat the disease [5]. However, SNV patterns were neither reported for human CAC nor for murine inflammation-triggered colorectal cancer yet. To gain further insights into the underlying processes of inflammation-associated tumor development and to allow the comparability between the murine AOM/DSS model and human CAC, differences between the SNV patterns of AOM/DSS-induced tumors and those of non-tumor samples were analyzed. Additionally, somatic variants were compared with SNVs patterns occurring due to evolutionary processes, which were visible as differences between the two mouse strains C57BL/6N and C57BL/6J.

A higher number of SNVs was observed in tumor compared to non-tumor samples in the current study. Thereby, the mutation rate increased with DSS dose. This can be explained by differences in the grade of inflammation between the three treatment sets (discussed in chapter 4.2.1) and is in line with previous observations [243, 244]. Higher number of mutations in the tumor samples might be due to an adaptation to changing conditions in the tumor environment, which can be caused by downregulation of DNA repair genes as well as the occurrence of hypoxia resulting in a higher amount of genetic instability, DNA damage, and increased repair of DNA double-strand breaks via the error-prone non-homologous end joining mechanism [245].

Compared to controls, a higher ratio of C>T base substitutions was observed in the tumor samples of the current study. This could be a consequence of AOM exposure [246]. One of the DNA adducts formed by AOM after several steps is O^6-methylguanine (O^6-meG), which leads by mispairing with thymine to a G>A base substitution (C>T on the complementary strand) during DNA replication [54, 246]. The C>T ratio was especially high in samples from mice treated with high DSS dose indicating that AOM has a stronger impact on base substitutions in combination with DSS. Besides AOM, the tumor type itself might contribute to the enrichment of C>T substitutions. This would be in line with the increased number of C>T SNVs observed in colorectal cancer in previous studies [54]. Moreover, in the comparison between the two mouse strains C57BL/6N and C57BL/6J, which reflects evolutionary influences, an increased amount of germline C>T base substitutions was visible. However, while in the investigated tumor samples, the somatic C>T SNVs were independent of the sequence context, germline C>T base substitutions were predominantly located at cytosines with adjacent guanine in the non-tumor samples. These germline differences are likely caused

by a high cytosine methylation rate at CpG islands, which results in an increased susceptibility to cytosine to thymine conversions via spontaneous deamination of 5'meC [241].

The increase of C>A substitutions observed in the post-inflamed samples derived from DSS- and AOM/DSS-treated mice in the current study might be a result of a higher level of oxidative stress, which is linked to chronic inflammation and multiple cancer types including CAC [28, 69]. One main product arising from oxidative DNA damage is 8-hydroxyguanine, which induces G>T mutations (C>A on the complementary strand) [114]. This mechanism might also contribute to the more equal distribution of the sequence context of C>D (D = A or G or T) substitutions observed in the investigated tumor samples. However, the number of C>A base substitutions caused by oxidative damage were probably superimposed by the effect of O^6-meG produced by AOM, why the relative proportion of somatic C>A substitutions decreased with increasing DSS dose (Figure S 31). Thus, a lower percentage of somatic C>A SNVs was observed in tumor samples from mice treated with medium or high DSS dose in the current study, although the real count might be still higher than in the third set (low DSS).

Somatic SNVs did not form mutational clusters in the investigated tumor samples pointing towards a missing *kataegis* effect in inflammation-triggered colorectal cancer. However, this needs to be verified by WGS to consider also intergenic regions, which were not covered by the applied WES. While missing *kataegis* events would be in line with acute myeloid leukemia and pilocytic astrocytoma, it would be in contrast to cancers of the breast, pancreas, lung, and liver [5].

No strand preference was observed for somatic SNVs in the tumor samples investigated in my studies. In contrast, genes on the Watson strand harbored more genomic differences between the mouse strains C57BL/6N and C57BL/6J than genes on the Crick strand. The lack of a strand bias of somatic SNVs is in line with MSS, but not MSI cancer types, which are characterized by asymmetric mutation rates between the leading (Crick) and lagging (Watson) strand [247]. An unequal distribution of variants between lagging and leading strands were also reported for bacteria: In these species, most genes, which are important for fitness and which are therefore under strong negative selection against mutations, are located on the leading strand of replication to avoid conflicts between replication and transcription machineries [196, 198]. These collisions can disturb gene expression and increase the amount of variants in many species including eukaryotes [197, 198]. It was proposed that higher rates of mutagenesis on the lagging strand guide evolutionary processes by fostering alterations in specific genes [197]. Thereby, a switch in gene orientation could be applied to select genes for rapid evolution [196, 197]. In addition, mRNA values were significantly higher for genes on the leading than for genes on the lagging strand in the investigated samples of the current study. Higher expression can result in a lower mutation rate (see next paragraph) and might contribute to the observed unequal distribution of the germline SNV types between the two

DNA strands. In contrast, somatic SNVs in the tumor samples were likely induced by other mechanisms, because they occurred independently of DNA strand and gene expression.

Somatic SNVs were equally distributed among genes of distinct expression classes in the tumor samples. In contrast, the amount of genomic differences between the mouse strains C57BL/6N and C57BL/6J was lower in highly expressed genes in my studies. The expression dependency is influenced by transcription-coupled nucleotide excision repair, which is a mechanism to fix genomic defects on the transcribed strand of expressed genes [6, 248]. In addition, alterations are more likely to occur in tightly packed DNA which corresponds to genes with a lower expression level [249]. However, SNV occurrence in AOM/DSS-triggered colorectal cancer seem to be independent from these mechanisms. This missing expression dependency is in contrast to several other cancer types, in which somatic variants arise predominantly in lowly expressed genes [6, 248].

As last feature of the variant pattern, no differences of somatic SNVs were observed between coding and transcribed strand in the investigated tumor samples. In contrast, a different mutation type distribution between the sense and the antisense strand of coding genes was detected for germline SNVs in the non-tumor samples. The strand bias during evolution is likely coupled with the transcription of genes. Frequent expression can for example increase the effects of single-strand deamination of cytosine on the sense strand during transcription [250]. In addition, the antisense strand is highly preferred by transcription-coupled repair, which also contributes to an asymmetric mutation type distribution between the strands [250]. As indicated above, this result demonstrates that the development of mutations is independent of gene transcription in AOM/DSS-triggered colorectal cancer and thus might point to a defective DNA repair mechanism in the tumor samples.

Taken together, all investigated parameters, except the substitution types, were characterized by a random distribution of variants in the tumor samples of the current inflammation-triggered colorectal cancer study. While this observation is in line with the SNV patterns observed in acute myeloid leukemia and pilocytic astrocytoma [5], it is in contrast to human CRC, which is characterized by strand biased SNVs due to a proofreading deficient DNA polymerase [251], and cancers of the breast, pancreas, liver, and lung, for which small foci of *kataegis* have been reported [5]. Even mutational signatures of germline variants, which accumulate during evolution, are usually unevenly distributed to, amongst others, select specific genes for adaptation [6, 54, 197, 248]. This indicates that variants in AOM/DSS-induced tumors arise not due to a directed adaptive process to changing environmental conditions, but result from a combination of treatment and tumor type specific processes. However, it cannot be excluded that the strong influence of the mutagen may superimpose other selective signatures, which may arise from tumor initiation and progression in the AOM/DSS mouse model.

4.2.3 Similarities between murine AOM/DSS-induced colorectal cancer and human CAC

In the current study, several approaches, including an integrative analysis based on WES and transcriptome data, were applied. The results enabled to identify commonalities as well as discrepancies between human CAC and the murine mouse model and suggested several genes and processes potentially associated with inflammation-triggered colorectal cancer. In this section, molecular features that are shared between the AOM/DSS mouse model and human CAC are described. First, alterations on gene level will be discussed followed by the consideration of cellular processes. This structure ensures to include also processes, which are altered in all tumor samples but differ by the sets of affected genes between the samples. Vice versa, investigations on gene level can reveal modified genes, even if the complete process is not affected. In addition, some genes might be involved in several pathways. The last part of this section includes molecular characteristics that differ between tumors at distinct colorectal localizations. These features were investigated for the first time in my studies.

Several gene modifications previously reported in the context of human CAC were also observed in the investigated AOM/DSS mouse model. However, not all of them were ranked as most promising alterations according to the applied filter criteria in my studies. Shared molecular changes included upregulation of genes encoding for the cyclin-dependent kinase inhibitor CDKN1C [60], the proinflammatory cytokine IL23 [66, 252], the transcription factor NF-κB, which is associated with immune and tumorigenesis processes [252], the signal transducer and transcription factor STAT3 [66, 252], and the cyclin-dependent kinase inhibitor CDKN2A (P16) [253]. Further examples for commonalities between human CAC and the applied mouse model were overexpression of the proinflammatory cytokine *IL17* [66, 252] as well as alterations of *LRP5*, which encodes for the low-density lipoprotein receptor-related protein that belongs to the Wnt pathway and is involved in the transfer of intracellular signals [60].

Moreover, an elevated mRNA level was observed for *Hsp90ab1* (*Hsp90*) in the investigated murine tumor samples as well as in human CAC [76]. The encoded chaperone is involved in the stabilization and control of key signaling proteins, including oncogenes, cell cycle regulators, proteins important for immune responses, and transcription factors [254]. The expression level of HSP90 correlates with the progression of CAC and the proliferative possibilities of cancer cells [76, 255]. Moreover, upregulation of *HSP90AB1* might also contribute to the lack of therapeutic response in certain cancer types [255]. Thus, specific chaperone activity inhibitors are currently evaluated for the treatment of different cancer types and inflammatory diseases in clinical and preclinical studies, respectively [254].

Tnf, which encodes for a proinflammatory cytokine, was highly upregulated, among the highest ranked genes in the performed network analyses, and thus a potential key player in AOM/DSS-induced colorectal cancer. An overexpression of *TNF* was also reported for human CAC [220]. However, while a previous mouse study demonstrated that inhibition of Tnf can reduce tumor development after treatment with AOM/DSS [72], TNF-α antagonists were shown to have no influence on the development of CAC in IBD patients [256].

One promising gene potentially associated with AOM/DSS-induced colorectal cancer was the mitogen-activated protein kinase 14 (*Mapk14*, protein: p38α). In the current study, *MAPK14* was characterized by aberrant splicing in the tumor cells, featured a high number of connected cancer or altered genes, and a high sum of cPI values of connected genes. In human CAC, differentially expression of MAPK14 was observed instead of alternative splicing [257]. However, both alteration types might cause the same effect. The protein p38α is a key regulator of the intestinal homeostasis and involved in the inflammatory response as well as in the regulation of the expression level of pro-survival and angiogenic cytokines [257–259]. Before tumor formation, inactivation of p38α leads to a defect epithelial barrier function and as consequence, to a higher amount of epithelial damage and an increased cancer risk [259]. In contrast to this tumor suppressor activity, p38α might be involved in oncogenic processes and regulation of tumorigenesis in cancer cells [259]. In particular, p38α contributes to numerous factors involved in cancer development at the cellular levels, such as cell metabolism, cell survival, invasion, metastasis, angiogenesis, inflammation, and chemoresistance of colorectal cancer cells and tissues [257, 258]. Thus, p38α is currently tested as therapeutic drug target for several cancer types including CAC [257, 260].

One striking regulator of murine inflammation-triggered colorectal cancer was the MAPK signaling pathway in the investigated samples. This process was enriched for hub genes in the protein-protein interaction network combining WES and transcriptome data as well as for differentially mutated genes. In addition, an altered expression was observed between tumors from mice treated with different DSS doses. In patients suffering from IBD or CAC, ERK/MAPK signaling is triggered by proinflammatory cytokines, while downregulation of the ERK MAPK process leads to a reduction of proliferation and an increase of apoptosis events in IECs in the tumor microenvironment [261]. Moreover, mutations in the MAPK pathway have been found in several human cancer types [262]. Besides the role in proliferation, the ERK/MAPK pathway is involved in cancer cell migration, cell survival, malignant invasion, angiogenesis, and tumor-immune system interactions [262]. In addition, therapeutic response is influenced by the interaction between the ERK/MAPK pathway and the Wnt signaling pathway [262], which was also one of the most striking processes in the investigated tumor samples and is described in the next paragraph.

The Wnt signaling pathway was characterized by an early upregulation in the investigated tumor samples and was enriched for hub genes of the combined protein-protein interaction network based on WES and transcriptome data in my studies. Wnt signaling is a key player in the maintenance of the intestinal epithelium, controls proliferation in normal and cancer cells [263], regulates cell differentiation and migration, while playing a role in polarity and asymmetric cell division [264]. Additionally, Wnt signaling can contribute to cancer cell growth and tumor metabolism [265]. Although mutations in Wnt pathway components occur first at a late stage in CAC [263], activation of Wnt signaling was, like observed in the investigated murine tumor samples, described as an early event in human CAC [65, 266]. It was proposed that upregulation of Wnt signaling is triggered by inflammation and repair mechanisms to increase the level of proliferation of intestinal stem cells. This might counteract inflammation. However, activation of Wnt signaling can also lead to dysplasia [66, 264]. Thus, it was suggested that Wnt signaling plays also a role in the initiation of the malignant transformation from colitis to cancer [264, 266]. Due to the additional contribution of Wnt signaling to the renewal of cancer stem cells and an increased tumor heterogeneity, Wnt signaling has been suggested as biomarker for CAC [266, 267], while inhibition of this process was proposed as potential therapy to suppress the development of CAC in IBD patients [266].

Several processes that are involved in the organization of the cell structure and adhesion were modified in the investigated tumor samples. As example, the KEGG pathway ‘Regulation of actin cytoskeleton’ was enriched for altered genes in different analyses of the current study. This process plays a central role in the regulation of cell movement [268]. Alterations of the actin cytoskeleton organization are involved in the transition from colitis to dysplasia as well as in tumor development including invasion and metastasis in CAC [268]. Also the molecular processes ‘Focal adhesion’ (aberrant spliced, differentially mutated genes), and ‘Adherens junction’ (aberrant spliced) were enriched for altered genes in the current study. Hints for modifications of these two processes were previously also reported for human CAC [269, 270]. Mediated by actin cytoskeleton molecules, intercellular and cell-matrix adhesion molecules control cell polarity, differentiation, proliferation, migration as well as invasion and are therefore relevant for the colonic epithelium homeostasis [271]. Dysfunction or loss of this interaction influence the intestinal epithelial structure and thus promote the formation and progression of colorectal cancer [271].

The molecular processes ‘Bacterial invasion of epithelial cells’ (aberrant spliced) and ‘Tight junctions’ (differentially mutated genes, aberrant spliced, among the highest ranked genes in the combined protein-protein interaction network) were both altered in the investigated tumor samples. Besides other functions, tight junctions are important for the maintenance of the intestinal homeostasis by avoiding attachment and infiltration of bacteria into the lamina propria via actin-rich microvilli [272]. As in the investigated murine tumor samples, tight junction genes

are altered in human CAC [269]. This points to an association of inflammation-triggered colorectal cancer with the invasion of bacteria into the inner mucus layer and a subsequent activation of the immune system. In line with these features, microbial dysbiosis is associated with the development of IBD and colorectal cancer [269, 273, 274]. Thereby, the bacterial composition differs even between human CRC and human CAC. As an example, a higher abundance of the Enterobacteriacae family and the Sphingomonas genus as well as lower counts of the Fusobacterium and the Ruminococcus genera were observed in CAC compared to CRC [275].

Several metabolic pathways were modified by downregulation, aberrant splicing or a high mutation rate in the investigated tumor samples from AOM/DSS treated mice. This included Glycolysis / Gluconeogenesis, amino sugar and nucleotide sugar metabolism, drug metabolism, citrate cycle (TCA cycle), starch and sucrose metabolism, ribosome biogenesis in eukaryotes as well as pentose and glucoronate interconversions. This is in line with other studies indicating alterations of metabolic processes in CAC [274, 276]. Otto Warburg postulated already a link between metabolism and cancer in 1956 [277]. To support proliferation processes, a shift from adenosine triphosphate (ATP) generation through oxidative phosphorylation to ATP production via aerobic glycolysis exists in cancer cells [2, 278, 279]. Further reported bioenergetics alterations of cancer cells are stimulations of the amino acid metabolism as well as the pentose phosphate pathway [279]. Intermediate metabolic products, for example ROS causing cell lesions, and persistent aerobic glycolysis are associated with the activation of oncogenes and loss of tumor suppressors [278]. Thus, reconstruction of metabolic pathways can lead to accelerated cell growth, angiogenesis, rapid proliferation, tumor cell survival, invasion, metastasis, and resistance to cancer treatment [279]. As consequence, metabolic reprogramming plays an important role during malignant transformation as well as tumorigenesis [278, 280] and is nowadays considered as hallmark of cancer [2]. The highly flexible metabolic processes enable tumor cells to adapt to new nutritional states in the microenvironment, which can influence the treatment response as well as the etiopathology (e.g. metastasis). Therefore, it is likely that anti-cancer metabolic drugs support therapies in future [279]. However, for the application of these drugs, it is necessary to consider cancer type specific differences [279] such as driver mutations, tissue of origin, nutrient availability, inflammatory cells, and cell-autonomous alterations caused by clonal expansion of mutants [280]. These factors can alter metabolic preferences and flexibility, why a more detailed analyses of the reported altered metabolic pathways in inflammation-induced colorectal cancer would be necessary.

As discussed in section 4.2, the tumor localization distribution observed in murine AOM/DSS-induced colorectal cancer in the current study was similar to those reported for human CAC [236]. In human CRC, the colorectal section of the tumor is associated with

molecular clinicopathological characteristics such as the 5-year survival rate, patient's age, and histological features [281]. In particular, in comparison to tumors in the distal colon or rectum, proximally situated tumors harbor a significantly larger tumor size, a higher tumor grade, a higher chemoresistance rate, and a higher risk of death and disease progression in human CRC [282, 283]. Thus, it was proposed to classify human CRC into right-sided colon cancer, left-sided colon cancer, and rectal cancer [281]. The described clinical differences were also reflected on a molecular level in the investigated tumor samples from AOM/DSS-treated mice in the current study. As an example, genes associated with different diseases indicating an altered immune system as well as obviously cancer-related processes, such as "Pathways in cancer", showed the highest expression level in tumors from the middle part of the colon followed by rectal located tumors. Since also the investigated tumors of the middle part of the colon were bigger in size, the current data point for the first time towards a defined cancerogenic molecular signature significantly associated with tumor location in inflammation-triggered colorectal cancer. However, it needs to be mentioned that the reported study is slightly biased, because all sequenced tumor samples from mice treated with a low DSS dose were located in the distal part of the colon, while all sequenced tumor samples from mice treated with high DSS dose were situated more proximal. Only samples from mice treated with medium DSS dose were selected from both positions, but for this group only one tumor transcriptome per mouse was sequenced.

In my studies, also genes involved in metabolic processes were significantly differentially expressed between tumors located in distinct colorectal sections, although, like described above, reconstruction of the metabolism is one hallmark of cancer [2]. This result indicates that also the interaction with the environment might be affected. Indeed, the composition of the tumor-associated microbiome depends on the tumor location [284].

Molecular processes with lowest expression in tumors located in the rectum followed by the expression level in tumors formed in the central part of the colon included 'Ubiquitin mediated proteolysis' and 'Progesterone-mediated oocyte maturation'. Both pathways exhibited the highest expression level in distal situated tumors in the current data. These processes were already linked to CRC or inflammatory response in previous studies [285–288]. In humans, 'Progesterone-mediated oocyte maturation' is especially striking in stage III CRC, but differentially expressed genes were also reported for samples from UC and early-stage CRC patients [289]. Moreover, this process showed an enrichment tendency for genes with altered expression level between wild type and interleukin 10 gene deficient mice, which is applied as mouse model of human IBD [288]. The exact role of the process 'Progesterone-mediated oocyte maturation' in tumorigenesis and inflammation is completely unknown and an association with the tumor localization reported for the first time in my studies. Proteolysis via the ubiquitin system is involved in colorectal tumorigenesis and important for cellular processes

like apoptosis, angiogenesis, differentiation, proliferation, cell cycle, and modulation of the immune and inflammatory responses [285, 287]. Thus, proteasome inhibitors were suggested to support existing therapies [285, 287]. Although this clearly points to a role of ubiquitin mediated proteolysis in inflammation-triggered colorectal cancer, a connection between tumor localization and regulation of this process was up to my knowledge observed for the first time.

Taken together, many molecular characteristics reported for human CAC were also reflected in the AOM/DSS mouse model. Besides alterations of genes and processes, this includes the distribution of tumor localizations. However, specific molecular features of a tumor site were described for the first time in my studies. As discussed in the next chapter, further genomic alterations are very likely to be shared by human CAC and the AOM/DSS mouse model.

4.2.4 Genes and processes potentially involved in inflammation-associated colorectal cancer

All genes and processes, which were among the highest ranked molecular features in the performed analyses of the current study, were associated with cancer or inflammation. However, a role in the development of human CAC is not known for all of them so far. In this section, genes and processes are discussed, which are likely to be involved in the development and progression of human CAC as well as in murine AOM/DSS-triggered colorectal cancer.

As an example, *Rac3* encoding for a GTPase was upregulated and among the highest ranked genes in the combined analysis of WES and transcriptome data. Indeed, it was previously reported that Rac (Rac1, Rac2, and Rac3) inhibition can counteract AOM/DSS-mediated colorectal cancer and DSS-triggered colitis in mice [290]. Moreover, *RAC3* is overexpressed in several human cancer types, such as aggressive breast carcinoma as well as prostate and brain cancers, and involved in metastatic processes [217]. Although alterations of *RAC3* were up to my knowledge not reported for human CAC yet and *RAC3* is not upregulated in UC patients, RAC3 was, together with RAC1 and RAC2, suggested as therapeutic target for the treatment of CAC [290].

Moreover, *Il6* encoding for a cytokine was among the highest ranked genes in the combined analysis of WES and transcriptome data in the current study. Il6 is a pro-tumorigenic cytokine that has an impact on epithelial and immune cells. Previous studies have shown that the protein is important at an early tumor stage in murine inflammation-triggered colorectal cancer, while anti-IL-6 receptor antibody treatment reduces the tumor burden in AOM/DSS-treated mice [291, 292]. Also in humans, the IL6 level was altered in several cancer studies, which included breast, prostate, lung, liver, and colon cancer patients [292, 293]. Thus, IL6 was suggested as target to treat human CAC [291, 292], although the role of IL6 in human CAC was up to my knowledge not revealed yet.

Foxo3 encoding for a transcription factor was downregulated and characterized by an especially high SNV density in three tumor samples in my studies. In mammals, FOXO factors are potential tumor suppressors that can regulate cell cycle arrest, DNA repair, and apoptosis [294]. Besides these functions, FOXO proteins might influence cancer development by interactions with oncogenes and tumor suppressors. Active Foxo3 counteracts tumorigenicity in nude mice, while in human patients with breast cancer, a shift of the FOXO3 localization from the nucleus to the cytoplasma is associated with decreased chances of survival. [294]. It is further known that several variants in *FOXO3* are strongly linked to longevity [295] and might therefore increase the survival of cancer cells. Although a role in CAC was not reported for FOXO3 yet, follow-up studies on this gene seem to be promising.

A striking gene within the investigated WES data was *Nfκbiζ*. In the current study, *Nfκbiζ* was characterized by the highest percentage of directly connected genes with a proinflammatory functional annotation (Figure S 42). The encoded protein NFKBIZ plays a role in inflammatory processes. In particular, the protein can regulate the central inflammatory protein NF-κB, activate natural killer cells, and recruit monocytes [296, 297]. Moreover, *Nfκbiζ* has been reported to be mutated in familial colorectal cancer [297] and is altered in IBD patients [296]. But although NFKBIZ was linked to cancer and inflammation by the mentioned studies, the precise role of NFKBIZ in human CAC remains largely unknown and has to be investigated in further functional studies.

Besides the well-known CAC-associated genes and processes described in section 4.2.3, *Acacb* was downregulated as well as a hub gene in the protein-protein interaction network based on transcriptome data in my studies and is thus likely to be involved in inflammation-triggered colorectal cancer. In humans, ACACB plays a role in obesity as well as several diseases, such as the metabolic syndrome and diabetes [298, 299]. In human CRC, *ACACB* is downregulated and contributes to therapy resistance against the drug cetuximab [299]. However, overexpression of *Acacb* leads to an elevated expression of proinflammatory cytokines, while downregulation of *Acacb* opposes this effect [298]. Due to the contrary effect in CRC and inflammation, it is difficult to draw conclusions to the behavior of ACACB in CAC. Nevertheless, it would be worth to investigate the function of ACACB in human CAC.

In my studies, *Ppara* was downregulated in the tumor samples, among the 60 differentially expressed genes with lowest p-value, and characterized as hub gene in the protein-protein interaction network based on transcriptome data. Moreover, *Ppara* harbored a high number of connections from the transcriptome subnetwork to the WES and the splicing subnetwork. PPARA is a regulator of the lipid and glucose metabolism as well as a key player in the maintenance of the gut mucosal immunity and the microbiome homeostasis [300]. Although a role of PPARA in human CAC was not reported so far, PPARA is known to suppress inflammatory and tumorigenesis processes [300–302]. As consequence, downregulation of

Ppara can lead to a higher abundance of inflammatory cytokines and an increased risk of inflammation in the intestine [300], while activation of *Ppara* is associated with less inflammation in the colon [301]. In addition, PPARA was mostly reported to act as tumor suppressor, although some previous studies are contrary [302].

A further striking gene was *Fras1*, which was upregulated and a hub gene within the protein-protein interaction network based on transcriptome data as well as in the protein-protein interaction network combining WES and transcriptome data. *FRAS1* encodes for an extracellular matrix (ECM) molecule and is overexpressed in several cancer types [303, 304]. Moreover, FRAS1 contributes to migration and invasion in cancer cell lines [303]. However, a role in inflammation or CAC is not known so far.

Based on the results of my studies, also *Erbb4* might be associated with inflammation-triggered colorectal cancer. In the investigated tumor samples, *Erbb4* was upregulated and differentially mutated in the investigated tumor samples, harbored a high sum of cPI values of connected genes and was highly connected with other genes including other potential cancer-associated genes, such as *Ctnnb1*, *Nos1*, and *Foxo3*. ERBB4 is a receptor tyrosine kinase and involved in the control of cell growth, proliferation, survival, and differentiation [221, 305]. A chronic high expression level of *ERBB4* can trigger cellular transformations and inhibition of apoptosis [221, 305]. As consequence, ERBB4 is involved in the development of cancer, while malignant transformation is reduced by suppression of ERBB4 in human CRC. Besides the upregulation of *ERBB4* in CRC, *ERBB4* is overexpressed in IBD [221, 305]. Thus, ERBB4 was proposed as therapeutic drug target for the treatment of CRC, although an association between ERBB4 and human CAC was not reported so far [221, 305].

Tep1 (*Pten*) was downregulated in the investigated tumor samples compared to controls and among the highest ranked genes in the protein-protein interaction network based on transcriptome data as well as in the network combining differentially mutated and differentially expressed genes. The encoded protein Tep1 harbors a protein and a lipid phosphatase activity. Using the lipid phosphatase activity, Tep1 is involved in cell survival and can act as tumor suppressor by inhibiting the PI3K-AKT-mTOR pathway. The protein phosphatase activity impacts cell cycle arrest and suppression of cell invasion [216]. In human sporadic cancer types including colon cancer, *TEP1* is one of the tumor suppressor genes, which are most commonly altered or affected by loss of function [216]. Besides the importance for tumor development, TEP1 has a key function in maintaining the balance between of pro- and anti-inflammatory processes [306] and might therefore be also altered in human CAC.

Besides genes potentially associated with inflammation-triggered colorectal cancer, the mTor (mammalian target of rapamycin) signaling pathway is worth mentioning. In the current study, this process was enriched for genes, which were more often mutated in tumor than in control

samples, and for high ranked genes in the protein-protein interaction network combining WES and transcriptome data. The mTor signaling pathway is an important component to regulate cell metabolism, cell cycle, proliferation, cell survival, stress response based on transcription, and the actin cytoskeleton [307]. Moreover, dysregulation of mTor signaling is associated with colon cancer and influences cancer-associated processes, such as tumor growth, angiogenesis, and metastasis [307–309]. Rapamycin, an inhibitor of mTOR, is already applied for the treatment of human colon cancer [309]. The mTor process is also upregulated in IBD patients, but a connection with human CAC was not reported so far [307–309].

Taken together, a clearly inflammatory and cancer promoting molecular signature was identified in the tumor samples from AOM/DSS-treated mice in my studies. Besides several genes and processes, which are known to be associated with human CAC (chapter 4.2.3), many molecular alterations are likely to be common between the human disease and the AOM/DSS mouse model. This assumption is based on a high ranking position in one of the performed analyses and an already reported function in cancer or inflammation for each gene or process. However, as described in the next section, some alterations are also contrary between human CAC and the AOM/DSS mouse model.

4.2.5 Differences between murine AOM/DSS-induced cancer and human CAC

Although human CAC and murine AOM/DSS-induced colorectal cancer share several characteristics such as inflammation-driven tumor development, some typical features of human CAC, including metastasizes or mucosal invasiveness, are lacking in the mouse model [50, 214]. These differences were also reflected on a molecular level in my studies. In particular, several alterations in genes and pathways, which were linked to human CAC in published studies, could not be confirmed or were even contrary in the investigated AOM/DSS mouse model. However, most of the results need to be proven in further experiments to exclude differences due to study design, applied technologies, or bioinformatic analysis types.

As example, mutations in *DPC4* (=*SMAD4*), *SOX9*, *EP300*, *BRAF*, *IL16*, *NRG1*, and *APC* as well as upregulation of *STK39* and inactivation of *DCC* by epigenetic changes, were reported for human CAC [59, 60, 66, 264], but neither the current data nor data of other studies investigating murine AOM/DSS-triggered colorectal cancer delivered any support for these alterations [62–64]. Most of these genes are known to be involved in invasive and metastatic processes in cancer. Thus, the molecular differences might explain the phenotypic discrepancies between human CAC and the AOM/DSS mouse model. As example, reduced activity of the tumor suppressor DCP4 leads to higher invasion and metastasis potential [310, 311]. However, instead of mutations like in human CAC [59], an upregulation of *Dpc4* was observed in the investigated tumor samples from AOM/DSS treated mice. Similar, the transcription factor Sox9 inhibits the formation of metastasis [312] and is characterized by a

high mutation frequency in human CAC [59], while an overexpression was observed in the murine mouse model in my studies. The metastatic capacities are also increased by lack of the tumor suppressor DCC and the histone acetyltransferase EP300 [313, 314]. These genes are altered in human CAC [59, 264] but not in the investigated tumor samples of the AOM/DSS mouse model. Also mutations in the serine/threonine protein kinase *Braf*, which contributes to invasion and metastasis in CRC [315], could not be confirmed with the AOM/DSS model in the current study. Another example is the serine/threonine kinase encoding gene *STK39.* While no alterations were detected in the tumor samples from AOM/DSS-treated mice in my studies, upregulation of *STK39* is associated with human CAC [60]. The expression level of *STK39* correlates with tumor progression including metastasis in other cancer types, such as non-small cell type lung cancer [316]. Among alterations, which were observed in human CAC, were also variants in the cytokine encoding gene *IL16* [59]. In other tumor types, it was shown that *IL16* overexpression triggers invasiveness and cell migration, while IL16 neutralizing antibodies lead to a lower amount of metastases [317]. In contrast, neither mutations nor an altered expression level was detected in tumor samples from AOM/DSS treated mice in my studies. Moreover, variants in the ERBB ligand *NRG1*, which can feature oncogenic characteristics after rearrangements, were observed in human CAC [59, 318]. For other cancer types, such as adenocarcinoma of the lung, *NRG1* rearrangements and fusions could be associated with the invasive potential of the cells [318]. In contrast, an upregulation of *NRG1* was observed in the investigated tumors from AOM/DSS treated mice. Also alterations in *Apc* were missing in my studies, although APC mutations occur in human CAC, albeit first at a late tumor stage [263]. APC is a tumor suppressor involved in repression of nearly all colorectal tumorigenesis steps, including invasion and metastasis [319].

Another noticeable gene is *Nos1*. *Nos1* was downregulated and differentially mutated in the investigated tumor samples compared to controls and was characterized as hub gene in the protein-protein interaction network based on WES data as well as in the protein-protein interaction network based on differentially expressed genes. Nos1 (nNos) is the largest producer of nitric oxide (NO) in an individual [320]. Besides the importance for the maintenance of the homeostasis, it also contributes to the development of cancer [320]. In particular, elevated NO production contributes to invasion, migration, and metastasis in tumor cells [321]. In line with these characteristics, *NOS1* is upregulated in cancer cells [320]. Although the *NOS1* expression level was not investigated in human CAC yet, the results observed in my studies seem to counteract aggressive tumor growth in the AOM/DSS mouse model.

The most central gene in the protein-protein interaction network based on WES data of the current study was *Ctnnb1*, which was additionally upregulated in the tumor samples. Interestingly, all SNVs in the tumor samples affected one of the codons 32-34, 37, 41 or 45, while a single synonymous SNV in a proximal non-tumor sample of an AOM/DSS-treated

mouse was identified in codon 63 in my studies. It was also demonstrated in previous studies that only mutations in the mentioned codons of the tumor samples (32-34, 37, 41, and 45) seem to be involved in the inactivation of Ctnnb1 in murine AOM/DSS-induced cancer [49]. Mutations in *CTNNB1* were also frequently observed in many human cancer types. Similar to the mouse model, most of the base substitutions occur at one of six positions in the third exon [215]. In contrast, mutations in the third exon of *CTNNB1* have been suggested to occur rarely during the progression of human CAC [322]. *Ctnnb1* encodes for the β-catenin protein, which contributes to cell-cell-adhesions processes and is involved in the transport of extracellular signals [215]. In a complex with E-cadherin, β-catenin is able to suppress tumor cell invasion and metastasis [215]. Mutations in *Ctnnb1* might cause a permanent active Wnt / β-catenin signaling pathway. In turn, this can increase cell proliferation and might trigger the development and progression of several cancer types, including gastrointestinal cancer [215]. Wnt signaling pathway is also of crucial importance for processes contributing to human CAC [65, 266]. Thus, other components of the pathway are likely to be responsible for the altered function of Wnt signaling in human CAC. For example, the Wnt signaling pathway can alternatively be activated by dysfunction of the tumor suppressor p53, which is a central upstream regulator of β-catenin [215].

It is especially notable that no mutations were detected in the tumor suppressor gene *Trp53* (human: *P53*) in the investigated samples, although *P53* mutations arise very frequently at an early stage in human CAC or even already in inflamed tissues [263]. The own observation is supported by other studies reporting a lack of *Trp53* alterations in murine AOM-induced tumors [61, 214]. Mutations in *P53* can not only lead to a loss of the tumor suppressor function but also to novel oncogenic characteristics supporting a more aggressive tumor type including metastasis [323]. Thus, Cooks et al. suggested DSS treatment of mice with a gain of function mutation in *Trp53* as mouse model for human CAC [68]. In contrast, an upregulation of non-mutated *Trp53* was detected in the tumor samples of AOM/DSS treated mice in my studies, although human CAC studies suggested reactivation of P53 with original tumor suppressor function as anti-cancer therapy [323].

In my studies, the oncogene *Kras* was noticed due to alternative splicing and a hub gene position in the protein-protein interaction network combining WES data with differentially expressed and alternative spliced genes. No somatic mutation was detected in *Kras* in the tumor samples, which is in line with a very low *Kras* mutation prevalence reported in previous studies investigating AOM/DSS-induced colorectal cancer [61]. In contrast, *Kras* mutations are associated with human CAC albeit with a lower frequency than in CRC [324]. Besides involvement in tumor initiation and progression, activating mutations in *KRAS* cause invasion and metastasis in CRC [325]. The splicing pattern was neither systematically investigated in human CAC nor in the mouse model so far.

Besides alterations occurring either exclusively in human CAC or exclusively in murine AOM/DSS-triggered colorectal cancer, several observations were even contrary between the human disease and the mouse model. Also these differences might contribute to the development of invasion and metastasis. As an example, *Lrp6* was found to be significantly downregulated in the investigated tumor samples compared to the controls, while an upregulation of *LRP6* is associated with the development of human CAC [60]. In colorectal cancer, invasion and metastasis is triggered by LRP6 [326]. Furthermore, *Hk1* was more often mutated and significantly lower expressed in tumor than in control samples in my studies. Although, no alterations were reported for human CAC so far, the observations are contrary to other cancer types in the gastrointestinal tract. *Hk1* encodes for the hexokinases 1, which stimulates the glucose intake and glucose phosphorylation to glucose-6-phosphate [327]. Tumor cells feature highly glycolytic characteristics and thus overexpress *HK1* in many cancer types [327]. Upregulation of *HK1* is also associated with an increased metastasis rate [327]. Moreover, in the current study, the kinase *Abl1* was downregulated and mutated in the tumor samples from AOM/DSS treated mice, among the highest ranked genes in the analysis combining WES and transcriptome results, and harbored a hub gene position in the protein-protein interaction network based on WES data. Although it is unknown whether the oncogene ABL1 is involved in the development of human CAC, the observed results are contrary to alterations reported for *ABL1* in other cancer and inflammatory diseases, in which *ABL1* is mostly overexpressed. In solid tumors, oxidative stress and decreased availability of nutrients can activate ABL1 and in turn, ABL1 might trigger invasion and metastasis [218, 219].

As last example, *Rac1* was downregulated and aberrantly spliced in the investigated tumor samples. Although *RAC1* was not reported to be differentially expressed in human CAC so far [59], the regulation is contrary to the function of RAC1 in other inflammatory and cancer diseases. Rac1 belongs to the Rho family of GTPases and regulates intracellular signaling pathways, is involved in the immune response, cell growth, cell cycle regulation, adhesion to the extracellular matrix, and reconstruction of the actin cytoskeleton [328]. In humans, *RAC1* overexpression is associated with IBD [290, 329] and plays a role in the malignant transformation and progression of several cancer types including colorectal adenocarcinoma. In particular, hyperactivation of RAC1 leads to accelerated tumor development including higher metastasis rate, invasion, higher mortality, more aggressive tumor cell growth, and poor differentiation [330, 331]. Therefore, RAC1 inhibition has been suggested as a promising application in the treatment of different cancer types including colorectal adenocarcinoma [330, 331]. Even in AOM/DSS treated mice, Rac1 hyperactivation leads to an increased risk for colitis development, while decreased Rac activity (Rac1, Rac2, and Rac3) causes eased colitis and reduced tumor formation [290, 329]. The discrepancy between the *Rac1* expression level in murine AOM/DSS-induced colorectal cancer and its function in inflammatory and cancer

diseases might be a further reason for missing metastasis and invasion in the AOM/DSS mouse model.

Besides molecular differences that might explain the lack of metastasis in the mouse model, genes contributing to the malignant transformation seem to be distinct between human CAC and murine AOM/DSS-triggered colorectal cancer. Watanabe et al. [60] suggested 40 genes that enable to predict whether a patient with UC will get cancer. Out of these genes, 29 were investigated in my studies. Although 17 genes reported by Watanabe et al. were differentially expressed between tumor samples from AOM/DSS treated mice and controls in my studies, none of them were involved in the malignant transformation and only the expression of the gene *Sla* was associated with the tumor stage. This clearly demonstrates that not all biomarkers are identical between human CAC and the AOM/DSS mouse model.

In addition to differences on gene level, also processes were differentially altered between the AOM/DSS mouse model and human CAC. One example for this discrepancy is the Notch signaling pathway, which was enriched for genes, which were more often mutated in tumor than in control samples in the current study. Notch signaling is conserved across species and involved in angiogenesis as well as in the maintenance and development of normal intestinal epithelia including regulation of cell fate determination, proliferation, cell survival, and differentiation of colonic goblet cell and stem cell populations [332–334]. Previous mouse studies have shown a downregulation of Notch signaling in AOM/DSS-induced cancer [335]. Although little is known about Notch signaling in human CAC, this is in contrast to human CRC. Here, overexpression of Notch signaling is associated with CRC as well as with a higher risk for the development of colorectal cancer in precancerous conditions [333, 334]. Moreover, Notch signaling can lead to induction of epithelial-mesenchymal transition, which is associated with loss of tumor cell adhesion as well as migration and as consequence, with the development of metastasis [334]. Thus, pharmacological inhibitors of the Notch signaling pathway were suggested as promising drugs for the treatment of human CRC [333, 336].

Besides the general investigation of altered genes and processes associated with inflammation-triggered colorectal cancer, it is important to consider gene expression levels in each tumor stage separately to truly understand the molecular background of inflammation-triggered colorectal cancer. This is in particular crucial for pathways that change the regulatory direction during the course of the disease. As an example, p53-signaling, which is associated with several cancer types including colorectal cancer [7, 59, 323], was upregulated during malignant transformation and characterized by a decline in the expression level during tumor progression in the investigated murine tumor samples. This indicates that each tumor stage is influenced by p53 signaling in a different manner and highlights the importance to investigate molecular differences between tumor stages. Another example points towards genes involved in the cell cycle. Although this process was clearly upregulated

in all investigated tumor samples, the expression level was lower in mice treated with a high DSS dose referring to an advanced tumor stage. Downregulation of the cell cycle might enable invasion and metastasis of cancer cells [337]. However, the expression decrease was not strong enough or additional factors are required for tumor invasion and metastasis, because both features were missing in the applied mouse model. This indicates that different activation points in time of specific processes might contribute to phenotypic characteristics and could also explain differences between human CAC and the corresponding mouse model. As an example, in human CAC, overexpression of MAPK signaling is involved in the inflammatory and dysplasia stage, but does not contribute to the cancerous phases [65]. In contrast, this process mainly influences late stage tumor growth in murine AOM/DSS-induced cancer in my studies. However, further studies will be necessary to compare molecular characteristics, which are associated with the tumor stage, between murine and human inflammation-triggered colorectal cancer.

In addition to the unequal cancer aggressiveness, further phenotypic differences between the AOM/DSS mouse model and human CAC exist. An occasional complication of AOM/DSS-treated mice with a high rectal tumor growth is an intestinal prolapse [214, 338]. In contrast, an intestinal prolapse is rare in human IBD patients and may arise only during the active phase [339]. The different susceptibility is based on a much shorter rectum length in mice than in humans [340]. However, the molecular consequences of an intestinal prolapse are mostly unknown. Two of the major factors contributing to the development of an intestinal prolapse are inflammation and diarrhea [340]. Diarrhea might have also an impact on deficiency of nutrients, modification of ionic transports, and diffusion of water in the intestinal lumen [341]. This in turn can result into an altered metabolism as well as a disturbed water and cellular ion homeostasis. Indeed, all of these processes were differentially expressed between tumors from mice with and those without an intestinal prolapse in the current study. The intestinal prolapse itself can result in mucosal hyperplasia and increased entry of environmental bacteria into the host [340, 341]. This in turn can trigger the activation of the immune system as well as changes in the microbial composition. In particular, differences in bacterial abundances caused by increased contact with the cage environment and the change of nutrient availability due to diarrhea may result in an altered microbial production or consumption of metabolites, which can influence the metabolism of the host by a distinct availability of specific nutrients [342, 343]. All of these known consequences of an intestinal prolapse were also reflected in the transcriptome data of the murine tumor samples investigated in my studies: Genes differentially expressed between tumors from mice with and those without intestinal prolapse were e.g. enriched for metabolic pathways, processes involved in the immune system, and genes responsible for thickened skin. Taken together, the consequences of an intestinal prolapse are clearly visible in the transcriptional signature of the

tumors and might influence the transferability of molecular alterations observed in the mouse model to human CAC.

Taken together, phenotypic differences between murine AOM/DSS-triggered colorectal cancer and human CAC, such as missing tumor aggressiveness in the mouse model and seldom occurrence of an intestinal prolapse in human CAC, are reflected on molecular level. Besides different alteration spectrums between the human disease and the mouse model, this includes the involvement of molecular features, which are characteristic for specific tumor stages. Thus, the AOM/DSS mouse model does not completely reflect human CAC and uniquely existing features should be considered while using this mouse model in future.

4.2.6 Specific splicing patterns in AOM/DSS-induced cancer

Alternative splicing is a major factor to enable protein diversity based on the genome [208]. Also cancer cells are characterized by aberrant splicing, which is involved in the dysregulation of cancer-associated processes and the progression of the disease [208]. This phenomenon existed also in the investigated tumor samples of the current study, in which alternative splicing was investigated in AOM/DSS-induced tumors for the first time. The observed main splicing type in the tumor samples of the current study was the usage of an alternative acceptor or donor splice site followed by intron retention. This is in line with previous studies reporting a shift from exon skipping, which is the most abundant splicing type in the majority of eukaryotic healthy tissues, to alternative splice site selection and intron retention in cancer cells [344]. IBD is, like normal cells, characterized by alternative exon usage [88], which was not increased in the investigated tumor samples. This indicates that the splicing effect of tumor cells exceeded the inflammatory influence or that inflammation had no impact on the splicing pattern in AOM/DSS-induced tumors.

Alternative splicing seems to occur in all cancer-associated processes, including metabolism, cell cycle control, cell survival, invasion, and metastasis [345]. Alternative splicing in cancer-relevant genes and processes was also observed in the tumor samples of my studies. However, not all splicing types existed in each cancer-associated process in the investigated AOM/DSS-induced tumor samples. This facilitates the hypothesis that the usage of different splicing forms is not a side effect of cancer-associated processes but another important level of molecular regulation affecting a broad spectrum of cancer-relevant pathways [344].

Processes affected by only one splicing type might indicate a specific regulation instead of a side effect. Interestingly, in the current study, specific splicing was observed for processes involved in malignant transformation and tumor progression, but was rarely detected in pathways with a strong impact mainly on metastasis and invasion. As example, the Wnt and MAPK signaling pathways were only affected by different exon usage in the tumor samples investigated in my studies. As described above, both processes are involved in several cancer

stages and do not have the main focus on invasion and metastasis development [65, 261, 262, 266]. Moreover, the NOD-like receptor (NLR) pathway harbored only more genes with alternative splice acceptor site positions than expected by chance. Altered NLR molecules might activate inflammatory processes and, due to the impact of NLRs on cell death as well as autophagy processes, the pathway is involved in tumor growth and progression [346].

Processes enriched for several kinds of splicing might indicate a more unspecific regulation. Interestingly, in the current study, these processes were mostly involved in metastases, which rarely exist in AOM/DSS-treated mice [214]. As an example, in the tumor samples of the current study, altered metabolism, which is, as discussed in section 4.2.3, a hallmark of cancer [2] was affected by nearly all splicing types. In particular, reconstruction of the metabolism is essential for invasion and metastasis [347]. Moreover, in my studies, nearly all splicing types differed between tumor and control samples for at least one process associated with cell adhesion. This included the apical junctional complex (AJC), consisting of tight junctions (TJs) and adherens junctions (AJs), and the actin cytoskeleton. The AJC is linked to the actin cytoskeleton, which acts as driver of CAC development [268, 348]. Disorganization of the actin cytoskeleton or AJC leads to disruption of cell morphology, cell adhesion, and epithelial functions. Alterations are also critical for malignant features, such as cell migration, invasion, and metastasis. Thus, disordered epithelial organization is considered as hallmark of cancer [348, 349]. Also genes associated with ECM receptor interaction and the FA complex showed a tendency for a more unspecific regulation in the current study. The ECM interacts with epithelial cells and focal adhesion complexes [350]. Changed activities of ECM and FA allow tumor cells to adapt the microenvironment and thus to support tumor cell survival and malignant progression [351]. In particular, reconstruction of the ECM is essential for the development of metastasis [351].

Taken together the splicing pattern observed in the tumor samples of my studies were clearly influenced by cancer-specific factors rather than inflammatory components. In AOM/DSS-triggered colorectal cancer, processes involved in malignant transformation and mostly non-aggressive tumor progression were regulated by specific splicing, while pathways involved in metastasis were characterized by unspecific splicing. Overall, the current study demonstrates that alternative splicing can be an additional important regulatory level of cancer-relevant genes and processes. Thus, the effect of alternative splicing should not be underestimated in cancer studies.

4.2.7 Conclusion of the murine AOM/DSS-induced colorectal cancer study

A comprehensive molecular characterization of the murine AOM/DSS-induced colorectal cancer model was performed using an integrative analysis based on WES and transcriptome data. The reported results clearly show that the AOM/DSS mouse model is an

inflammation-triggered colonic tumorigenesis model, in which a large part of the molecular background is shared with human CAC. Common alterations included an upregulation of several cytokines and transcription factors (e.g. *Tnf*, *CDKN1C*, and *NF-κB*), hyperactivation of the Wnt signaling pathway, an altered metabolism as well as modified processes involved in cell structure. Moreover, the analysis revealed molecular features of the model, which were not investigated in human CAC so far. This includes (i) a random distribution of SNV positions, (ii) association of gene expression patterns with tumor localization, DSS dose, and clinical signs such as intestinal prolapse, and (iii) the identification of further genes and processes potentially associated with inflammation-associated colorectal cancer, such as *Nfκbiζ*, *Foxo3*, and *Ppara*. The integrated compendium of the genetic and transcriptomic alterations in AOM/DSS-induced murine colorectal cancer will thus serve as a comprehensive resource for the interpretation of novel findings in this widely used model, which will ultimately also lead to a deeper understanding of the biology of human CAC. However, despite the huge amount of commonalities between murine AOM/DSS-induced colorectal cancer and the human CAC, also striking molecular differences should be noted. As an example, several genes were not mutated in murine inflammation-triggered colorectal cancer, although variants were reported for human CAC (e.g. *Trp53*). Vice versa, mutations in *CTNNB1* are rare in human CAC, while mutated *Ctnnb1* seems to harbor a central role in the mouse model. Moreover, the expression level of several genes were contrary between human CAC and AOM/DSS-triggered colorectal cancer (e.g. *Lrp6*). Differences in the molecular signatures might contribute to distinct cancer phenotypes, such as lack of metastasis in the mouse model, and should be taken into consideration when using AOM/DSS-induced cancer as a model for human CAC. Despite these limitations, it was demonstrated that the AOM/DSS mouse model can be applied to decipher genes and processes potentially involved in the development of human CAC. Thus, the AOM/DSS mouse model can be very useful to understand the molecular background of CAC and to develop novel therapies for the treatment of human CAC.

4.3 Conclusion

Cancer is among the diseases with the highest number of deaths in economically developed countries. Thereby, molecular factors have a high impact on etiopathology and therapy response. To decipher genomic mechanisms underlying tumor development, two gastrointestinal cancer types, namely human sporadic gastric cancer and murine inflammation-triggered colorectal cancer, were investigated.

In the GC study, one microsatellite stable as well as one microsatellite unstable gastric carcinoma were investigated using WES and WGS. Despite the low sample number, many alterations reported by previous studies could be confirmed, while novel features were additionally revealed. The two investigated tumor samples harbored clearly distinct mutational

signatures. Thus, molecular patterns underlying GC might be potential biomarkers in future. In particular, the observed high heterogeneity should be considered in treatment and diagnostic approaches for GC.

Besides the investigation of genomic mechanisms underlying GC, a comprehensive molecular characterization of the AOM/DSS mouse model was performed using an integrative analysis approach based on WES and whole transcriptome sequencing. This revealed commonalities between the mouse model and human CAC as well as novel insights into inflammation-triggered colorectal cancer, such as a specific variant signature and expression patterns associated with tumor localization. However, also discrepancies between the human disease and murine AOM/DSS-triggered colorectal cancer were observed, which might explain phenotypic differences, such as missing metastasis in AOM/DSS-treated mice. Thus, the current study demonstrate advantages and limitations of the AOM/DSS mouse model and is an important resource for the interpretation of results observed in murine AOM/DSS experiments in future. This is crucial to draw conclusions for human CAC and is the first step to enable the cure of CAC.

Taken together, both studies revealed novel insights into the molecular background of two gastrointestinal cancer types. The improved understanding of genomic mechanisms contributing to these cancer types is an important step towards the development of novel treatment and diagnostic approaches. Moreover, both studies demonstrate the importance of NGS for clinical applications.

5 References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61:69–90.

2. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.

3. ICGC TJH, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature. 2010;464:993–8.

4. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458:719–24.

5. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.

6. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010;463:191–6.

7. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513:202–9.

8. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100:57–70.

9. Pfeifer GP. Environmental exposures and mutational patterns of cancer genomes. Genome Med. 2010;2:54.

10. Vogelstein B, Kinzler KW. The multistep nature of cancer. Trends Genet TIG. 1993;9:138–41.

11. Cooper GM. The Development and Causes of Cancer. 2. ed. Washington, DC: ASM Press; 2000.

12. Rakoff-Nahoum S. Why Cancer and Inflammation? Yale J Biol Med. 2006;79:123–30.

13. Sobin LH, Gospodarowicz MK, Wittekind C. TNM Classification of Malignant Tumours. 7 edition. Chichester, West Sussex, UK ; Hoboken, NJ: Wiley-Blackwell; 2009.

14. Lee YY, Derakhshan MH. Environmental and lifestyle risk factors of gastric cancer. Arch Iran Med. 2013;16:358–65.

15. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000;343:78–85.

16. Liang H, Kim YH. Identifying Molecular Drivers of Gastric Cancer through Next-generation Sequencing. Cancer Lett. 2013;340:241–6.

17. Dicken BJ, Bigam DL, Cass C, Mackey JR, Joy AA, Hamilton SM. Gastric Adenocarcinoma. Ann Surg. 2005;241:27–39.

18. Bertuccio P, Chatenoud L, Levi F, Praud D, Ferlay J, Negri E, et al. Recent patterns in gastric cancer: A global overview. Int J Cancer. 2009;125:666–73.

19. Song Z, Wu Y, Yang J, Yang D, Fang X. Progress in the treatment of advanced gastric cancer. Tumor Biol. 2017;39:101042831771462.

20. DeVita VT Jr, Lawrence TS. DeVita, Hellman, and Rosenberg's Cancer: Principles and Practice of Oncology. Revised. Philadelphia: Lippincott Raven; 2011.

21. Hu B, El Hajj N, Sittler S, Lammert N, Barnes R, Meloni-Ehrig A. Gastric cancer: Classification, histology and application of molecular pathology. J Gastrointest Oncol. 2012;3:251–61.

22. Werner M, Becker KF, Keller G, Höfler H. Gastric adenocarcinoma: pathomorphology and molecular pathology. J Cancer Res Clin Oncol. 2001;127:207–16.

23. Bevan S, Houlston RS. Genetic predisposition to gastric cancer. QJM. 1999;92:5–10.

24. Piazuelo MB, Correa P. Gastric cáncer: Overview. Colomb Médica CM. 44:192–201.

25. Shokal U, Sharma PC. Implication of microsatellite instability in human gastric cancers. Indian J Med Res. 2012;135:599–613.

26. Beghelli S, de Manzoni G, Barbi S, Tomezzoli A, Roviello F, Di Gregorio C, et al. Microsatellite instability in gastric cancer is associated with better prognosis in only stage II cancers. Surgery. 2006;139:347–56.

27. Elinav E, Nowarski R, Thaiss CA, Hu B, Jin C, Flavell RA. Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. Nat Rev Cancer. 2013;13:759–71.

28. Reuter S, Gupta SC, Chaturvedi MM, Aggarwal BB. Oxidative stress, inflammation, and cancer: How are they linked? Free Radic Biol Med. 2010;49:1603–16.

29. Triantafillidis JK, Nasioulas G, Kosmidis PA. Colorectal Cancer and Inflammatory Bowel Disease: Epidemiology, Risk Factors, Mechanisms of Carcinogenesis and Prevention Strategies. Anticancer Res. 2009;29:2727–37.

30. Colotta F, Allavena P, Sica A, Garlanda C, Mantovani A. Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. Carcinogenesis. 2009;30:1073–81.

31. Coussens LM, Werb Z. Inflammation and cancer. Nature. 2002;420:860–7.

32. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. Nature. 2011;474:307–17.

33. Chaudhari S, Desai JS, Adam A, Mishra P. Inflammatory Bowel Disease: An Idiopathic Disease and its Treatment. Int J Pharma Res Rev. 2014;3:106–14.

34. Panaccione R. Mechanisms of Inflammatory Bowel Disease. Gastroenterol Hepatol. 2013;9:529–32.

35. Eaden JA, Abrams KR, Mayberry JF. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut. 2001;48:526–35.

36. Canavan C, Abrams KR, Mayberry J. Meta-analysis: colorectal and small bowel cancer risk in patients with Crohn's disease. Aliment Pharmacol Ther. 2006;23:1097–104.

37. Frolkis A, Dieleman LA, Barkema HW, Panaccione R, Ghosh S, Fedorak RN, et al. Environment and the inflammatory bowel diseases. Can J Gastroenterol. 2013;27:e18–24.

38. Halme L, Paavola-Sakki P, Turunen U, Lappalainen M, Färkkilä M, Kontula K. Family and twin studies in inflammatory bowel disease. World J Gastroenterol WJG. 2006;12:3668–72.

39. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119–24.

40. Munkholm P. Review article: the incidence and prevalence of colorectal cancer in inflammatory bowel disease. Aliment Pharmacol Ther. 2003;18 Suppl 2:1–5.

41. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. Cell. 2010;140:883–99.

42. Sunkara S, Swanson G, Forsyth CB, Keshavarzian A, Sunkara S, Swanson G, et al. Chronic Inflammation and Malignancy in Ulcerative Colitis, Chronic Inflammation and Malignancy in Ulcerative Colitis. Ulcers Ulcers. 2011;2011, 2011:e714046.

43. Beaugerie L, Itzkowitz SH. Cancers Complicating Inflammatory Bowel Disease. N Engl J Med. 2015;372:1441–52.

44. Choi PM, Zelig MP. Similarity of colorectal cancer in Crohn's disease and ulcerative colitis: implications for carcinogenesis and prevention. Gut. 1994;35:950–4.

45. Watanabe T, Konishi T, Kishimoto J, Kotake K, Muto T, Sugihara K, et al. Ulcerative colitis-associated colorectal cancer shows a poorer survival than sporadic colorectal cancer: a nationwide Japanese study. Inflamm Bowel Dis. 2011;17:802–8.

46. Ren L-L, Fang J-Y. Should we sound the alarm? Dysplasia and colitis-associated colorectal cancer. Asian Pac J Cancer Prev APJCP. 2011;12:1881–6.

47. Rhodes JM, Campbell BJ. Inflammation and colorectal cancer: IBD-associated and sporadic cancer compared. Trends Mol Med. 2002;8:10–6.

48. De Robertis M, Massi E, Poeta M, Carotti S, Morini S, Cecchetelli L, et al. The AOM/DSS murine model for the study of colon carcinogenesis: From pathways to diagnosis and therapy studies. J Carcinog. 2011;10:9.

49. Tanaka T. Development of an inflammation-associated colorectal cancer model and its application for research on carcinogenesis and chemoprevention. Int J Inflamm. 2012;2012:658786.

50. Neufert C, Becker C, Neurath MF. An inducible mouse model of colon carcinogenesis for the analysis of sporadic and inflammation-driven tumor progression. Nat Protoc. 2007;2:1998–2004.

51. Rosenberg DW, Giardina C, Tanaka T. Mouse models for the study of colon carcinogenesis. Carcinogenesis. 2009;30:183–96.

52. Megaraj V, Ding X, Fang C, Kovalchuk N, Zhu Y, Zhang Q-Y. Role of Hepatic and Intestinal P450 Enzymes in the Metabolic Activation of the Colon Carcinogen Azoxymethane in Mice. Chem Res Toxicol. 2014;27:656–62.

53. Tanaka T, Kohno H, Suzuki R, Yamada Y, Sugie S, Mori H. A novel inflammation-related mouse colon carcinogenesis model induced by azoxymethane and dextran sodium sulfate. Cancer Sci. 2003;94:965–73.

54. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446:153–8.

55. TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–8.

56. Wong SS, Kim K-M, Ting JC, Yu K, Fu J, Liu S, et al. Genomic landscape and genetic heterogeneity in gastric adenocarcinoma revealed by whole-genome sequencing. Nat Commun. 2014;5:5477.

57. Wang K, Yuen ST, Xu J, Lee SP, Yan HHN, Shi ST, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat Genet. 2014;46:573–82.

58. Liu J, McCleland M, Stawiski EW, Gnad F, Mayba O, Haverty PM, et al. Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. Nat Commun. 2014;5:3830.

59. Robles AI, Traverso G, Zhang M, Roberts NJ, Khan MA, Joseph C, et al. Whole-Exome Sequencing Analyses of Inflammatory Bowel Disease-Associated Colorectal Cancers. Gastroenterology. 2016;150:931–43.

60. Watanabe T, Kobunai T, Toda E, Kanazawa T, Kazama Y, Tanaka J, et al. Gene expression signature and the prediction of ulcerative colitis-associated colorectal cancer by DNA microarray. Clin Cancer Res Off J Am Assoc Cancer Res. 2007;13 2 Pt 1:415–20.

61. Pan Q, Lou X, Zhang J, Zhu Y, Li F, Shan Q, et al. Genomic variants in mouse model induced by azoxymethane and dextran sodium sulfate improperly mimic human colorectal cancer. Sci Rep. 2017;7.

62. Suzuki R, Miyamoto S, Yasui Y, Sugie S, Tanaka T. Global gene expression analysis of the mouse colonic mucosa treated with azoxymethane and dextran sodium sulfate. BMC Cancer. 2007;7:84.

63. Gao Y, Li X, Yang M, Zhao Q, Liu X, Wang G, et al. Colitis-accelerated colorectal cancer and metabolic dysregulation in a mouse model. Carcinogenesis. 2013;:bgt135.

64. Li X, Gao Y, Yang M, Zhao Q, Wang G, Yang Y mei, et al. Identification of Gene Expression Changes from Colitis to CRC in the Mouse CAC Model. PLoS ONE. 2014;9:e95347.

65. Tang A, Li N, Li X, Yang H, Wang W, Zhang L, et al. Dynamic activation of the key pathways: linking colitis to colorectal cancer in a mouse model. Carcinogenesis. 2012;33:1375–83.

66. Subramaniam R, Mizoguchi A, Mizoguchi E. Mechanistic roles of epithelial and immune cell signaling during the development of colitis-associated cancer. Cancer Res Front. 2016;2:1–21.

67. Hussain SP, Amstad P, Raja K, Ambs S, Nagashima M, Bennett WP, et al. Increased p53 mutation load in noncancerous colon tissue from ulcerative colitis: a cancer-prone chronic inflammatory disease. Cancer Res. 2000;60:3333–7.

68. Cooks T, Pateras IS, Tarcic O, Solomon H, Schetter AJ, Wilder S, et al. Mutant p53 Prolongs NF-κB Activation and Promotes Chronic Inflammation and Inflammation-Associated Colorectal Cancer. Cancer Cell. 2013;23:634–46.

69. Waldner MJ, Neurath MF. Mechanisms of Immune Signaling in Colitis-Associated Cancer. CMGH Cell Mol Gastroenterol Hepatol. 2015;1:6–16.

70. Al Obeed OA, Alkhayal KA, Al Sheikh A, Zubaidi AM, Vaali-Mohammed M-A, Boushey R, et al. Increased expression of tumor necrosis factor-α is associated with advanced colorectal cancer stages. World J Gastroenterol. 2014;20:18390–6.

71. Van Der Kraak L, Gros P, Beauchemin N. Colitis-associated colon cancer: Is it in your genes? World J Gastroenterol. 2015;21:11688–99.

72. Popivanova BK, Kitamura K, Wu Y, Kondo T, Kagaya T, Kaneko S, et al. Blocking TNF-α in mice reduces colorectal carcinogenesis associated with chronic colitis. J Clin Invest. 2008;118:560–70.

73. Lasry A, Zinger A, Ben-Neriah Y. Inflammatory networks underlying colorectal cancer. Nat Immunol. 2016;17:230–40.

74. Itzkowitz SH, Yio X. Inflammation and cancer IV. Colorectal cancer in inflammatory bowel disease: the role of inflammation. Am J Physiol Gastrointest Liver Physiol. 2004;287:G7-17.

75. Kim ER, Chang DK. Colorectal cancer in inflammatory bowel disease: The risk, pathogenesis, prevention and diagnosis. World J Gastroenterol WJG. 2014;20:9872–81.

76. Tomasello G, Sciumè C, Rappa F, Rodolico V, Zerilli M, Martorana A, et al. Hsp10, Hsp70, and Hsp90 immunohistochemical levels change in ulcerative colitis after therapy. Eur J Histochem EJH. 2011;55.

77. Corvinus FM, Orth C, Moriggl R, Tsareva SA, Wagner S, Pfitzner EB, et al. Persistent STAT3 Activation in Colon Cancer Is Associated with Enhanced Cell Proliferation and Tumor Growth. Neoplasia N Y N. 2005;7:545–55.

78. Plewka D, Plewka A, Miskiewicz A, Morek M, Bogunia E. Nuclear factor-kappa B as potential therapeutic target in human colon cancer. J Cancer Res Ther. 2018;14:516.

79. Schatoff EM, Leach BI, Dow LE. Wnt Signaling and Colorectal Cancer. Curr Colorectal Cancer Rep. 2017;13:101–10.

80. Drecoll E, Nitsche U, Bauer K, Berezowska S, Slotta-Huspenina J, Rosenberg R, et al. Expression analysis of heat shock protein 90 (HSP90) and Her2 in colon carcinoma. Int J Colorectal Dis. 2014;29:663–71.

81. Schmidt EM, Lamprecht S, Blaj C, Schaaf C, Krebs S, Blum H, et al. Targeting tumor cell plasticity by combined inhibition of NOTCH and MAPK signaling in colon cancer. J Exp Med. 2018;215:1693–708.

82. International Human Genome Sequencing Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004;431:931–45.

83. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet TIG. 2014;30:418–26.

84. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.

85. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta. 2014;1842:1932–41.

86. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010;11:31–46.

87. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.

88. Häsler R, Sheibani-Tezerji R, Sinha A, Barann M, Rehman A, Esser D, et al. Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. Gut. 2016.

89. Pertea M. The human transcriptome: an unfinished story. Genes. 2012;3:344–60.

90. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32.

91. Piskol R, Ramaswami G, Li JB. Reliable Identification of Genomic Variants from RNA-Seq Data. Am J Hum Genet. 2013;93:641–51.

92. Liu M, Guan X-Y. Allele imbalance in the transcriptome of human hepatocellular carcinoma: stress-induced gene plays a role. Sci Proc. 2014;1:382.

93. Pinter SF, Colognori D, Beliveau BJ, Sadreyev RI, Payer B, Yildirim E, et al. Allelic Imbalance Is a Prevalent and Tissue-Specific Feature of the Mouse Transcriptome. Genetics. 2015;200:537–49.

94. Park E, Williams B, Wold BJ, Mortazavi A. RNA editing in the human ENCODE RNA-seq data. Genome Res. 2012;22:1626–33.

95. Samuels DC, Han L, Li J, Quanghu S, Clark TA, Shyr Y, et al. Finding the lost treasures in exome sequencing data. Trends Genet TIG. 2013;29:593–9.

96. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J Hum Genet. 2014;59:5–15.

97. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011;48:580–9.

98. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7:111–8.

99. Bodi K, Perera AG, Adams PS, Bintzler D, Dewar K, Grove DS, et al. Comparison of Commercially Available Target Enrichment Methods for Next-Generation Sequencing. J Biomol Tech JBT. 2013;24:73–86.

100. Arnheim N, Calabrese P. Understanding what determines the frequency and pattern of human germline mutations. Nat Rev Genet. 2009;10:478–88.

101. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.

102. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 2014;42:D764–70.

103. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Molecular Cell Biology. 4. ed. New York, NY: Freeman; 2000.

104. Kondrashov A. Genetics: The rate of human mutation. Nature. 2012;488:467–8.

105. Uchimura A, Higuchi M, Minakuchi Y, Ohno M, Toyoda A, Fujiyama A, et al. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res. 2015;25:1125–34.

106. Watson IR, Takahashi K, Futreal PA, Chin L. Emerging patterns of somatic mutations in cancer. Nat Rev Genet. 2013;14:703–18.

107. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000;156:297–304.

108. Guo C, McDowell IC, Nodzenski M, Scholtens DM, Allen AS, Lowe WL, et al. Transversions have larger regulatory effects than transitions. BMC Genomics. 2017;18:394.

109. Choudhuri S. Chapter 2 - Fundamentals of Molecular Evolution*. In: Bioinformatics for Beginners. Oxford: Academic Press; 2014. p. 27–53.

110. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014;24:52–60.

111. Natarajan AT, Darroudi F, Jha AN, Meijers M, Zdzienicka MZ. Ionizing radiation induced DNA lesions which lead to chromosomal aberrations. Mutat Res. 1993;299:297–303.

112. Westcott PMK, Halliwill KD, To MD, Rashid M, Rust AG, Keane TM, et al. The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. Nature. 2015;517:489–92.

113. Jia P, Pao W, Zhao Z. Patterns and processes of somatic mutations in nine major cancers. BMC Med Genomics. 2014;7:11.

114. Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. J Biol Chem. 1992;267:166–72.

115. Pfeifer GP, Besaratinia A. Mutational spectra of human cancer. Hum Genet. 2009;125:493–506.

116. Moolenbeek C, Ruitenberg EJ. The “Swiss roll”: a simple technique for histological studies of the rodent intestine. Lab Anim. 1981;15:57–9.

117. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014;42 Database issue:D756-763.

118. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009;37 Database issue:D412-416.

119. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.

120. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.

121. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.

122. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–83.

123. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34:57–65.

124. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009;1:13.

125. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42 Database issue:D980-985.

126. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42 Database issue:D1001-1006.

127. Waldman YY, Geiger T, Ruppin E. A Genome-Wide Systematic Analysis Reveals Different and Predictive Proliferation Expression Signatures of Cancerous vs. Non-Cancerous Cells. PLOS Genet. 2013;9:e1003806.

128. Wilcoxon F. Individual Comparisons by Ranking Methods. Biom Bull. 1945;1:80–3.

129. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. J Am Stat Assoc. 1952;47:583–621.

130. Fisher RA. On the Interpretation of χ2 from Contingency Tables, and the Calculation of P. J R Stat Soc. 1922;85:87–94.

131. R Core Team. R: A language and environment for statistical computing. 2015. http://www.R-project.org/.

132. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.

133. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.

134. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2010;26:589–95.

135. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinforma Oxf Engl. 2011;27:863–4.

136. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.

137. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.

138. 't Hoen PAC, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol. 2013;31:1015–22.

139. Fisher SRA. Statistical methods for research workers. Fourteenth Edition Revised. Enlarged 14th edition. Edinburgh: Oliver & Boyd; 1970.

140. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

141. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.

142. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

143. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009;25:2078–9.

144. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.

145. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11:863–74.

146. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.

147. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32:894–9.

148. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.

149. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.

150. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677–81.

151. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinforma Oxf Engl. 2009;25:2865–71.

152. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2012;:bbs038.

153. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149:979–93.

154. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data. 2015.

155. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinforma Oxf Engl. 2018.

156. Lada AG, Dhar A, Boissy RJ, Hirano M, Rubel AA, Rogozin IB, et al. AID/APOBEC cytosine deaminase induces genome-wide kataegis. Biol Direct. 2012;7:47; discussion 47.

157. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinforma Oxf Engl. 2015;31:3673–5.

158. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016;17.

159. Blokzijl F, Janssen R, Boxtel RV, Cuppen E. MutationalPatterns: an integrative R package for studying patterns in base substitution catalogues. bioRxiv. 2016;:071761.

160. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12:656–64.

161. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009.

162. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinforma Oxf Engl. 2015;31:166–9.

163. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

164. Wickham H. ggplot2. New York, NY: Springer New York; 2009.

165. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562–78.

166. Klostermeier UC, Barann M, Wittig M, Häsler R, Franke A, Gavrilova O, et al. A tissue-specific landscape of sense/antisense transcription in the mouse intestine. BMC Genomics. 2011;12:305.

167. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456:470–6.

168. Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, et al. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Mol Syst Biol. 2008;4.

169. Reimand J, Arak T, Vilo J. g:Profiler--a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. 2011;39 Web Server issue:W307-315.

170. Zambelli F, Pesole G, Pavesi G. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. Nucleic Acids Res. 2009;37 Web Server issue:W247-252.

171. Esser D, Holze N, Haag J, Schreiber S, Krüger S, Warneke V, et al. Interpreting whole genome and exome sequencing data of individual gastric cancer samples. BMC Genomics. 2017;18:517.

172. Esser D, Falk-Paulsen M, Ito G, Kuiper J, Aden K, Billmann-Born S, et al. Mutational and transcriptional landscapes of murine inflammation-induced colorectal cancer. submitted.

173. Karakas B, Bachman KE, Park BH. Mutation of the PIK3CA oncogene in human cancers. Br J Cancer. 2006;94:455–9.

174. Zang ZJ, Cutcutache I, Poon SL, Zhang SL, McPherson JR, Tao J, et al. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet. 2012;44:570–4.

175. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol. 2016;34:155–63.

176. Toguchida J, Yamaguchi T, Dayton SH, Beauchamp RL, Herrera GE, Ishizaki K, et al. Prevalence and spectrum of germline mutations of the p53 gene among patients with sarcoma. N Engl J Med. 1992;326:1301–8.

177. Zheng R, Blobel GA. GATA Transcription Factors and Cancer. Genes Cancer. 2010;1:1178–88.

178. Dulak AM, Schumacher S, van Lieshout J, Imamura Y, Fox C, Shim B, et al. Gastrointestinal adenocarcinomas of the esophagus, stomach and colon exhibit distinct patterns of genome instability and oncogenesis. Cancer Res. 2012;72:4383–93.

179. Zisman-Rozen S, Fink D, Ben-Izhak O, Fuchs Y, Brodski A, Kraus MH, et al. Downregulation of Sef, an inhibitor of receptor tyrosine kinase signaling, is common to a variety of human carcinomas. Oncogene. 2007;26:6093–8.

180. Her C, Zhao N, Wu X, Tompkins JD. MutS homologues hMSH4 and hMSH5: diverse functional implications in humans. Front Biosci J Virtual Libr. 2007;12:905–11.

181. Clark N, Wu X, Her C. MutS Homologues hMSH4 and hMSH5: Genetic Variations, Functions, and Implications in Human Diseases. Curr Genomics. 2013;14:81–90.

182. Sun, Wanpeng Y Jian. Tumor Suppressor RIZ1 in Carcinogenesis. J Carcinog Mutagen. 2014;05.

183. Xie W, Li X, Chen X, Huang S, Huang S. Decreased expression of PRDM2 (RIZ1) and its correlation with risk stratification in patients with myelodysplastic syndrome. Br J Haematol. 2010;150:242–4.

184. Pan K-F, Lu Y-Y, Liu W-G, Zhang L, You W-C. Detection of frameshift mutations of RIZ in gastric cancers with microsatellite instability. World J Gastroenterol WJG. 2004;10:2719–22.

185. Polyak K, Xia Y, Zweier JL, Kinzler KW, Vogelstein B. A model for p53-induced apoptosis. Nature. 1997;389:300–5.

186. Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat Genet. 2002;30:315–20.

187. Wang LJ, Jin HC, Wang X, Lam EKY, Zhang JB, Liu X, et al. ZIC1 is downregulated through promoter hypermethylation in gastric cancer. Biochem Biophys Res Commun. 2009;379:959–63.

188. Zhong J, Chen S, Xue M, Du Q, Cai J, Jin H, et al. ZIC1 modulates cell-cycle distributions and cell migration through regulation of sonic hedgehog, PI3K and MAPK signaling pathways in gastric cancer. BMC Cancer. 2012;12:290.

189. Saiki Y, Yoshino Y, Fujimura H, Manabe T, Kudo Y, Shimada M, et al. DCK is frequently inactivated in acquired gemcitabine-resistant human cancer cells. Biochem Biophys Res Commun. 2012;421:98–104.

190. Kinzler KW, Nilbert MC, Vogelstein B, Bryan TM, Levy DB, Smith KJ, et al. Identification of a gene located at chromosome 5q21 that is mutated in colorectal cancers. Science. 1991;251:1366–70.

191. Li W-D, Li Q-R, Xu S-N, Wei F-J, Ye Z-J, Cheng J-K, et al. Exome sequencing identifies an MLL3 gene germ line mutation in a pedigree of colorectal cancer and acute myeloid leukemia. Blood. 2013;121:1478–9.

192. Gong Y, Zack TI, Morris LGT, Lin K, Hukkelhoven E, Raheja R, et al. Pan-cancer genetic analysis identifies PARK2 as a master regulator of G1/S cyclins. Nat Genet. 2014;46:588–94.

193. Lin D-C, Xu L, Ding L-W, Sharma A, Liu L-Z, Yang H, et al. Genomic and functional characterizations of phosphodiesterase subtype 4D in human cancers. Proc Natl Acad Sci. 2013;110:6109–14.

194. Jaccard P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. New Phytol. 1912;11:37–50.

195. Hershberg R. Mutation—The Engine of Evolution: Studying Mutation and Its Role in the Evolution of Bacteria. Cold Spring Harb Perspect Biol. 2015;7.

196. Domingues V. Evolution: Replication-transcription conflict promotes gene evolution. Nat Rev Genet. 2013;14:302–3.

197. Paul S, Million-Weaver S, Chattopadhyay S, Sokurenko E, Merrikh H. Accelerated gene evolution through replication-transcription conflicts. Nature. 2013;495:512–5.

198. Price MN, Alm EJ, Arkin AP. Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. Nucleic Acids Res. 2005;33:3224–34.

199. Goodman WA, Omenetti S, Date D, Di Martino L, De Salvo C, Kim G-D, et al. KLF6 contributes to myeloid cell plasticity in the pathogenesis of intestinal inflammation. Mucosal Immunol. 2016;9:1250–62.

200. Reeves HL, Narla G, Ogunbiyi O, Haq AI, Katz A, Benzeno S, et al. Kruppel-like factor 6 (KLF6) is a tumor-suppressor gene frequently inactivated in colorectal cancer. Gastroenterology. 2004;126:1090–103.

201. Kostaras E, Sflomos G, Pedersen NM, Stenmark H, Fotsis T, Murphy C. SARA and RNF11 interact with each other and ESCRT-0 core proteins and regulate degradative EGFR trafficking. Oncogene. 2013;32:5220–32.

202. Liu S, Tang H, Zhu J, Ding H, Zeng Y, Du W, et al. High expression of Copine 1 promotes cell growth and metastasis in human lung adenocarcinoma. Int J Oncol. 2018.

203. Jamsai D, Watkins DN, O'Connor AE, Merriner DJ, Gursoy S, Bird AD, et al. In vivo evidence that RBM5 is a tumour suppressor in the lung. Sci Rep. 2017;7:16323.

204. Kent OA, Sandí M-J, Burston HE, Brown KR, Rottapel R. An oncogenic KRAS transcription program activates the RHOGEF ARHGEF2 to mediate transformed phenotypes in pancreatic cancer. Oncotarget. 2017;8:4484–500.

205. Chiesa M, Guillamot M, Bueno MJ, Malumbres M. The Cdc14B phosphatase displays oncogenic activity mediated by the Ras-Mek signaling pathway. Cell Cycle Georget Tex. 2011;10:1607–17.

206. Hong S-B, Oh H, Valera VA, Stull J, Ngo D-T, Baba M, et al. Tumor suppressor FLCN inhibits tumorigenesis of a FLCN-null renal cancer cell line and regulates expression of key molecules in TGF-beta signaling. Mol Cancer. 2010;9:160.

207. Liu P, Kao TP, Huang H. CDK1 promotes cell proliferation and survival via phosphorylation and inhibition of FOXO1 transcription factor. Oncogene. 2008;27:4733–44.

208. Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. 2016;35:2413–27.

209. Koralewski TE, Krutovsky KV. Evolution of exon-intron structure and alternative splicing. PloS One. 2011;6:e18055.

210. Friedel RH, Friedel CC, Bonfert T, Shi R, Rad R, Soriano P. Clonal expansion analysis of transposon insertions by high-throughput sequencing identifies candidate cancer genes in a PiggyBac mutagenesis screen. PloS One. 2013;8:e72338.

211. Matsubara A, Sekine S, Kushima R, Ogawa R, Taniguchi H, Tsuda H, et al. Frequent GNAS and KRAS mutations in pyloric gland adenoma of the stomach and duodenum. J Pathol. 2013;229:579–87.

212. Rodriguez JA, Huerta-Yepez S, Law IKM, Baay-Guzman GJ, Tirado-Rodriguez B, Hoffman JM, et al. Diminished Expression of Corticotropin-Releasing Hormone Receptor 2 in Human Colon Cancer Promotes Tumor Growth and Epithelial-to-Mesenchymal Transition via Persistent Interleukin-6/Stat3 Signaling. Cell Mol Gastroenterol Hepatol. 2015;1:610–30.

213. Dasgupta S, Rajapakshe K, Zhu B, Nikolai BC, Yi P, Putluri N, et al. Metabolic enzyme PFKFB4 activates transcriptional coactivator SRC-3 to drive breast cancer. Nature. 2018;556:249.

214. Thaker AI, Shaker A, Rao MS, Ciorba MA. Modeling colitis-associated cancer with azoxymethane (AOM) and dextran sulfate sodium (DSS). J Vis Exp JoVE. 2012.

215. Gao C, Wang Y, Broaddus R, Sun L, Xue F, Zhang W. Exon 3 mutations of CTNNB1 drive tumorigenesis: a review. Oncotarget. 2017;9:5492–508.

216. Hollander MC, Blumenthal GM, Dennis PA. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. Nat Rev Cancer. 2011;11:289–301.

217. Donnelly SK, Cabrera R, Mao SPH, Christin JR, Wu B, Guo W, et al. Rac3 regulates breast cancer invasion and metastasis by controlling adhesion and matrix degradation. J Cell Biol. 2017;216:4331–49.

218. Greuber EK, Smith-Pearson P, Wang J, Pendergast AM. Role of ABL Family Kinases in Cancer: from Leukemia to Solid Tumors. Nat Rev Cancer. 2013;13:559–71.

219. Khatri A, Wang J, Pendergast AM. Multifunctional Abl kinases in health and disease. J Cell Sci. 2016;129:9–16.

220. Rafa H, Benkhelifa S, AitYounes S, Saoula H, Belhadef S, Belkhelfa M, et al. All-Trans Retinoic Acid Modulates TLR4/NF-κB Signaling Pathway Targeting TNF-α and Nitric Oxide Synthase 2 Expression in Colonic Mucosa during Ulcerative Colitis and Colitis Associated Cancer. Mediators Inflamm. 2017;2017.

221. Williams CS, Bernard JK, Demory Beckler M, Almohazey D, Washington MK, Smith JJ, et al. ERBB4 is over-expressed in human colon cancer and enhances cellular transformation. Carcinogenesis. 2015;36:710–8.

222. Röcken C, Behrens H-M, Böger C, Krüger S. Clinicopathological characteristics of RHOA mutations in a Central European gastric cancer cohort. J Clin Pathol. 2016;69:70–5.

223. Krøigård AB, Thomassen M, Lænkholm A-V, Kruse TA, Larsen MJ. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. PLOS ONE. 2016;11:e0151664.

224. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med. 2013;5:91.

225. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A. 2015;112:5473–8.

226. Jensen LB, Bartlett JMS, Witton CJ, Kirkegaard T, Brown S, Müller S, et al. Frequent amplifications and deletions of G1/S-phase transition genes, CCND1 and MYC in early breast cancers: a potential role in G1/S escape. Cancer Biomark Sect Dis Markers. 2009;5:41–9.

227. Waerner T, Gardellin P, Pfizenmaier K, Weith A, Kraut N. Human RERE is localized to nuclear promyelocytic leukemia oncogenic domains and enhances apoptosis. Cell Growth Differ Mol Biol J Am Assoc Cancer Res. 2001;12:201–10.

228. Chadwick RB, Jiang GL, Bennington GA, Yuan B, Johnson CK, Stevens MW, et al. Candidate tumor suppressor RIZ is frequently involved in colorectal carcinogenesis. Proc Natl Acad Sci U S A. 2000;97:2662–7.

229. Guo W-J, Zeng M-S, Yadav A, Song L-B, Guo B-H, Band V, et al. Mel-18 acts as a tumor suppressor by repressing Bmi-1 expression and down-regulating Akt activity in breast cancer cells. Cancer Res. 2007;67:5083–9.

230. Chen C, Liu Y, Rappaport AR, Kitzing T, Schultz N, Zhao Z, et al. MLL3 Is a Haploinsufficient 7q Tumor Suppressor in Acute Myeloid Leukemia. Cancer Cell. 2014;25:652–65.

231. Barber GN. VSV-tumor selective replication and protein translation. Oncogene. 2005;24:7710–9.

232. Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome Res. 2014.

233. Kozeretska IA, Demydov SV, Ostapchenko LI. Mobile genetic elements and cancer. From mutations to gene therapy. Exp Oncol. 2011;33:198–205.

234. Ohnishi S, Ma N, Thanan R, Pinlaor S, Hammam O, Murata M, et al. DNA damage in inflammation-related carcinogenesis and cancer stem cells. Oxid Med Cell Longev. 2013;2013:387014.

235. Gu H, Huang T, Shen Y, Liu Y, Zhou F, Jin Y, et al. Reactive Oxygen Species-Mediated Tumor Microenvironment Transformation: The Mechanism of Radioresistant Gastric Cancer. Oxidative Medicine and Cellular Longevity. 2018.

236. Goldstone R, Itzkowitz S, Harpaz N, Ullman T. Dysplasia is more common in the distal than proximal colon in ulcerative colitis surveillance. Inflamm Bowel Dis. 2012;18:832–7.

237. Kim JJ, Shajib MdS, Manocha MM, Khan WI. Investigating Intestinal Inflammation in DSS-induced Model of IBD. J Vis Exp JoVE. 2012.

238. Egger B, Bajaj-Elliott M, MacDonald TT, Inglin R, Eysselein VE, Büchler MW. Characterisation of acute murine dextran sodium sulphate colitis: cytokine profile and dose dependency. Digestion. 2000;62:240–8.

239. Brooks AN, Turkarslan S, Beer KD, Lo FY, Baliga NS. Adaptation of cells to new environments. Wiley Interdiscip Rev Syst Biol Med. 2011;3:544–61.

240. Gilad Y, Oshlack A, Rifkin SA. Natural selection on gene expression. Trends Genet TIG. 2006;22:456–61.

241. Turner BM. Epigenetic responses to environmental change and their evolutionary implications. Philos Trans R Soc Lond B Biol Sci. 2009;364:3403–18.

242. Gerstung M, Pellagatti A, Malcovati L, Giagounidis A, Porta MGD, Jädersten M, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nat Commun. 2015;6:5901.

243. Loeb KR, Loeb LA. Significance of multiple mutations in cancer. Carcinogenesis. 2000;21:379–85.

244. Ferguson LR. Chronic inflammation and mutagenesis. Mutat Res. 2010;690:3–11.

245. Galhardo RS, Hastings PJ, Rosenberg SM. Mutation as a Stress Response and the Regulation of Evolvability. Crit Rev Biochem Mol Biol. 2007;42:399–435.

246. Suaeyun R, Kinouchi T, Arimochi H, Vinitketkumnuen U, Ohnishi Y. Inhibitory effects of lemon grass (Cymbopogon citratus Stapf) on formation of azoxymethane-induced DNA adducts and aberrant crypt foci in the rat colon. Carcinogenesis. 1997;18:949–55.

247. Andrianova M, Bazykin GA, Nikolaev S, Seplyarskiy V. Human mismatch repair system corrects errors produced during lagging strand replication more effectively. 2016.

248. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.

249. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488:504–7.

250. Frank AC, Lobry JR. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene. 1999;238:65–77.

251. Segovia R, Tam AS, Stirling PC. Dissecting genetic and environmental mutation signatures with model organisms. Trends Genet. 2015;31:465–74.

252. Kinugasa T, Akagi Y. Status of colitis-associated cancer in ulcerative colitis. World J Gastrointest Oncol. 2016;8:351–7.

253. Risques RA, Lai LA, Himmetoglu C, Ebaee A, Li L, Feng Z, et al. Ulcerative colitis-associated colorectal cancer arises in a field of short telomeres, senescence, and inflammation. Cancer Res. 2011;71:1669–79.

254. Tukaj S, Węgrzyn G. Anti-Hsp90 therapy in autoimmune and inflammatory diseases: a review of preclinical studies. Cell Stress Chaperones. 2016;21:213–8.

255. Csermely P, Schnaider T, Soti C, Prohászka Z, Nardai G. The 90-kDa molecular chaperone family: structure, function, and clinical applications. A comprehensive review. Pharmacol Ther. 1998;79:129–68.

256. Andersen NN, Pasternak B, Basit S, Andersson M, Svanström H, Caspersen S, et al. Association Between Tumor Necrosis Factor-α Antagonists and Risk of Cancer in Patients With Inflammatory Bowel Disease. JAMA. 2014;311:2406–13.

257. Grossi V, Peserico A, Tezil T, Simone C. p38α MAPK pathway: A key factor in colorectal cancer therapy and chemoresistance. World J Gastroenterol WJG. 2014;20:9744–58.

258. Wagner EF, Nebreda ÁR. Signal integration by JNK and p38 MAPK pathways in cancer development. Nat Rev Cancer. 2009;9:537–49.

259. Gupta J, del Barco Barrantes I, Igea A, Sakellariou S, Pateras IS, Gorgoulis VG, et al. Dual function of p38α MAPK in colon cancer: suppression of colitis-associated tumor initiation but requirement for cancer cell survival. Cancer Cell. 2014;25:484–500.

260. Wakeman D, Schneider JE, Liu J, Wandu WS, Erwin CR, Guo J, et al. Deletion of p38-alpha MAPK within the Intestinal Epithelium Promotes Colon Tumorigenesis. Surgery. 2012;152:286–93.

261. Lu J, Zeng H, Liang Z, Chen L, Zhang L, Zhang H, et al. Network modelling reveals the mechanism underlying colitis-associated colon cancer and identifies novel combinatorial anti-cancer targets. Sci Rep. 2015;5:14739.

262. Burotto M, Chiou VL, Lee J-M, Kohn EC. The MAPK pathway across different malignancies: A new perspective. Cancer. 2014;120:3446–56.

263. Grivennikov SI. Inflammation and colorectal cancer: colitis-associated neoplasia. Semin Immunopathol. 2013;35:229–44.

264. Romano M, Francesco FD, Zarantonello L, Ruffolo C, Ferraro GA, Zanus G, et al. From Inflammation to Cancer in Inflammatory Bowel Disease: Molecular Perspectives. Anticancer Res. 2016;36:1447–60.

265. Sherwood V. WNT Signaling: an Emerging Mediator of Cancer Cell Metabolism? Mol Cell Biol. 2015;35:2–10.

266. Shenoy AK, Fisher RC, Butterworth EA, Pi L, Chang L-J, Appelman HD, et al. Transition From Colitis to Cancer: High Wnt Activity Sustains The Tumor-Initiating Potential Of Colon Cancer Stem Cell Precursors. Cancer Res. 2012;72:5091–100.

267. Claessen MMH, Schipper MEI, Oldenburg B, Siersema PD, Offerhaus GJA, Vleggaar FP. WNT-pathway activation in IBD-associated colorectal carcinogenesis: potential biomarkers for colonic surveillance. Cell Oncol Off J Int Soc Cell Oncol. 2010;32:303–10.

268. Kanaan Z, Qadan M, Eichenberger MR, Galandiuk S. The Actin-Cytoskeleton Pathway and Its Potential Role in Inflammatory Bowel Disease-Associated Human Colorectal Cancer. Genet Test Mol Biomark. 2010;14:347–53.

269. Mees ST, Mennigen R, Spieker T, Rijcken E, Senninger N, Haier J, et al. Expression of tight and adherens junction proteins in ulcerative colitis associated colorectal carcinoma: upregulation of claudin-1, claudin-3, claudin-4, and beta-catenin. Int J Colorectal Dis. 2009;24:361–8.

270. Pekow J, Hutchison AL, Meckel K, Harrington K, Deng Z, Talasila N, et al. miR-4728-3p Functions as a Tumor Suppressor in Ulcerative Colitis-associated Colorectal Neoplasia Through Regulation of Focal Adhesion Signaling. Inflamm Bowel Dis. 2017;23:1328–37.

271. Buda A, Pignatelli M. Cytoskeletal network in colon cancer: from genes to clinical application. Int J Biochem Cell Biol. 2004;36:759–65.

272. Artis D. Epithelial-cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. Nat Rev Immunol. 2008;8:411–20.

273. Dejea C, Wick E, Sears CL. Bacterial oncogenesis in the colon. Future Microbiol. 2013;8:445–60.

274. Yang Y, Jobin C. Novel insights into microbiome in colitis and colorectal cancer. Curr Opin Gastroenterol. 2017;33:422–7.

275. Richard ML, Liguori G, Lamas B, Brandi G, da Costa G, Hoffmann TW, et al. Mucosa-associated microbiota dysbiosis in colitis associated cancer. Gut Microbes. 2018;9:131–42.

276. Sobczak M, Wlazłowski M, Zatorski H, Sałaga M, Fichna J. Current overview of colitis-associated colorectal cancer. Open Life Sci. 2014;9.

277. Warburg O. On the origin of cancer cells. Science. 1956;123:309–14.

278. Dang CV. Links between metabolism and cancer. Genes Dev. 2012;26:877–90.

279. Phan LM, Yeung S-CJ, Lee M-H. Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies. Cancer Biol Med. 2014;11:1–19.

280. Boroughs LK, DeBerardinis RJ. Metabolic pathways promoting cancer cell survival and growth. Nat Cell Biol. 2015;17:351–9.

281. Yang J, Du X lin, Li S ting, Wang B yuan, Wu Y ying, Chen Z ling, et al. Characteristics of Differently Located Colorectal Cancers Support Proximal and Distal Classification: A Population-Based Study of 57,847 Patients. PLoS ONE. 2016;11.

282. Minoo. Characterization of rectal, proximal and distal colon cancers based on clinicopathological, molecular and protein profiles. Int J Oncol. 2010;37.

283. Loupakis F, Yang D, Yau L, Feng S, Cremolini C, Zhang W, et al. Primary Tumor Location as a Prognostic Factor in Metastatic Colorectal Cancer. J Natl Cancer Inst. 2015;107:dju427.

284. Purcell RV, Visnovska M, Biggs PJ, Schmeier S, Frizelle FA. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. Sci Rep. 2017;7:11590.

285. Ciechanover A, Orian A, Schwartz AL. The ubiquitin-mediated proteolytic pathway: Mode of action and clinical implications. J Cell Biochem. 2000;77:40–51.

286. Lu P, Liu R, Ma E-M, Yang T-J, Liu J-L. Functional Analysis of B7-H3 in Colonic Carcinoma Cells. Asian Pac J Cancer Prev. 2012;13:3899–903.

287. Voutsadakis IA. The ubiquitin-proteasome system in colorectal cancer. Biochim Biophys Acta BBA - Mol Basis Dis. 2008;1782:800–8.

288. Russ AE, Peters JS, McNabb WC, Barnett MPG, Anderson RC, Park Z, et al. Gene Expression Changes in the Colon Epithelium Are Similar to Those of Intact Colon during Late Inflammation in Interleukin-10 Gene Deficient Mice. PLoS ONE. 2013;8:e63251.

289. Wu D, Li Q, Song G, Lu J. Identification of disrupted pathways in ulcerative colitis-related colorectal carcinoma by systematic tracking the dysregulated modules. J BUON Off J Balk Union Oncol. 2016;21:366–74.

290. Guo Y, Xiong J, Wang J, Wen J, Zhi F. Inhibition of Rac family protein impairs colitis and colitis-associated cancer in mice. Am J Cancer Res. 2018;8:70–80.

291. Han J, Xi Q, Meng Q, Liu J, Zhang Y, Han Y, et al. Interleukin-6 promotes tumor progression in colitis‑associated colorectal cancer through HIF-1α regulation. Oncol Lett. 2016;12:4665–70.

292. Grivennikov S, Karin E, Terzic J, Mucida D, Yu G-Y, Vallabhapurapu S, et al. IL-6 and STAT3 are required for survival of intestinal epithelial cells and development of colitis associated cancer. Cancer Cell. 2009;15:103–13.

293. Heikkilä K, Ebrahim S, Lawlor DA. Systematic review of the association between circulating interleukin-6 (IL-6) and cancer. Eur J Cancer Oxf Engl 1990. 2008;44:937–45.

294. Greer EL, Brunet A. FOXO transcription factors at the interface between longevity and tumor suppression. Oncogene. 2005;24:7410–25.

295. Flachsbart F, Caliebe A, Kleindorp R, Blanché H, von Eller-Eberstein H, Nikolaus S, et al. Association of FOXO3A variation with human longevity confirmed in German centenarians. Proc Natl Acad Sci U S A. 2009;106:2700–5.

296. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47:979–86.

297. Esteban-Jurado C, Vila-Casadesús M, Garre P, Lozano JJ, Pristoupilova A, Beltran S, et al. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. Genet Med. 2015;17:131–42.

298. Kobayashi M, Watada H, Kawamori R, Maeda S. Overexpression of acetyl-coenzyme A carboxylase beta increases proinflammatory cytokines in cultured human renal proximal tubular epithelial cells. Clin Exp Nephrol. 2010;14:315–24.

299. Yu C, Hong H, Lu J, Zhao X, Hu W, Zhang S, et al. Prediction of Target Genes and Pathways Associated With Cetuximab Insensitivity in Colorectal Cancer. Technol Cancer Res Treat. 2018;17. doi:10.1177/1533033818806905.

300. Manoharan I, Suryawanshi A, Hong Y, Ranganathan P, Shanmugam A, Ahmad S, et al. Homeostatic PPARα Signaling Limits Inflammatory Responses to Commensal Microbiota in the Intestine. J Immunol. 2016;196:4739–49.

301. Wang D, DuBois RN. Peroxisome proliferator-activated receptors in chronic inflammation and colorectal cancer. Gastroenterol Clin North Am. 2010;39:697–707.

302. Gou Q, Gong X, Jin J, Shi J, Hou Y. Peroxisome proliferator-activated receptors (PPARs) are potential drug targets for cancer therapy. Oncotarget. 2017;8:60704–9.

303. Zhan Q, Huang R-F, Liang X-H, Ge M-X, Jiang J-W, Lin H, et al. FRAS1 knockdown reduces A549 cells migration and invasion through downregulation of FAK signaling. Int J Clin Exp Med. 2014;7:1692–7.

304. Xu J, Min W, Liu X, Xie C, Tang J, Yi T, et al. Identification of FRAS1 as a human endometrial carcinoma-derived protein in serum of xenograft model. Gynecol Oncol. 2012;127:406–11.

305. Frey MR, Edelblum KL, Mullane MT, Liang D, Polk DB. The ErbB4 growth factor receptor is required for colon epithelial cell survival in the presence of TNF. Gastroenterology. 2009;136:217–26.

306. Schabbauer G, Matt U, Günzl P, Warszawska J, Furtner T, Hainzl E, et al. Myeloid PTEN Promotes Inflammation but Impairs Bactericidal Activities during Murine Pneumococcal Pneumonia. J Immunol. 2010;185:468–76.

307. Pópulo H, Lopes JM, Soares P. The mTOR signalling pathway in human cancer. Int J Mol Sci. 2012;13:1886–918.

308. Khare V, Dammann K, Asboth M, Krnjic A, Jambrich M, Gasche C. Overexpression of PAK1 Promotes Cell Survival in Inflammatory Bowel Diseases and Colitis-associated Cancer. Inflamm Bowel Dis. 2015;21:287–96.

309. Nagarajan A, Malvi P, Wajapeyee N. Oncogene-Directed Alterations in Cancer Cell Metabolism. Trends Cancer. 2016;2:365–77.

310. Ohtaki N, Yamaguchi A, Goi T, Fukaya T, Takeuchi K, Katayama K, et al. Somatic alterations of the DPC4 and Madr2 genes in colorectal cancers and relationship to metastasis. Int J Oncol. 2001;18:265–70.

311. Isaksson-Mettävainio M, Palmqvist R, Forssell J, Stenling R, Oberg A. SMAD4/DPC4 expression and prognosis in human colorectal cancer. Anticancer Res. 2006;26:507–10.

312. Prévostel C, Rammah-Bouazza C, Trauchessec H, Canterel-Thouennon L, Busson M, Ychou M, et al. SOX9 is an atypical intestinal tumor suppressor controlling the oncogenic Wnt/ß-catenin signaling. Oncotarget. 2016;7:82228–43.

313. Saito M, Yamaguchi A, Goi T, Tsuchiyama T, Nakagawara G, Urano T, et al. Expression of DCC protein in colorectal tumors and its relationship to tumor progression and metastasis. Oncology. 1999;56:134–41.

314. Mees ST, Mardin WA, Wendel C, Baeumer N, Willscher E, Senninger N, et al. EP300--a miRNA-regulated metastasis suppressor gene in ductal adenocarcinomas of the pancreas. Int J Cancer. 2010;126:114–24.

315. Korphaisarn K, Kopetz S. BRAF-directed Therapy in Metastatic Colorectal Cancer. Cancer J Sudbury Mass. 2016;22:175–8.

316. Li Z, Zhu W, Xiong L, Yu X, Chen X, Lin Q. Role of high expression levels of STK39 in the growth, migration and invasion of non-small cell type lung cancer cells. Oncotarget. 2016;7:61366–77.

317. Donati K, Sépult C, Rocks N, Blacher S, Gérard C, Noel A, et al. Neutrophil-Derived Interleukin 16 in Premetastatic Lungs Promotes Breast Tumor Cell Seeding. Cancer Growth Metastasis. 2017;10.

318. Drilon A, Somwar R, Mangatt BP, Edgren H, Desmeules P, Ruusulehto A, et al. Response to ERBB3-Directed Targeted Therapy in NRG1-Rearranged Cancers. Cancer Discov. 2018;8:686–95.

319. Hankey W, Frankel WL, Groden J. Functions of the APC tumor suppressor protein dependent and independent of canonical WNT signaling: implications for therapeutic targeting. Cancer Metastasis Rev. 2018;37:159–72.

320. Resende F, Titze-de-Almeida S, Titze-de-Almeida R. Function of neuronal nitric oxide synthase enzyme in temozolomide-induced damage of astrocytic tumor cells. Oncol Lett. 2018.

321. Vannini F, Kashfi K, Nath N. The dual role of iNOS in cancer. Redox Biol. 2015;6:334–43.

322. Aust DE, Terdiman JP, Willenbucher RF, Chang CG, Molinaro-Clark A, Baretton GB, et al. The APC/beta-catenin pathway in ulcerative colitis-related colorectal carcinomas: a mutational analysis. Cancer. 2002;94:1421–7.

323. Li X-L, Zhou J, Chen Z-R, Chng W-J. p53 mutations in colorectal cancer- molecular pathogenesis and pharmacological reactivation. World J Gastroenterol WJG. 2015;21:84–93.

324. Du L, Kim JJ, Shen J, Chen B, Dai N. KRAS and TP53 mutations in inflammatory bowel disease-associated colorectal cancer: a meta-analysis. Oncotarget. 2017;8:22175–86.

325. Boutin AT, Liao W-T, Wang M, Hwang SS, Karpinets TV, Cheung H, et al. Oncogenic Kras drives invasion and maintains metastases in colorectal cancer. Genes Dev. 2017;31:370–82.

326. Yao Q, An Y, Hou W, Cao Y-N, Yao M-F, Ma N-N, et al. LRP6 promotes invasion and metastasis of colorectal cancer through cytoskeleton dynamics. Oncotarget. 2017;8:109632–45.

327. Smith TA. Mammalian hexokinases and their abnormal expression in cancer. Br J Biomed Sci. 2000;57:170–8.

328. Bosco EE, Mulloy JC, Zheng Y. Rac1 GTPase: a “Rac” of all trades. Cell Mol Life Sci CMLS. 2009;66:370–4.

329. Muise AM, Walters T, Xu W, Shen-Tu G, Guo C-H, Fattouh R, et al. Single nucleotide polymorphisms that increase expression of the guanosine triphosphatase RAC1 are associated with ulcerative colitis. Gastroenterology. 2011;141:633–41.

330. Bid HK, Roberts RD, Manchanda PK, Houghton PJ. RAC1: An Emerging Therapeutic Option for Targeting Cancer Angiogenesis and Metastasis. Mol Cancer Ther. 2013;12.

331. Espina C, Céspedes MV, García-Cabezas MA, del Pulgar MTG, Boluda A, Oroz LG, et al. A Critical Role for Rac1 in Tumor Progression of Human Colorectal Adenocarcinoma Cells. Am J Pathol. 2008;172:156–66.

332. Lobry C, Oh P, Aifantis I. Oncogenic and tumor suppressor functions of Notch in cancer: it’s NOTCH what you think. J Exp Med. 2011;208:1931–5.

333. Qiao L, Wong BCY. Role of Notch signaling in colorectal cancer. Carcinogenesis. 2009;30:1979–86.

334. Capaccione KM, Pine SR. The Notch signaling pathway as a mediator of tumor survival. Carcinogenesis. 2013;34:1420–30.

335. Piazzi G, D’Argenio G, Prossomariti A, Lembo V, Mazzone G, Candela M, et al. Eicosapentaenoic acid free fatty acid prevents and suppresses colonic neoplasia in colitis-associated colorectal cancer acting on Notch signaling and gut microbiota. Int J Cancer. 2014;135:2004–13.

336. Yuan X, Wu H, Xu H, Xiong H, Chu Q, Yu S, et al. Notch signaling: An emerging therapeutic target for cancer treatment. Cancer Lett. 2015;369:20–7.

337. Kohrman AQ, Matus DQ. Divide or Conquer: Cell Cycle Regulation of Invasive Behavior. Trends Cell Biol. 2017;27:12–25.

338. Zheng H, Lu Z, Wang R, Chen N, Zheng P. Establishing the colitis-associated cancer progression mouse models. Int J Immunopathol Pharmacol. 2016;29:759–63.

339. Edwards FC, Truelove SC. Course and prognosis of ulcerative colitis: Part III Complications. Gut. 1964;5:1–15.

340. Pettan-Brewer C, Treuting PM. Practical pathology of aging mice. Pathobiol Aging Age Relat Dis. 2011;1. doi:10.3402/pba.v1i0.7202.

341. Scaldaferri F, Pizzoferrato M, Lopetuso LR, Musca T, Ingravalle F, Sicignano LL, et al. Nutrition and IBD: Malnutrition and/or Sarcopenia? A Practical Guide. Gastroenterol Res Pract. 2017;2017.

342. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, et al. Host-gut microbiota metabolic interactions. Science. 2012;336:1262–7.

343. Belkaid Y, Hand TW. Role of the microbiota in immunity and inflammation. Cell. 2014;157:121–41.

344. Kim E, Goren A, Ast G. Insights into the connection between cancer and alternative splicing. Trends Genet. 2008;24:7–10.

345. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev. 2010;24:2343–64.

346. Saxena M, Yeretssian G. NOD-Like Receptors: Master Regulators of Inflammation and Cancer. Front Immunol. 2014;5.

347. Elia I, Doglioni G, Fendt S-M. Metabolic Hallmarks of Metastasis Formation. Trends Cell Biol. 2018;28:673–84.

348. Gehren AS, Rocha MR, de Souza WF, Morgado-Díaz JA. Alterations of the apical junctional complex and actin cytoskeleton and their role in colorectal cancer progression. Tissue Barriers. 2015;3:e1017688.

349. Ivanov AI, Parkos CA, Nusrat A. Cytoskeletal Regulation of Epithelial Barrier Function During Inflammation. Am J Pathol. 2010;177:512–24.

350. Nagano M, Hoshino D, Koshikawa N, Akizawa T, Seiki M. Turnover of Focal Adhesions and Cancer Cell Migration. Int J Cell Biol. 2012;2012:e310616.

351. Cox TR, Erler JT. Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer. Dis Model Mech. 2011;4:165–78.

352. Buhard O, Cattaneo F, Wong YF, Yim SF, Friedman E, Flejou J-F, et al. Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors. J Clin Oncol Off J Am Soc Clin Oncol. 2006;24:241–51.

353. Lauren P. THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA. AN ATTEMPT AT A HISTO-CLINICAL CLASSIFICATION. Acta Pathol Microbiol Scand. 1965;64:31–49.

354. Namikawa T, Hanazaki K. Mucin phenotype of gastric cancer and clinicopathology of gastric-type differentiated adenocarcinoma. World J Gastroenterol WJG. 2010;16:4634–9.

6 Supplement

List of supplemental figures

List of supplemental tables

6.1 Supplementary material and methods

6.1.1 Used chemicals, kits, consumables, and devices

Buffer or solution	Composition or company
PBS	8 g/l sodium chloride, 0.2 g/l potassium chloride, 1.56 g/l disodium phosphate, 0.24 g/l monopotassium phosphate, pH = 7.4

Table S 1 Applied media and their composition

Chemical	Company
Azoxymethane	Sigma Aldrich, Munich, Germany
Chloroform	Sigma Aldrich, Munich, Germany
Dextran sulfate sodium	Sigma Aldrich, Munich, Germany
Eosin	Roth, Karlsruhe, Germany
Ethanol	Merck, Darmstadt, Germany
Formalin	Sigma Aldrich, Munich, Germany
Hemalum	Carl Roth, Karlsruhe, Germany
Hematoxylin solution	Th Geyer, Renningen, Germany
Isopropanol	Merck, Darmstadt, Germany
TRIzol	Invitrogen / Life Technologies, Darmstadt, Germany
Xylene	Sigma Aldrich, Munich, Germany

Table S 2 Applied chemicals and their distributing company

Kit	Company
DNeasy® Blood & Tissue kit	Qiagen, Hilden, Germany
Multiplex PCR Master Mix	QIAGEN, Hilden, Germany
QIAamp DNA mini kit	Qiagen, Hilden, Germany
QIAGEN Multiplex PCR Master Mix	Qiagen, Hilden, Germany
Ribo-Zero™ Human/Mouse/Rat rRNA Removal Kit	Illumina, SanDiego, USA
Roti Histokit	Carl Roth, Karlsruhe, Germany
SOLiD paired 50/35 v4 run	Applied Biosystems / Life Technologies, Darmstadt, Germany
SOLiD PCR Kit	Applied Biosystems / Life Technologies, Darmstadt, Germany
Sure Select Target Enrichment Human All Exon v2	Agilent, Santa Clara, USA
SureSelectXT Mouse All Exon kit	Agilent, Santa Clara, USA
TruSeq DNA sample prep kit	Illumina, SanDiego, USA
TruSeq PE Cluster Kit v2.5	Illumina, SanDiego, USA
TruSeq PE Cluster Kit v3	Illumina, SanDiego, USA
TruSeq SBS Kit (200cycles)	Illumina, SanDiego, USA
TruSeq Stranded mRNA Sample Prep Kit	Illumina, SanDiego, USA

Table S 3 Overview of applied kits and their distributing company

Devices / consumables	Company
0.5 ml, 1.5 ml and 2 ml micro tubes	Sarstedt, Nümbrecht, Germany
ABI Prism 310 Genetic Analyzer	Applied Biosystems / Life Technologies, Darmstadt, Germany
Axio Imager Z1	ZEISS, Oberkochen, Germany
cBot	Illumina, SanDiego, USA
Centrifuge for Eppendorf tubes: Fresco 21	Thermo Scientific, Bremen, Germany
Certomat MV, vortex mixer	B. Braun Biotech Internat., Melsungen, Germany
HiSeq 2500	Illumina, SanDiego, USA
Leica RM 2255 microtome	Leica Microsystems, Wetzlar, Germany
NanoDrop ND-1000 spectrophotometer	PeqLab Biotechnologie GmbH, Erlangen, Germany
Pipette tips and filter tips	Sarstedt, Nümbrecht, Germany
Pyromark Q24	Qiagen, Hilden, Germany
Serological pipettes	B. Braun, Melsungen, Germany
SOLiD™ 4 System	Applied Biosystems / Life Technologies, Darmstadt, Germany
Vortex-Genie 2 Variable Speed	Sartorius, Göttingen, Germany

Table S 4 Overview of used devices and consumables including the distributing companies

6.1.2 Microsatellite instability assay (GC study)

Microsatellite instability (MSI) was determined by comparison of the allelic profiles of the mononucleotide repeat markers BAT-25, BAT-26, NR-21, NR-24, and NR-27 in tumor and corresponding control tissue as described in Burhard et al. [352]. In brief, all markers were coamplified in a pentaplex PCR assay with the QIAGEN Multiplex PCR Master Mix (QIAGEN, Hilden, Germany) following the manufacturer´s recommendations for amplification of microsatellite loci. The amplified loci were analyzed on an ABI Prism 310 Genetic Analyzer (Applied Biosystems, Darmstadt, Germany). Samples were judged as MSI, if the tumor showed instability in at least two of the five analyzed microsatellites.

6.1.3 Histology and TNM classification (GC study)

Tissue specimens were fixed in formalin and embedded in paraffin. Deparaffinized sections were stained with hematoxylin and eosin. Tumors were classified according to the Laurén classification [353] and the mucin phenotype [354]. All cases included in the GC study were re-examined by the two surgical pathologists Prof. Dr. Christoph Röcken and Dr. Viktoria Warneke. The pathological TNM-stage of all study patients was determined according to the 7th edition of the UICC guidelines [13] and was based solely on surgical pathological examination including classification of distant metastases (pM-category).

6.1.4 SNV validation in gastric cancer samples

Gene	Mutation	Primer	Sequence
DROSHA	p.Q1089X	DROSHA-PCR-for	5´-GGGAGAAGGAATTTTACAAAACAC-3´
		DROSHA-PCR-rev	5´-TCCAATTGCTTCTTCAAACTCA-3´
		DROSHA-Sequencing	5´-TATTTCTATTTTCCTGTAGC-3´
MSH4	p.A174T	MSH4-PCR-for	5´-TTGTAGAAGGGAGAGGACTTGC-3´
		MSH4-PCR-rev	5´-CCTTTGCATATGTTGTGTTGTCTG-3´
		MSH4-Sequencing	5´-GTTTTTTAAATCAATACTTG -3´
RERE	p.Q500X	RERE-PCR-for	5´-CCCACCCCCACTATGTGC-3´
		RERE-PCR-rev	5´-GGCAGGCGTACCCCTTCA-3´
		RERE-Sequencing	5´-TACCCCTTCAGCTCC-3´
ROS1	p.Q925R	ROS1-PCR-for	5´-CTCTGTTTTGGAACCAGCCAGATT-3´
		ROS1-PCR-rev	5´-AGGGGCTTAAGGGATGTCTGAATA-3´
		ROS1-Sequencing	5´-GAACCAGCCAGATTTAAT-3´
TACC2	p.T502M	TACC2-PCR-for	5´-TTGAGATCCCAGCCAGTGCTAT-3´
		TACC2-PCR-rev	5´-CTTACCACCTCCACCCCTGAA-3´
		TACC2-Sequencing	5´-TGAACTTACGTCTTTAGGG-3´
TYRO3	p.R333H	TYRO-PCR-for	5´-AGGGCAGGGGTCTTAGCAATCT-3´
		TYRO-PCR-rev	5´-ACTTCTTCCCACTCCAAGATGAGG-3´
		TYRO-Sequencing	5´-TGAGGCCTGAATCTGT -3´

Table S 5 PCR and sequencing primers for validation of SNVs in GC study

6.1.5 Parameters for variant calling in human samples

The following parameters were used for the variant calling with GATK v1.3, Samtools v0.1.16 and diBayes:

GATK v1.3 was performed with the following steps and parameters: CountCovariates (training set: dbSNP build 132; covariates: ReadGroupCovariate, QualityScoreCovariate, DinucCovariate, CycleCovariate) → TableRecalibration → AnalyzeCovariates (ignore = 5) → UnifiedGenotyper (stand_call_conf = 50.0; stand_emit_conf = 10.0; dcov = 800; glm = SNV) → VariantRecalibrator (mode = SNV; maxGaussians = 4; percentBadVariants = 0.05; an: QD, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, MQ; training set: hapmap, 1000 Genomes, dbSNV build 132) → ApplyRecalibration (ts_filter_level = 99.0). Small InDels were called with diBayes with medium stringency in the human WES data and Samtools v0.1.16 with default parameters in WGS data. All variants called on chromosome Y were excluded.

For SNV calling with Samtools v0.1.16, the default settings were used.

DiBayes was executed with medium stringency, minimum mapping and pairing quality of eight, minimum color quality value of seven, and a minimum coverage of three reads with different start points. A support of the variant from reads on both strand was not required. Reads with an alignment-length/read-length ratio less than 0.85 were excluded.

6.1.6 Definition of genomic regions

Class	Description
Exonic	Overlaps with coding region
Splicing	Less than 2 bp distance to a splicing junction
ncRNA	Overlap with transcript without coding annotation
UTR5	Overlap with 5' untranslated region
UTR3	Overlaps with 3' untranslated region
Intronic	Overlaps with intron
Upstream (neighborhood)	Overlap with 1-kb region upstream of transcription start site
Downstream (neighborhood)	Overlap with 1-kb region downstream of transcription end site
Intergenic	Intergenic region

Table S 6 Description of genomic regions

6.1.7 Parameters for variant calling in murine samples

In GATK, the following steps were applied before SNV calling to improve the quality of the generated alignments: First, the functions RealignerTargetCreator and InDelRealigner were used to perform a local realignment around InDel positions to minimize the number of mismatching bases across all reads. The subsequently applied recalibration of base quality scores (function BaseRecalibration) was used to correct for quality variation due to machine cycle and sequence bias. SNV positions were called with the analysis type UnifiedGenotyper. The recalibration of the probability values of each SNV call and the quality filtering of the variants was performed with the options VariantRecalibrator and ApplyRecalibrator.

The program VarScan v.2.3.5 with the mpileup2snp option for SNVs and the mpileup2indel setting for InDels was applied to an mpileup file created with Samtools v0.1.19.

To call SNVs and small InDels with Samtools v.0.1.19, first an mpileup file was created with the option mpileup provided by Samtools v.0.1.19. Based on this result the tool bcftools were applied to call potential variant sites. These variants were filtered using the script vcfutils.pl out of the bcftools package. Thereby, the maximum read depth was set to 300.

The programs Breakdancer v1.2 and Pindel v0.2.4d were used with default settings.

6.1.8 Quality filter criteria for large structural variants

All large structural variants were filtered with split reads and read pairs (Figure 2-3). This section describes the criteria, which had to be fulfilled to support a large structural variant. A large insertion was supported by a split read, if one part of the read mapped to the assembled contig (inserted sequence) and one part aligned adjacent to the insert position in the known genomic reference. An interchromosomal translocation was supported by a split read, if both subsequences aligned in less than 200 bp distance from the chromosomal breakpoints, respectively. A split read indicated an inversion, if one subsequence aligned outside of the

inverted region with a maximum distance of 50 bp to the breakpoint, and the other read part mapped within the inverted chromosomal part. Thereby, the alignment positions of the two subsequences were required to be on opposite strands. A large deletion was supported by a split read, if one read part mapped before the deletion start and the second after the deletion end. Thereby, a gap of at least five base pairs between the two alignment positions was required and the alignment positions had to be in less than two base pairs distance to the chromosomal start and end position of the deletion, respectively. A tandem duplication was supported by a split read, if the chromosomal alignment positions of the second hit was before the chromosomal alignment position of the first hit. Thereby, both subsequences had to point to the same end of the chromosome. Moreover, the alignment positions of both subsequences were required to be within the called duplicated region or within the adjacent 20 bp.

A read pair endorsed an insertion, if one read mapped to the known genomic reference sequence in a distance smaller than 800 bp of the detected insert position and the other read aligned to the inserted contig sequence. An interchromosomal translocation was supported by a read pair, if one read mapped close to the first chromosomal breakpoint (< 1000 bp distance) and the second read aligned in the neighborhood of the second chromosomal breakpoint (< 1000 bp distance). A read pair supported an inversion, if both reads were located on the same strand and the alignment positions were less than 700 bp away from the respective breakpoints. A tandem duplication was supported by a read pair, if both reads pointed towards different ends of the same chromosome and not towards each other. Moreover, the alignments positions had to be within or adjacent to the duplicated region (< 200 bp distance). Exclusively tandem duplication of a size ≤ 10,000 bp were investigated.

6.1.9 Workflow of human gastric cancer

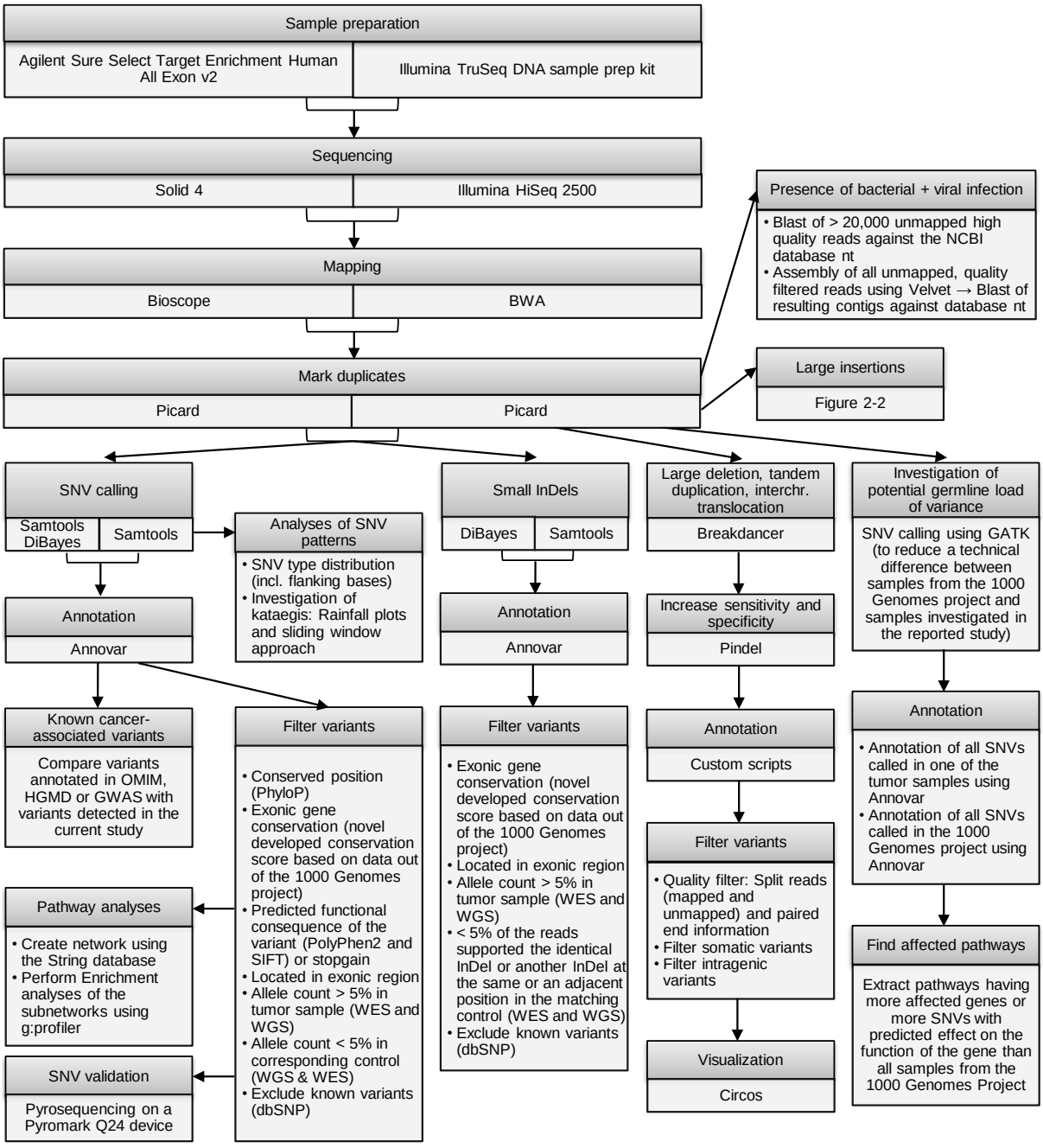


Figure S 1 Study workflow of the GC project

6.2 Supplementary results

6.2.1 Human gastric cancer

6.2.1.1 Sequencing results and variant calling

Exome	Patient 1 tumor	Patient 1 control	Patient 2 tumor	Patient 2 control
# reads	160,307,884	179,749,208	195,373,036	198,844,362
# mapped reads	102,160,406 (63.7%)	128,266,225 (71.4%)	144,208,668 (73.8%)	142,174,101 (71.5%)
Base pairs on target	1,568,736,212 (61.6 %)	2,000,025,200 (61.1 %)	2,308,669,582 (65.0 %)	1,687,354,871 (57.1 %)
# duplicates	24,859,346 (24.3%)	35,662,087 (27.8%)	44,053,956 (30.6%)	54,374,335 (38.2%)
Mean coverage	33.9	43.3	50.0	36.5
Median coverage	28	36	41	34
Mean template size	195.8	201.8	180.9	214.1

Table S 7 WES sequencing statistics

Genome	Patient 1 tumor	Patient 1 control	Patient 2 tumor	Patient 2 control
# reads	2,818,634,692	3,117,880,513	6,269,201,959	3,165,759,194
# mapped reads	2,803,136,338 (99.5%)	2,898,407,733 (93.0%)	5,965,975,293 (95.7%)	2,852,988,076 (90.1%)
# duplicates	1,310,186,458 (44.7%)	424,327,327 (14.6%)	4,287,757,921 (71.9%)	1,433,819,596 (50.3%)
Mean coverage	44.9	24.3	47.3	79.4
Median coverage	47	25	50	84
Mean template size	285.4	237.5	184.6	272.2

Table S 8 WGS sequencing statistics

Sample	# called SNVs
MSI WGS	19,784
MSS WGS	5,116
MSI WES	705
MSS WES	1,847

Table S 9 Number of called somatic SNVs

6.2.1.2 Differences between sequencing technologies and SNV calling programs

To find a valid approach to reduce the false-positive and false-negative rates of reported SNVs, the robustness of different SNV-callers as well as different sequencing strategies were investigated. The SNV caller comparison was carried out for two sequencing approaches and technologies. The WES data were produced with a SOLiD paired 50/35 v4 run. The WGS data were based on the Illumina HiSeq 2500 technology. In the WES data, the SNV caller GATK, diBayes, and Samtools were compared, while in the WGS results, SNVs detected with GATK and Samtools were analyzed. In the WES data, the highest number of SNVs were called with diBayes followed by GATK and Samtools (Figure 3-2). Between 42% and 49% of the WES

SNVs were detected by all calling-programs (diBayes, Samtools, and GATK) and additional ~20% by two callers (Figure S 2, Figure S 3, Figure S 4). 28-36% were exclusively found with diBayes (Figure S 2). In the WGS approach, the overlap between Samtools and GATK was between 65% and 88% (Figure S 2), while Samtools exhibited the highest amount of detected cross platform variants.

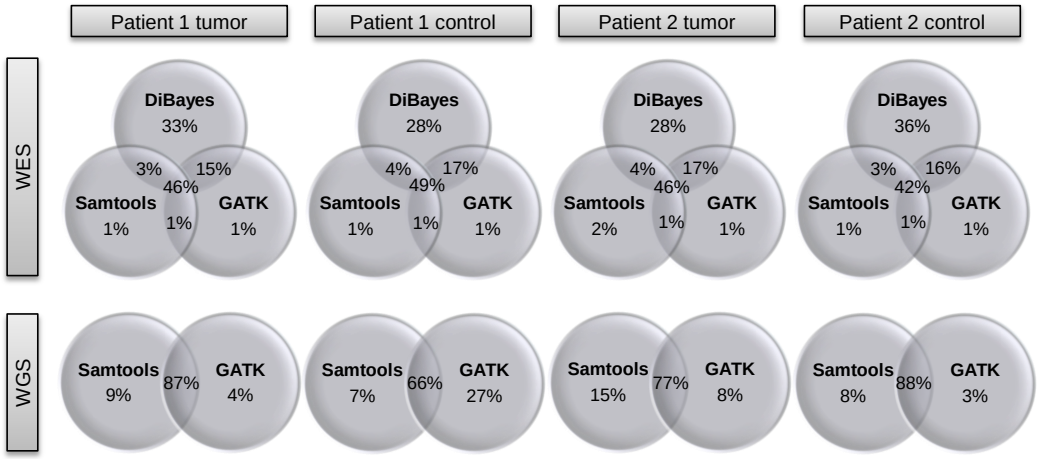


Figure S 2 Overlap of SNVs called by different programs

In the next step, the number of SNVs, which could be validated with the allele counts of the other sequencing technology, respectively, were investigated for each program. An SNV was defined as confirmed, if at least 20% of the reads sequenced with a second technology at this position supported the variant. For this analysis, exclusively positions were considered, which were covered in both sequencing approaches. The highest number of verified variants was observed for SNVs called with GATK. The lowest number could be validated for SNVs called with diBayes (Figure S 3). The results were similar to a 10% read support threshold (data not shown).

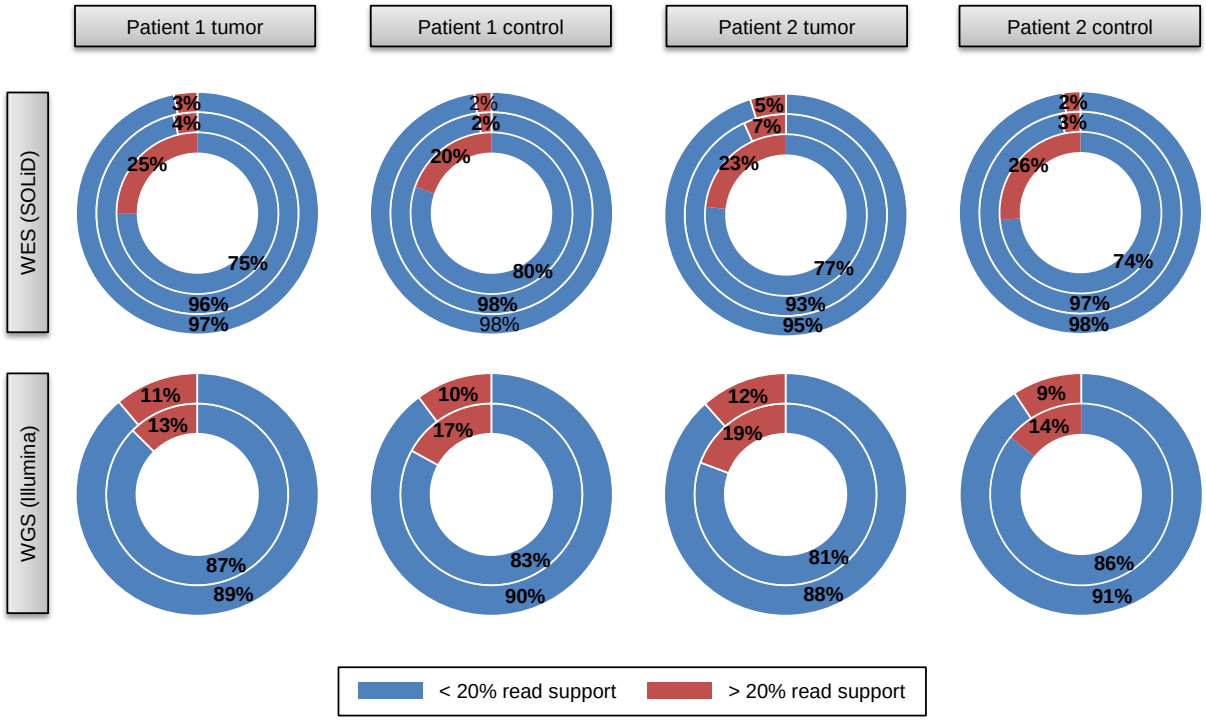
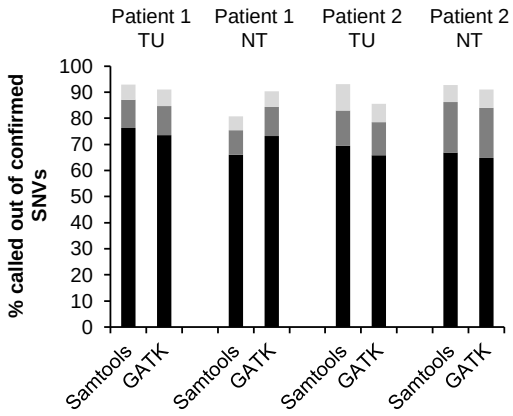


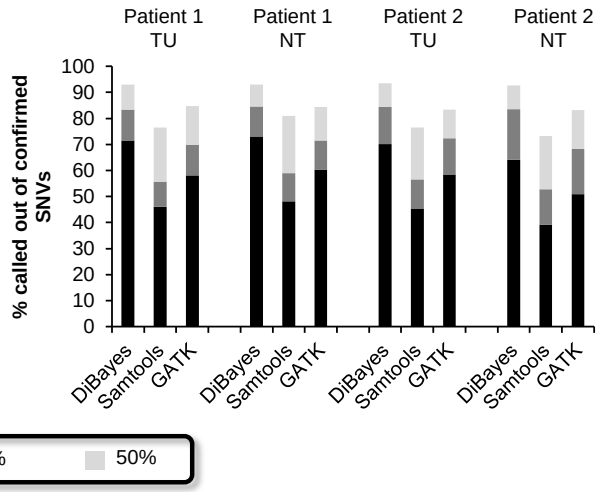
Figure S 3 Specificity of SNV caller
The figure shows the number of verified SNVs for each SNV caller. Exclusively positions covered in both data sets were included in the analysis. In the WES data, the circles indicate from outside to inside the results of the SNV callers GATK, Samtools, and diBayes. In the WGS data, the results of GATK are outside and the one of Samtools inside. The percentages of SNVs, which have less than 20% read support in the other sequencing technology, respectively, are marked in red, the percentages of confirmed SNVs are shown in blue.

Furthermore, it was investigated how many SNVs of all cross platform variants were called by the different programs. Three stringency levels were applied: The SNVs were called confirmed, if in the data of the second technology, (i) 10%, (ii) 20% or (iii) 50% of the reads supported the variant. In the WES approach, the program diBayes detected the highest number of verified SNVs, while this was the case for Samtools in the WGS data (Figure S 4).

A



B



C

	Patient 1 TU	Patient 1 NT	Patient 2 TU	Patient 2 NT
WGS -Samtools	67,426 / 63,229 / 30,967	65,751 / 62,894 / 29,815	67,969 / 64,310 / 34,095	78,203 / 74,402 / 34,643
WGS – GATK	65,010 / 61,420 / 30,324	72,990 / 70,246 / 33,336	64,420 / 60,855 / 31,352	75,691 / 72,530 / 34,048
WES – DiBayes	63,249 / 60,504 / 30,944	72,779 / 70,446 / 34,944	68,645 / 65,459 / 34,216	74,895 / 72,090 / 34,653
WES – Samtools	40,667 / 40,363 / 25,447	47,894 / 49,151 / 36,259	44,308 / 43,826 / 28,385	45,759 / 45,572 / 27,353
WES - GATK	51,276 / 50,577 / 28,209	60,156 / 59,624 / 31,131	57,220 / 56,097 / 31,468	59,430 / 58,950 / 31,120

Figure S 4 Number of detected SNVs per SNV caller
(A)+(B) The figures show for each SNV caller the percentage of called SNVs out of set of confirmed variants. The SNVs were called confirmed, if in the data of the second technology, (i) 10%, (ii) 20% or (iii) 50% of the reads supported the variant. (A) Analyses based on SNVs detected in WGS data. (B) Analysis based on SNVs detected in WES data. (C) Table summarizing the number of confirmed SNV counts.

A comparison between strictly filtered (detected with all callers) SNVs in the WES data (called with Samtools and GATK and diBayes) and those detected in the WGS data (called with Samtools and GATK) is shown in Figure S 5.

Figure S 5 Comparison between strict filtered WES SNVs and WGS SNVs
The figure shows a comparison between strict filtered SNVs called with all three callers (Samtools, GATK, and diBayes) in the WES data and those detected with Samtools and GATK in the WGS data

A large number of SNVs called in the data of one of the sequencing technology could not be confirmed with the second sequencing approach, respectively. Due to a lack of coverage in many intergenic regions in the WES data, more SNVs called in the WGS data could not be confirmed with the WES results. But even in the set of SNVs at positions covered in WES and WGS, the percentage of somatic SNVs called in the WES and validated in the WGS data was markedly higher than the percentage of SNVs called in the WGS and validated in the WES data (Figure S 6).

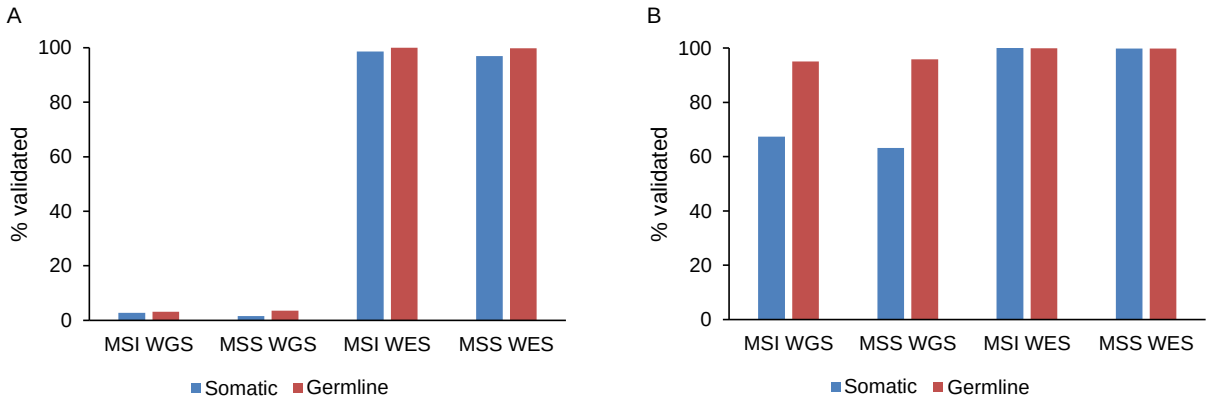


Figure S 6 Comparison between WES and WGS SNV validation rates
Ratio of SNVs, which were called in one sequencing data set and could be validated in the second. (A) Analysis based on the total number of called SNVs. (B) Analysis exclusively based on SNVs, which were covered in both data sets. Please note that in this case, 'somatic' refers to SNVs, which were not detected in the corresponding non-tumor samples sequenced with the same technology.

Next, the NGS approaches were compared, namely WES on the SOLiD 4 and WGS on the Illumina HiSeq 2500. To test the false positive and false negative rate, all exonic SNVs called with Samtools were investigated in all sample pairs (Figure S 7). Around 12% of the intragenic positions, at which an SNV was called in the WGS, were uncovered in the WES data. In the reverse direction, this was for only two SNV positions in one sample the case. The number of SNVs, which were uncovered in the WES, was especially high for C>A, C>G as well as T>G base substitutions and low for T>C variants. Around 5% of WGS SNVs at positions, which were also covered in the WES, could not be confirmed with a 5% allele support threshold. Like for the uncovered SNV positions, the false positive rate was in C>A, C>G, and T>G SNVs higher and in T>C lower than expected. Vice versa, less than 0.7% of SNVs called in the WES data could not be confirmed in the WGS results. In comparison to the total number of called SNVs, unconfirmed SNVs were especially often T>A or T>G base substitutions.

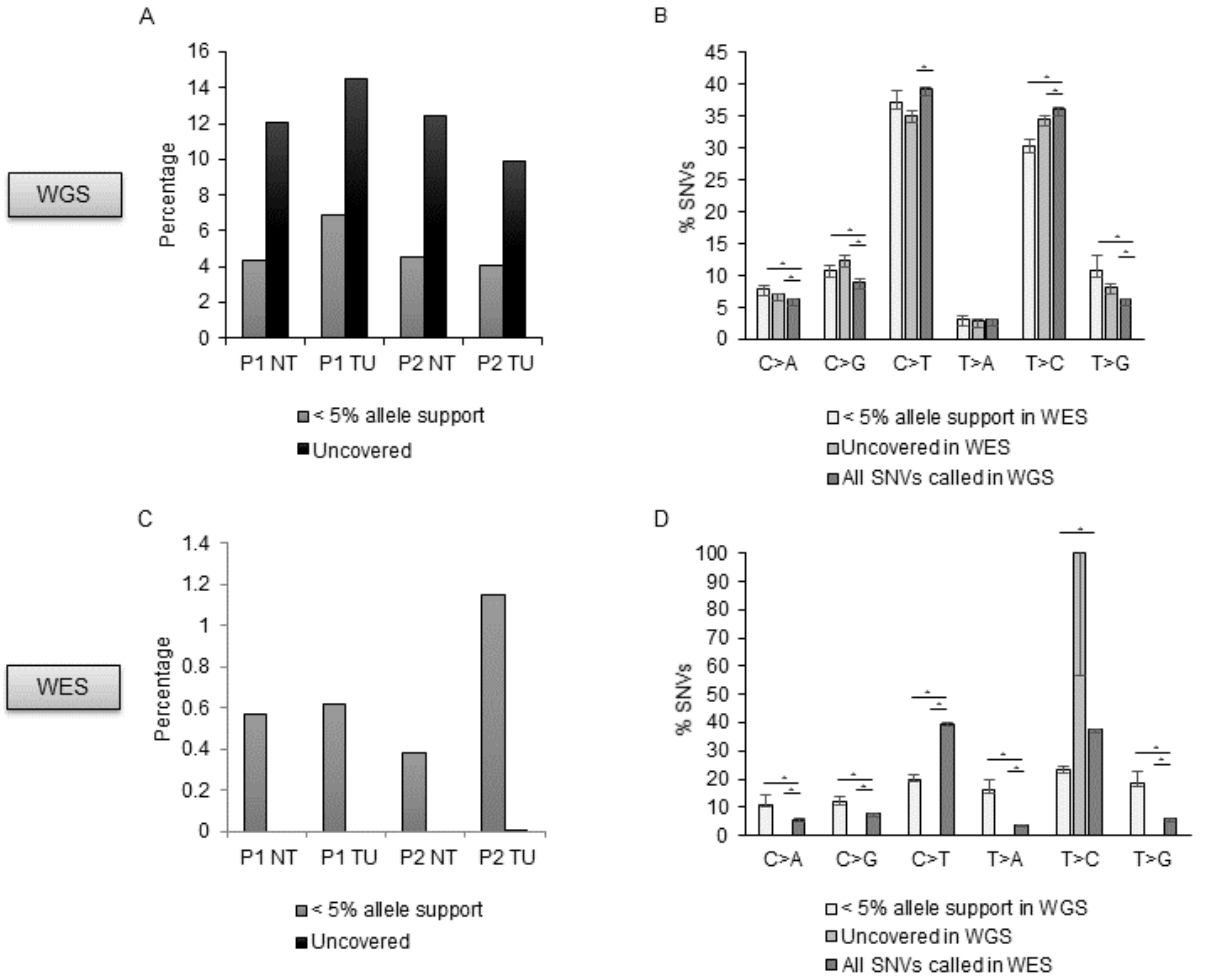


Figure S 7 False positive and false negative rates in WES and WGS data
All exonic SNVs called with Samtools were investigated in all sample pairs and compared between WES and WGS. (A) Percentage of SNVs, which were called in the WGS data and which were either uncovered or not supported in the WES results. (B) SNV type distributions for three classes of SNVs called in the WGS data: All SNVs called in WGS data, SNVs uncovered in WES, and SNVs covered but not supported in WES. (C) Percentage of SNVs, which were called in WES and which were either uncovered or not supported in the WGS data. (D) SNV type distributions for three classes of SNVs called in the WES data: All SNVs called in WES data, SNVs uncovered in WGS, and SNVs covered but not supported in WGS.

External validation with pyrosequencing of a subset of eight SNVs called in the WES data and not supported by the WGS data failed entirely. Six of these somatic SNVs were detected in both WES tumor samples (Table S 18, Table S 19). This clearly point to specific problems in either the SOLiD sequencing platform and / or library preparation protocol. Most of the variants in genes listed in the COSMIC cancer gene census list were called only either in WGS or WES (Table S 10, Table S 11), and only confirmed by mutant allele fractions with the other sequencing approach. This included non-synonymous variants in the genes *PIK3CA*, *JAK2*, *GATA3*, *ROS1*, *ARID1A*, and *BRAF*. Only mutations in *DROSHA* were detected by all methods. All reported observations strengthen the benefit of applying a cross platform approach.

6.2.1.3 Mutational landscape of somatic SNVs and InDels in GC samples

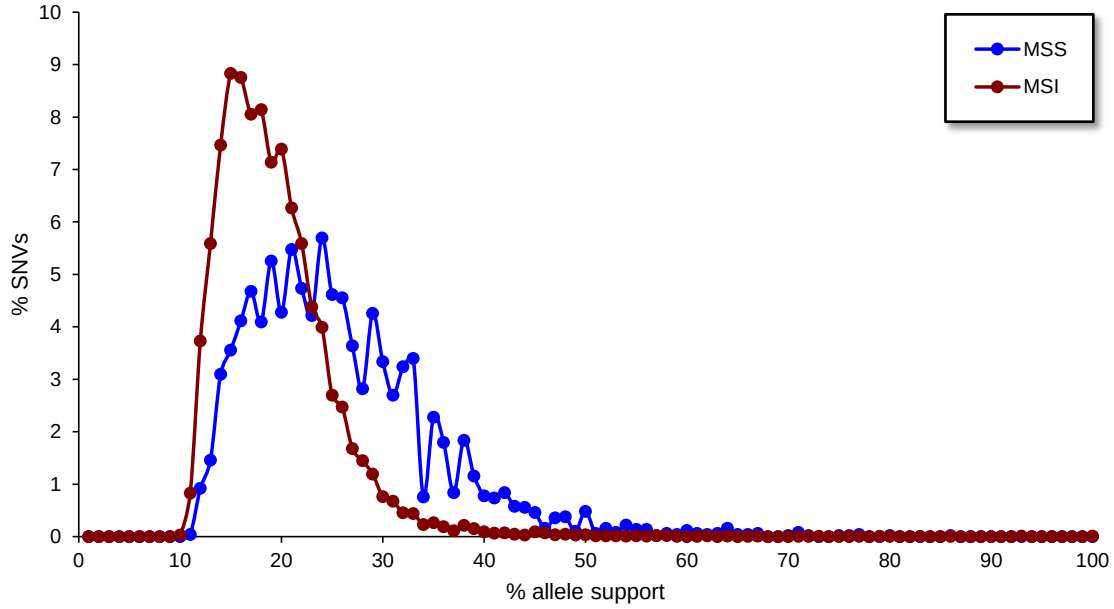


Figure S 8 Estimation of sample purity
Peak shows the mutant allele fractions for heterozygous SNVs. This results in a tumor ratio, which is twice as high as the peak.

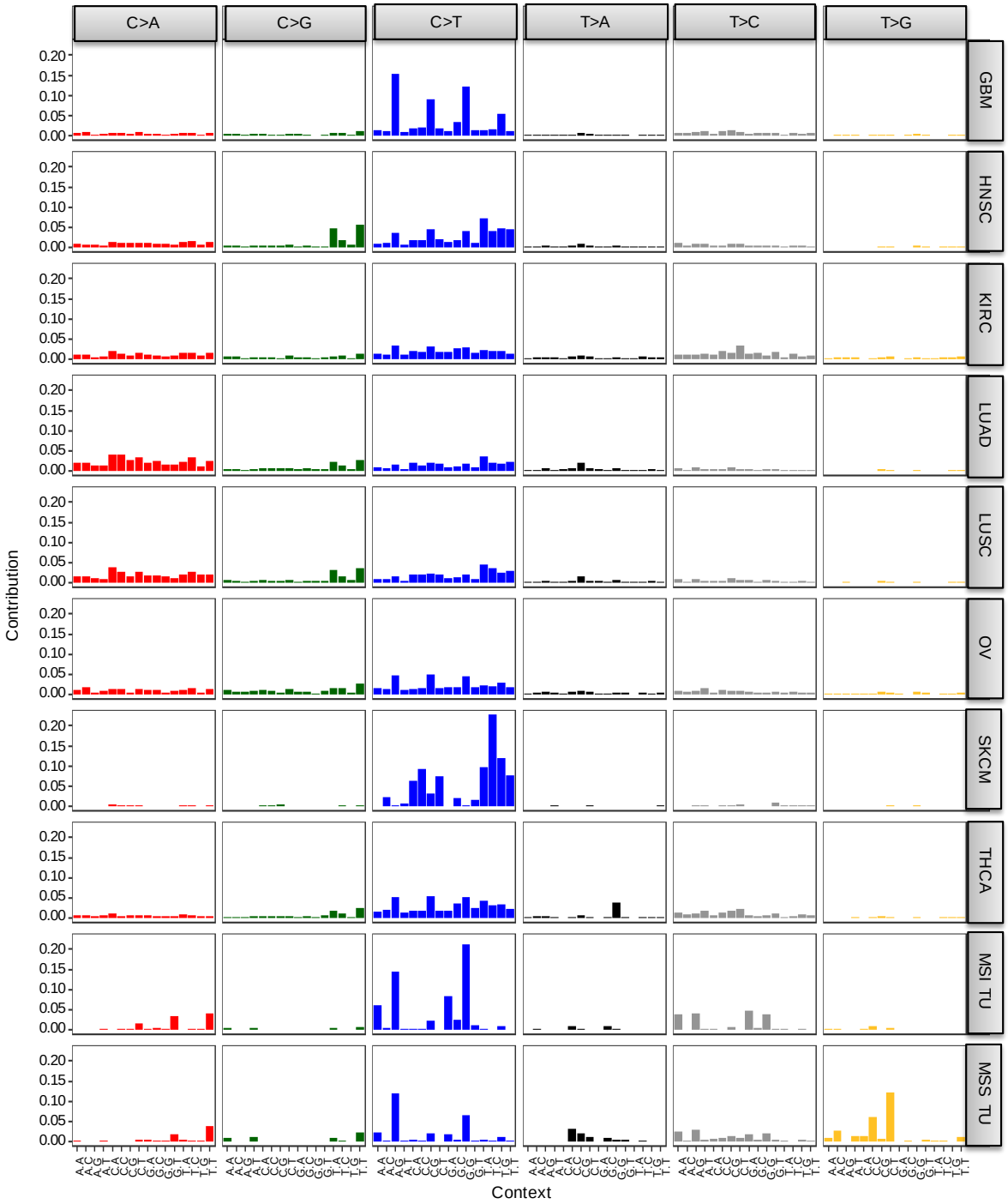


Figure S 9 Comparison between somatic SNV spectrums of investigated GC samples with other cancer types

The somatic SNV spectrums of investigated GC tumor samples were compared with those of TCGA WES samples. MSI TU and MSS TU indicate the SNV type distribution of the investigated tumor samples. The TCGA spectrums include the following cancer types: glioblastoma (GBM), head-neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian cancer (OV), skin cutaneous melanoma (SKCM), and thyroid cancer (THCA).

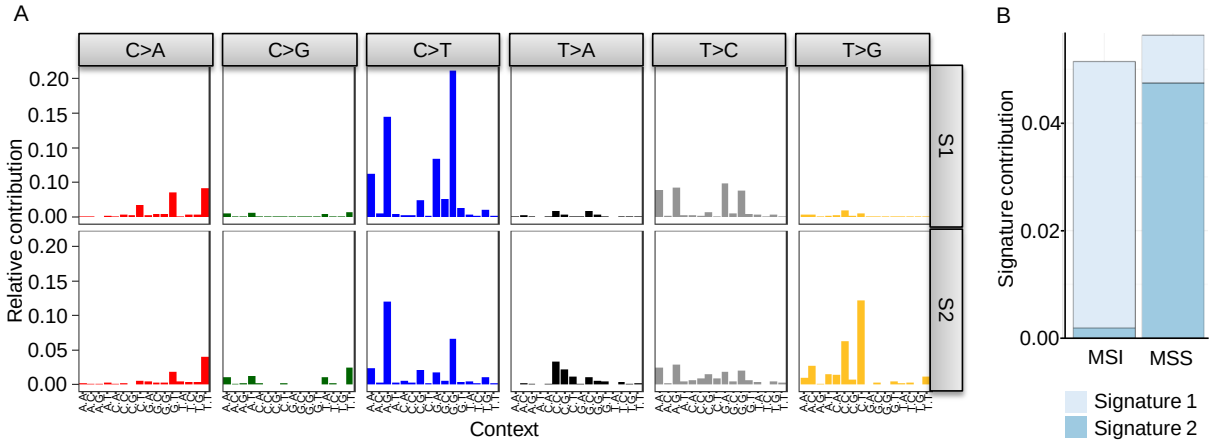


Figure S 10 Identified mutational signatures in GC samples
(A) Description of SNV signatures identified in the investigated tumor samples. (B) Contribution of the identified signatures to the investigated samples.

Chr	Pos	Gene	Effect	Ref nt	Observ. nt	Ref aa	Observ. aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	G-st	E-st	E-bs	COSMIC
1	8424848	*RERE*	nonsense	G	A	Q	X	500	0.17	0.74	0.9996	0.1114	yes	yes	yes	no
1	14108317	*PRDM2*	non-syn	G	A	D	N	1142	0	0.7	0.9998	2.0446	no	no	yes	no
1	19449524	*UBR4*	non-syn	G	A	R	C	3207	0	1	0.9996	0.4922	yes	no	no	no
1	26448895	*PDIK1L*	non-syn	G	A	E	K	285	0.01	1	0.9997	0.0012	no	no	yes	no
1	76272758	*MSH4*	non-syn	G	A	A	T	174	0	1	0.9997	1.0239	no	no	yes	no
1	149859238	*HIST2H2AB*	non-syn	T	C	T	A	77		0.57	0.9976	0	yes	no	no	no
1	156268899	*VHLL*	nonsense	C	A	E	X	28		0.45	0.2208	0.0443	no	no	yes	no
1	161011635	*USF1*	non-syn	G	A	A	V	93	0.01	0.06	0.9996	0.0094	no	yes	yes	no
1	183095328	*LAMC1*	non-syn	G	A	G	S	959	0	0.99	0.9997	1.3632	no	no	yes	no
2	100623269	*AFF3*	non-syn	G	A	A	V	258	0	0.96	0.9996	1.0244	no	no	yes	yes
3	33427020	*FBXL2*	non-syn	G	A	A	T	330	0	0.93	0.9998	0.0168	yes	no	no	no
3	52548457	*STAB1*	non-syn	G	A	R	H	1208	0.01	0.99	0.9996	2.9546	yes	no	no	no
3	107517481	*BBX*	non-syn	G	T	S	I	792	0	1	0.9998	0.048	yes	no	no	no
3	122003133	*CASR*	non-syn	G	A	G	S	778	0.04	0.98	0.9997	1.8264	no	no	yes	no
3	178936082	*PIK3CA*	non-syn	G	A	E	K	542	0.04	0.89	0.9997	0.2407	no	no	yes	yes
4	111397920	*ENPEP*	non-syn	G	A	G	D	117	0	1	0.9997	2.3104	no	no	yes	no
5	31423048	*DROSHA*	nonsense	G	A	Q	X	1052	0.01			0.6416	yes	yes	yes	yes
5	177642326	*AGXT2L2*	nonsense	G	A	Q	X	345	0.35	0.69	0.9767	0.3738	yes	no	no	no
7	99797878	*STAG3*	non-syn	G	T	R	L	566	0	0.18	0.9997	0.6811	yes	no	no	no
9	5126730	*JAK2*	non-syn	G	A	R	H	1113	0	1	0.9995	0.0818	yes	no	no	yes
9	104356830	*PPP3R2*	non-syn	G	A	T	M	128	0.01	0	0.1469	0.0099	yes	no	no	no
9	115805897	*ZFP37*	nonsense	G	T	S	X	334	0.04	0.74	0.9994	1.8508	no	no	yes	no
9	127298179	*NR6A1*	non-syn	G	A	P	S	348	0.02	1	0.9997	0.0768	yes	no	no	no
9	129594868	*ZBTB43*	non-syn	G	A	R	H	27	0	0.98	0.9998	0.0047	yes	no	no	no
10	12191899	*SEC61A2*	non-syn	C	T	T	M	112	0.04	0.25	0.998	0.0064	yes	no	no	no
11	12183953	*MICAL2*	non-syn	G	A	C	Y	84	0	1	0.9998	0.3538	yes	no	no	no
11	73022234	*ARHGEF17*	nonsense	C	T	R	X	851	1	0.72	0.8847	0.6167	no	no	yes	no
12	56755353	*APOF*	nonsense	G	A	R	X	213	1			1.0174	yes	no	no	no
12	69981309	*CCT2*	nonsense	C	T	R	X	10	1	0.73	0.9863	0.1673	yes	no	no	no

Chr	Pos	Gene	Effect	Ref nt	Observ. nt	Ref aa	Observ. aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	G-st	E-st	E-bs	COSMIC
12	109278801	*DAO*	non-syn	G	A	G	R	7	0	1	0.9995	0.0704	yes	no	no	no
14	24839069	*NFATC4*	nonsense	C	A	Y	X	155	0	0.6	0.177	1.3149	no	no	yes	no
15	64506145	*CSNK1G1*	non-syn	T	G	H	P	208	0	1	0.9985	0.0076	no	no	yes	no
18	18619508	*ROCK1*	nonsense	G	A	R	X	326		0.73	0.99	0.136	yes	no	no	no
18	31432849	*NOL4*	non-syn	C	T	R	Q	340	0	0.79	0.9991	0.0073	no	no	yes	no
19	9271885	*ZNF317*	nonsense	C	T	R	X	490	1	0.7	0.7709	0.4575	yes	no	no	no
20	16486763	*KIF16B*	non-syn	G	A	R	C	258	0	1	0.9996	1.5824	yes	no	no	no
20	31671526	*BPIFB4*	nonsense	G	T	G	X	175	0.03	0.73	0.9977	11.229	no	no	yes	no
20	43243244	*PKIG*	non-syn	G	A	R	Q	16	0.04	1	0.9996	0.0157	yes	no	no	no
20	62560874	*DNAJC5*	non-syn	C	T	A	V	106	0.15	0.9	0.9989	0	no	no	yes	no
X	24552101	*PDK3*	non-syn	G	A	R	H	378	0	0.98	0.9997	0.0114	no	yes	yes	no
X	106459930	*CXorf41*	non-syn	G	T	M	I	61	0.34	1	0.9692	0.0029	no	no	yes	no

Table S 10 Somatic SNVs in patient 1 (damaging and conserved)

The table displays all somatic SNVs, which did not exist in the database dbSNP, were called with at least one variant caller, supported by more than 5% of the reads in the WGS as well as in the WES tumor data, and existed in maximum 5% of the reads in the WGS as well as in the WES data of the matching control samples. Furthermore, all SNVs result (i) in a nonsense amino acid change or (ii) were predicted as damaging to the protein function by Sift or PolyPhen2 and were either at a conserved position (PhyloP) or in a conserved gene (ECS). Genes listed in the databases ESP or ExAc are not shown. The following abbreviations were used: Chr = chromosome, Pos = position, Observ. = observed, Ref = reference, nt = nucleotide, aa = amino acid, G-st = called in WGS data with Samtools, E-st = called in WES data with Samtools, E-bs = called in WES data with diBayes / Bioscope, non-syn= non-synonymous.

Chr	Pos	Gene	Effect	Ref nt	Observ. nt	Ref aa	Observ. aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	G-st	E-st	E-bs	COSMIC
10	8106046	*GATA3*	non-syn	T	G	L	R	290	0	0.98	0.9979	0.0018	No	no	yes	yes
14	100363521	*EML1*	nonsense	C	G	Y	X	239	0	0.74	0.9984	1.7618	Yes	yes	yes	no
16	81398615	*GAN*	non-syn	G	A	A	T	425	0.01	0.74	0.9996	0.0272	yes	yes	no	no
17	74276151	*QRICH2*	non-syn	G	A	P	S	1405	0.01	1	0.9996	5.6252	no	no	yes	no
19	46094563	*GPR4*	non-syn	C	T	V	M	188	0.02	0.01	0.9979	0.0055	yes	no	no	no
20	377023	*TRIB3*	non-syn	G	C	A	P	256	0.01	0.84	0.9997	0.8909	no	no	yes	no

Table S 11 Somatic SNVs in patient 2 (damaging and conserved)

The table displays all somatic SNVs, which did not exist in the database dbSNP, were called with at least one variant caller, supported by more than 5% of the reads in the WGS as well as in the WES tumor data, and existed in maximum 5% of the reads in the WGS as well as in the WES data of the matching control samples. Furthermore, all SNVs result (i) in a nonsense amino acid change or (ii) were predicted as damaging by Sift or PolyPhen2 and were either at a conserved position (PhyloP) or in a conserved gene (ECS). Genes listed in the databases ESP or ExAc are not shown. The following abbreviations were used: Chr = chromosome, Pos = position, Observ. = observed, Ref = reference, nt = nucleotide, aa = amino acid, G-st = called in WGS data with Samtools, E-st = called in WES data with Samtools, E-bs = called in WES data with diBayes / Bioscope, non-syn = non-synonymous.

Chr	Pos	Gene	Effect	Ref nt	Observ. nt	Ref aa	Observ. aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	G-st	E-st	E-bs	COSMIC
1	42047914	*HIVEP3*	non-syn	T	G	N	T	852	0	0.99	0.9976	3.1492	no	no	yes	no
1	75805297	*SLC44A5*	non-syn	T	C	D	G	24	0	1	0.9983	0.0532	no	no	yes	no
1	113246418	*RHOC*	non-syn	C	T	A	T	2	0.03	0	0.9984	0.0243	no	no	yes	no
1	153615804	*CHTOP*	non-syn	C	T	R	C	170	0	0.85	0.9993	0.07	yes	no	no	no
1	156146493	*SEMA4A*	non-syn	C	A	P	Q	532	0.04	0.64	0.9979	0.2635	no	no	yes	no
1	183087260	*LAMC1*	non-syn	C	T	R	C	657	0	1	0.9979	1.3632	yes	yes	yes	no
1	222832120	*MIA3*	non-syn	C	T	A	V	1555	0.13	0.98	0.9991	1.5735	no	yes	yes	no
2	21233882	*APOB*	non-syn	T	C	H	R	1953	0.88	0.82	0.2371	2.6255	yes	no	yes	no
2	71654216	*ZNF638*	non-syn	A	G	I	M	1739	0	0.82	0.9145	1.2099	yes	no	no	no
2	103068269	*IL18RAP*	non-syn	A	C	R	S	476	0	1	0.9392	0.0611	yes	yes	no	no
2	170088278	*LRP2*	non-syn	A	G	C	R	1725	0	0.88	0.9986	1.3948	no	no	yes	no
2	208994242	*CRYGC*	non-syn	G	A	R	W	59	0.1	1	0.9392	0.2053	yes	yes	yes	no
2	238275374	*COL6A3*	non-syn	G	A	A	V	1212		0.57	0.9994	2.0665	yes	no	no	no
3	9780797	*BRPF1*	non-syn	T	G	S	R	238	0.02	0.02	0.792	0.0143	yes	no	no	no
3	47098793	*SETD2*	non-syn	G	A	H	Y	2161	0	1	0.9957	0.871	no	yes	yes	yes
3	111697936	*ABHD10*	non-syn	G	A	A	T	10	0.06	0.46	0.9877	0.3632	yes	no	yes	no
3	119451232	*C3orf15*	non-syn	G	T	Q	H	370	0	0.99	0.9684	1.6671	no	no	yes	no
3	127335771	*MCM2*	non-syn	C	T	A	V	528	0	1	0.9989	0.2061	yes	no	no	no
3	183210369	*KLHL6*	non-syn	T	C	T	A	493	0.04	0.02	0.8691	0.0498	yes	no	yes	no
4	81124574	*PRDM8*	non-syn	C	T	A	V	653	0.02	0.13	0.9983	0.9291	yes	no	no	no
4	129809878	*SCLT1*	non-syn	G	A	R	C	654	0	1	0.9995	2.3453	yes	no	no	no
5	1798796	*MRPL36*	non-syn	C	T	R	Q	85	0.17	0.95	0.9467	0.2744	no	no	yes	no
5	71739960	*ZNF366*	non-syn	C	T	E	K	620	0.06	0.78	0.9994	0.6001	no	no	yes	no
5	140166393	*PCDHA1*	non-syn	G	T	S	I	173	0.01	0.55	0.8597	2.0306	no	yes	yes	no
5	140167635	*PCDHA1*	non-syn	C	T	A	V	587	0.12	0.5	0.882	2.0306	yes	no	no	no
6	26091725	*HFE*	non-syn	G	A	R	Q	87	0.08	1	0.7991	0.604	no	no	yes	no
6	109774937	*MICAL1*	non-syn	C	T	V	M	124	0	0.25	0.999	2.9512	yes	no	no	no
7	5545126	*FBXL18*	non-syn	C	T	V	I	52	0	0.73	0.9975	0.0464	yes	no	no	no
7	103029485	*SLC26A5*	non-syn	G	T	A	D	463	0	0.99	0.9995	0.0526	yes	no	no	no
7	148288147	*C7orf33*	non-syn	G	A	A	T	44	0	0.01	0.9597	0.6387	no	no	yes	no
7	150647363	*KCNH2*	non-syn	G	A	P	L	424	0	0.95	0.9987	0.3248	yes	yes	yes	no
7	151962282	*MLL3*	non-syn	G	A	A	V	342	0.29	0.96	0.9989	0.2818	no	no	yes	no

Chr	Pos	Gene	Effect	Ref nt	Observ. nt	Ref aa	Observ. aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	G-st	E-st	E-bs	COSMIC
8	10583282	*SOX7*	non-syn	G	A	T	M	378	0	1	0.9994	0.1973	yes	no	no	no
8	68044263	*CSPP1*	non-syn	C	T	P	L	587	0.02	0.42	0.8906	0.4499	yes	no	no	no
9	32635480	*TAF1L*	non-syn	C	T	G	D	33		0.84	0.9092	0.2873	yes	no	no	no
9	96846997	*PTPDC1*	non-syn	C	T	A	V	62	0	0.89	0.9992	0.1736	no	no	yes	no
10	120810812	*EIF3A*	non-syn	G	A	R	C	740	0	0.73	0.9994	0.2848	yes	no	no	no
10	123971211	*TACC2*	non-syn	C	T	T	M	502	0	1	0.998	2.9859	yes	yes	yes	no
10	128192969	*C10orf90*	non-syn	A	G	V	A	267	0.01	0.02	0.9365	2.6803	yes	no	no	no
11	8122493	*TUB*	non-syn	C	T	R	C	446	0	1	0.999	0.0723	no	no	yes	no
11	9530327	*ZNF143*	non-syn	C	T	R	W	437	0	1	0.922	2.14	yes	no	yes	no
12	430249	*KDM5A*	non-syn	C	T	R	Q	818	0.78	0.77	0.9992	0.5499	yes	no	no	yes
12	10783802	*STYK1*	non-syn	G	T	A	D	98	0.02	0	0.9047	1.7457	yes	no	no	no
12	56221845	*DNAJC14*	non-syn	G	A	R	C	200	0	1	0.9877	0.0371	no	yes	yes	no
12	91371921	*EPYC*	non-syn	C	T	G	D	95	0.12	0.84	0.9991	0.478	yes	no	yes	no
13	24864923	*SPATA13*	non-syn	A	G	D	G	369	0	0.78	0.9983	0.3041	yes	no	no	no
13	52667265	*NEK5*	non-syn	T	C	H	R	378	0.22	0.98	0.7847	0.6349	yes	no	no	no
13	98668017	*IPO5*	non-syn	A	C	E	D	843	0.02	0.99	0.9857	0.0206	yes	no	no	no
13	101881755	*NALCN*	non-syn	T	C	T	A	539	0.01	0.97	0.9975	0.0414	yes	yes	yes	no
13	102568880	*FGF14*	non-syn	A	C	L	R	39	0.01	0.56	0.9983	0.0501	no	no	yes	no
15	25584307	*UBE3A*	non-syn	C	T	A	T	846	0.01	0.62	0.9982	0.0128	yes	yes	no	no
15	28391389	*HERC2*	non-syn	G	A	R	W	3668	0	1	0.9995	0.1469	yes	no	no	no
15	41860451	*TYRO3*	non-syn	G	A	R	H	333	0.1	0.88	0.0693	0.7283	yes	yes	yes	no
15	42002919	*MGA*	non-syn	G	A	R	H	819	0			0.873	yes	no	no	no
15	62273608	*VPS13C*	non-syn	A	G	V	A	657	0	0.02	0.9991	1.101	yes	no	no	no
15	80743328	*ARNT2*	non-syn	C	T	R	C	47	0	0.99	0.9976	0.2047	yes	yes	yes	no
16	15811127	*MYH11*	non-syn	G	A	R	W	1792	0	1	0.9956	0.5946	yes	no	no	yes
16	27268842	*NSMCE1*	non-syn	C	T	R	H	17	0.05	0.72	0.9984	3.5112	no	yes	yes	no
16	50744711	*NOD2*	non-syn	G	C	V	L	297	0.14	0.17	0.9996	0.75	no	no	yes	no
16	67264651	*FHOD1*	non-syn	C	T	R	H	904	0.02	0.75	0.8367	0.2477	yes	no	no	no
16	80667028	*CDYL2*	non-syn	C	T	R	H	241	0	1	0.9983	0.0578	no	no	yes	no
17	47656505	*NXPH3*	non-syn	G	A	R	Q	201	1	0.02	0.9992	0.004	no	no	yes	no
17	48601125	*MYCBPAP*	non-syn	G	A	D	N	582	0.06	0.49	0.9929	6.6175	no	no	yes	no
17	61882450	*DDX42*	non-syn	A	G	E	G	214	0.02	0.83	0.9993	0.1547	no	no	yes	no

Chr	Pos	Gene	Effect	Ref nt	Observ. nt	Ref aa	Observ. aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	G-st	E-st	E-bs	COSMIC
18	48452143	*ME2*	non-syn	C	T	R	C	397	0	0.95	0.9381	0.4046	no	no	yes	no
19	15164741	*CASP14*	non-syn	G	T	K	N	125	0	1	0.1449	0.0447	no	no	yes	no
19	18375992	*KIAA1683*	non-syn	C	A	Q	H	786	0	1	0.0734	9.8172	yes	no	no	no
19	40362730	*FCGBP*	non-syn	C	A	D	Y	5114	0.01	0.97	0.9985	2.9936	yes	no	no	no
19	40408082	*FCGBP*	non-syn	G	T	L	M	1547		0.91	0.8254	2.9936	no	no	yes	no
19	47207446	*PRKD2*	non-syn	C	T	R	Q	133	0	1	0.9987	1.8043	yes	no	no	no
19	49713590	*TRPM4*	non-syn	C	T	R	C	941	0.02	1	0.9989	0.1508	yes	yes	yes	no
19	57176223	*ZNF835*	non-syn	C	T	C	Y	115	0			3.6035	yes	no	no	no
20	76969	*DEFB125*	non-syn	A	G	T	A	128	0.04	0.14	0.0011	0.1743	yes	no	no	no
20	33575653	*MYH7B*	non-syn	A	T	D	V	493	0	0.72	0.9967	1.7309	no	yes	yes	no
20	42159033	*L3MBTL1*	non-syn	C	T	T	M	299	0	1	0.999	0.1449	yes	no	no	no
21	15746149	*HSPA13*	non-syn	A	T	V	E	402	0	0.93	0.9989	0.0183	no	no	yes	no
21	35468198	*SLC5A3*	non-syn	C	A	S	Y	234	0.01	0	0.999	1.0678	no	no	yes	no
21	43708011	*ABCG1*	non-syn	C	T	A	V	329	0.03	0.14	0.9968	0.0142	yes	yes	yes	no
22	32853263	*BPIFC*	non-syn	C	A	R	S	37	0.01	0.98	0.2285	4.5579	no	no	yes	no
22	35943079	*RASD2*	non-syn	T	G	S	A	75	0	1	0.9972	0.017	yes	no	no	no
22	50987896	*KLHDC7B*	non-syn	T	G	L	R	434	0	1	0.9965	3.3646	no	yes	yes	no
X	47918103	*ZNF630*	non-syn	A	C	C	W	576	0	1	0.9819	0.2213	no	no	yes	no
X	84329375	*APOOL*	non-syn	G	T	K	N	232	0.31			0.0087	yes	no	no	no
X	99661761	*PCDH19*	non-syn	C	T	R	H	612	0.02	0.99	0.9989	0.0276	no	yes	yes	no
X	138880854	*ATP11C*	non-syn	A	C	N	K	256	0	1	0.9987	1.0978	no	no	yes	no
X	151908787	*CSAG1*	non-syn	C	A	P	H	9	0	0.79	0.8333	6.6042	yes	no	no	no
X	151935493	*MAGEA3*	non-syn	A	C	L	R	225	0	1	0.9731	0.6124	yes	no	no	no

Table S 12 Somatic SNVs in patient 1 (damaging or conserved, additional SNVs to Table S 10)

Somatic SNVs, which did not exist in the database dbSNP, were called with at least one variant caller, and resulted either in a nonsense amino acid change or were predicted as damaging to the protein function by Sift or PolyPhen2 or were at a conserved position or in a conserved gene. Genes listed in the databases ESP or ExAc are not shown. The following abbreviations were used: Chr = chromosome, Pos = position, Observ. = observed, Ref = reference, nt = nucleotide, aa = amino acid, G-st = called in WGS data with Samtools, E-st = called in WES data with Samtools, E-bs = called in WES data with diBayes / Bioscope, non-syn = non-synonymous.

Chr	Pos	Gene	Effect	Ref nt	Observ. nt	Ref aa	Observ. aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	G-st	E-st	E-bs	COSMIC
1	16343573	*HSPB7*	non-syn	T	C	E	G	110	0.00	0.64	0.9968	0.0272	no	no	yes	no
1	21926027	*RAP1GAP*	non-syn	A	G	V	A	643	1.00	0.47	0.9957	11.463	no	no	yes	no
1	22047604	*USP48*	non-syn	G	T	L	I	607	0.05	0.00	0.9996	0.0346	yes	no	no	no
1	22329501	*CELA3A*	non-syn	G	T	G	C	17	0.00	0.72	0.9975	106.867	no	no	yes	no
1	57185921	*C1orf168*	non-syn	T	G	S	R	686	0.07	0.56	0.9687	0.4269	no	yes	yes	no
1	91859934	*HFM1*	non-syn	T	G	L	F	70	0.00	0.16	0.9389	19.611	no	no	yes	no
1	109801069	*CELSR2*	non-syn	T	C	V	A	1109	0.61	0.45	0.9964	0.7822	no	no	yes	no
1	158151974	*CD1D*	non-syn	T	G	L	V	161	0.29	0.00	0.0001	0.0579	no	no	yes	no
1	175048706	*TNN*	non-syn	T	C	V	A	216	0.33	0.55	0.8329	46.155	no	no	yes	no
1	196367787	*KCNT2*	non-syn	A	C	I	M	400	0.00	0.68	0.9767	0.0143	no	no	yes	no
1	196887442	*CFHR4*	non-syn	T	A	F	Y	301	0.00	0.39	0.9871	56.184	no	no	yes	no
2	14774302	*FAM84A*	non-syn	A	C	S	R	67	0.05	0.00	0.9982	0.0000	no	no	yes	no
2	86333449	*PTCD3*	non-syn	T	A	L	M	27	0.00	0.00	0.0070	0.1793	no	no	yes	no
2	107041568	*RGPD3*	non-syn	A	C	I	S	952	0.02			13.425	no	no	yes	no
2	116599859	*DPP10*	non-syn	T	G	F	V	770	0.00	0.97	0.9985	19.456	no	no	yes	no
2	196891498	*DNAH7*	non-syn	A	G	V	A	218	0.01	0.81	0.9991	19.891	no	yes	no	no
2	215813869	*ABCA12*	non-syn	A	C	F	C	1968	0.00	1.00	0.9764	11.316	no	no	yes	no
2	225661686	*DOCK10*	non-syn	A	C	L	V	1608	0.00			0.5502	no	no	yes	no
3	48685833	*CELSR3*	non-syn	A	C	L	R	2280	0.00	0.79	0.9964	0.4052	no	no	yes	no
4	876583	*GAK*	non-syn	G	A	R	W	477	0.00	1.00	0.0843	0.5924	no	no	yes	no
4	13571747	*BOD1L*	non-syn	T	G	K	T	3015	0.00	0.99	0.9968	13.509	no	no	yes	no
5	141694306	*SPRY4*	non-syn	T	A	Q	L	123	0.64	0.95	0.9982	0.0349	no	no	yes	no
6	33137191	*COL11A2*	non-syn	G	C	P	R	1149	0.04	0.75	0.9989	0.7828	no	no	yes	no
6	117687277	*ROS1*	non-syn	T	C	Q	R	925	0.21	0.92	0.8620	15.183	no	no	yes	yes
7	106938626	*COG5*	non-syn	T	G	D	A	456	0.01	0.83	0.8933	0.9487	yes	no	no	no
7	141536919	*PRSS37*	non-syn	A	C	I	S	187	0.86	0.99	0.9993	41.145	no	no	yes	no
8	65509363	*CYP7B1*	non-syn	A	C	F	V	453	0.06	0.47	0.9767	0.3939	no	no	yes	no
9	125512506	*OR1L6*	non-syn	T	C	I	T	127	0.00	0.98	0.9969	380.865	no	no	yes	no
10	1263020	*ADARB2*	non-syn	T	C	H	R	518	0.02	1.00	0.9970	0.7311	no	no	yes	no
10	118645926	*KIAA1598*	non-syn	C	A	G	C	609	0.02			0.0199	no	no	yes	yes
10	135347299	*CYP2E1*	non-syn	A	G	I	V	289	0.03	0.00	0.0142	0.6131	yes	no	no	no
11	64663944	*ATG2A*	non-syn	T	C	Q	R	1806	0.10	0.99	0.9942	15.500	no	no	yes	no

Chr	Pos	Gene	Effect	Ref nt	Observ. nt	Ref aa	Observ. aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	G-st	E-st	E-bs	COSMIC
11	118518752	*PHLDB1*	non-syn	T	C	V	A	1158	0.00	0.57	0.9976	0.2071	no	no	yes	no
11	118986890	*C2CD2L*	non-syn	T	A	F	Y	683	0.03	0.99	0.9980	0.2956	no	no	yes	no
12	40012871	*ABCD2*	non-syn	C	T	A	T	183	0.01	0.74	0.9980	0.0799	yes	no	no	no
14	25043464	*CTSG*	non-syn	T	G	K	T	194	0.00	0.98	0.9676	15.316	no	yes	yes	no
14	37132597	*PAX9*	non-syn	C	T	T	M	167	0.02	0.46	0.9984	11.336	yes	no	no	no
16	150473	*NPRL3*	non-syn	T	C	S	G	43	0.00			0.1839	no	no	yes	no
16	31090892	*ZNF646*	non-syn	G	A	V	I	1083	0.35	0.05	0.9997	18.997	no	no	yes	no
16	67709818	*GFOD2*	non-syn	A	G	M	T	133	0.05	0.82	0.9981	0.0107	no	yes	yes	no
16	68023265	*DPEP2*	non-syn	T	G	K	T	344	0.03	0.00	0.9742	44.821	no	yes	yes	no
16	84224888	*ADAD2*	non-syn	C	G	R	G	18	0.00	0.99	0.9674	42.752	no	no	yes	no
16	88713569	*CYBA*	non-syn	A	T	F	Y	48	0.04	0.87	0.9957	162.433	no	no	yes	no
17	14205307	*HS3ST3B1*	non-syn	C	T	R	C	158	0.00	0.74	0.9978	0.2017	yes	no	no	no
17	36499576	*GPR179*	non-syn	G	A	R	C	33	0.00	1.00	0.9909	0.6524	no	no	yes	no
17	36895869	*PCGF2*	non-syn	A	G	V	A	60	0.00	0.99	0.9652	0.0237	no	no	yes	no
17	37374365	*STAC2*	non-syn	T	C	N	S	51	0.00	0.99	0.9977	0.0257	no	yes	yes	no
17	38508279	*RARA*	non-syn	A	C	Q	P	99	0.00	0.37	0.9966	0.0206	no	no	yes	yes
17	48277129	*COL1A1*	non-syn	A	G	C	R	95	0.00	0.79	0.9983	10.697	no	no	yes	yes
18	3879864	*DLGAP1*	stoploss	T	C	X	G	69		0.98	0.9981	0.6117	no	no	yes	no
18	9887707	*TXNDC2*	non-syn	A	G	T	A	411		0.79	0.2339	79.892	no	no	yes	no
19	9578567	*ZNF560*	non-syn	T	A	E	D	352	0.13	0.87	0.7827	0.2806	no	no	yes	no
19	11568960	*ELAVL3*	non-syn	G	A	T	M	210	0.37	0.00	0.9836	0.0012	no	no	yes	no
19	36342206	*NPHS1*	non-syn	C	A	G	W	119	0.00	0.94	0.9432	12.396	no	no	yes	no
19	39871319	*SAMD4B*	non-syn	G	C	R	P	581	0.00	1.00	0.9992	0.0188	no	no	yes	no
19	55241094	*KIR3DL3*	non-syn	T	C	L	P	264	0.02	0.31	0.0186	16.339	no	no	yes	no
22	40800438	*SGSM3*	non-syn	C	G	I	M	115	0.00	0.09	0.8535	0.1747	no	no	yes	no
X	12736568	*FRMPD4*	non-syn	T	C	F	S	1208	0.00	0.07	0.9979	0.0212	no	no	yes	no

Table S 13 Somatic SNVs in patient 2 (damaging or conserved, additional SNVs to Table S 11)
Somatic SNVs, which did not exist in the database dbSNP, were called with at least one variant caller, and resulted either in a nonsense amino acid change or were predicted as damaging by Sift or PolyPhen2 or were at a conserved position or in a conserved gene. Genes listed in the databases ESP or ExAc are not shown. The following abbreviations were used: Chr = chromosome, Pos = position, Ref = reference, Observ. = observed, nt = nucleotide, aa = amino acid, G-st = called in WGS data with Samtools, E-st = called in WES data with Samtools, E-bs = called in WES data with diBayes / Bioscope, non-syn = non-synonymous.

Chr	Start	End	Gene	Type	Ref nt	Observed nt	Ref aa	Pos aa	Called in WES?	Called in WGS?	COSMIC
1	1560788	1560788	*MIB2*	frameshift insertion	-	G	W	332	uncovered	yes	no
1	10725600	10725600	*CASZ1*	frameshift insertion	-	G	P	15	uncovered	yes	no
1	27088787	27088787	*ARID1A*	frameshift insertion	-	G	Q	799	no	yes	yes
1	39833833	39833833	*MACF1*	frameshift deletion	A	-	E	2702	no	yes	no
1	40366751	40366751	*MYCL1*	frameshift insertion	-	G	R	119	uncovered	yes	no
1	47904569	47904569	*FOXD2*	frameshift insertion	-	G	G	254	uncovered	yes	no
1	50884762	50884762	*DMRTA2*	frameshift insertion	-	C	L	402	uncovered	yes	no
1	59156071	59156071	*MYSM1*	frameshift deletion	T	-	K	79	yes	yes	no
1	92643415	92643415	*KIAA1107*	frameshift deletion	A	-	K	363	uncovered	yes	no
1	156255317	156255317	*TMEM79*	frameshift insertion	-	C	E	100	yes	yes	no
1	214814551	214814554	*CENPF*	frameshift deletion	AAAT	-		957	yes	yes	no
1	225528370	225528370	*DNAH14*	frameshift insertion	-	A	E	3456	uncovered	yes	no
1	235345419	235345419	*ARID4B*	frameshift deletion	T	-	T	853	yes	yes	no
1	249141827	249141827	*ZNF672*	frameshift insertion	-	C	R	118	uncovered	yes	no
2	17698737	17698737	*RAD51AP2*	frameshift deletion	T	-	T	316	no	yes	no
2	24300578	24300580	*TP53I3*	non-frameshift deletion	GGA	-		224	yes	yes	no
2	73520690	73520690	*EGR4*	frameshift deletion	C	-	G	22	uncovered	yes	no
2	97824365	97824365	*ANKRD36*	frameshift insertion	-	T	N	454	uncovered	yes	no
2	109421454	109421454	*CCDC138*	frameshift insertion	-	A	L	282	no	yes	no
2	152320541	152320541	*RIF1*	frameshift deletion	A	-	K	1503	yes	yes	no
2	166011130	166011130	*SCN3A*	frameshift deletion	A	-	F	404	yes	yes	no
2	178257585	178257585	*AGPS*	frameshift insertion	-	G	A	23	uncovered	yes	no
2	186662103	186662103	*FSIP2*	frameshift deletion	A	-	K	3503	uncovered	yes	no
2	217559252	217559252	*IGFBP5*	frameshift deletion	G	-	R	83	uncovered	yes	no
2	219757717	219757717	*WNT10A*	frameshift insertion	-	C	G	326	uncovered	yes	no
2	231738163	231738163	*ITM2C*	frameshift deletion	G	-	R	51	no	yes	no
2	242626190	242626190	*DTYMK*	frameshift insertion	-	C	A	3	uncovered	yes	no
3	45267441	45267441	*TMEM158*	frameshift insertion	-	G	G	27	uncovered	yes	no
3	50647874	50647875	*CISH*	frameshift deletion	TG	-		13324	uncovered	yes	no
3	126194513	126194513	*ZXDC*	frameshift insertion	-	G	A	66	uncovered	yes	no

Chr	Start	End	Gene	Type	Ref nt	Observed nt	Ref aa	Pos aa	Called in WES?	Called in WGS?	COSMIC
3	184039592	184039592	*EIF4G1*	frameshift deletion	C	-	A	211	yes	yes	no
3	194991598	194991598	*C3orf21*	frameshift insertion	-	G	A	64	uncovered	yes	no
4	15005749	15005749	*CPEB2*	frameshift insertion	-	G	G	484	uncovered	yes	no
4	25278792	25278792	*PI4K2B*	frameshift deletion	T	-	F	477	yes	yes	no
4	42153875	42153875	*BEND4*	frameshift insertion	-	G	A	96	uncovered	yes	no
4	71508472	71508472	*ENAM*	frameshift deletion	A	-	P	443	no	yes	no
4	109748336	109748336	*COL25A1*	frameshift deletion	T	-	K	573	no	yes	no
4	142643137	142643137	*IL15*	frameshift insertion	-	A	L	57	yes	yes	no
4	164466808	164466808	*03.01.2017*	frameshift deletion	A	-	W	154	yes	yes	no
5	60628709	60628709	*ZSWIM6*	frameshift insertion	-	G	R	204	uncovered	yes	no
5	73981222	73981222	*HEXB*	frameshift insertion	-	G	P	46	uncovered	yes	no
5	145883505	145883506	*TCERG1*	frameshift deletion	GA	-		868	yes	yes	no
5	169309816	169309816	*FAM196B*	frameshift insertion	-	G	T	363	uncovered	yes	no
5	169677853	169677853	*LCP2*	frameshift deletion	T	-	T	454	yes	yes	no
6	22570251	22570251	*HDGFL1*	frameshift deletion	G	-	A	149	uncovered	yes	no
6	25491971	25491971	*LRRC16A*	frameshift deletion	T	-	N	359	uncovered	yes	no
6	30122095	30122096	*TRIM10*	frameshift deletion	GT	-		366	yes	yes	no
6	31868562	31868562	*ZBTB12*	frameshift deletion	G	-	P	174	no	yes	no
6	42073643	42073643	*C6orf132*	frameshift insertion	-	G	P	669	uncovered	yes	no
6	138751899	138751899	*NHSL1*	frameshift insertion	-	G	I	1199	uncovered	yes	no
6	150263278	150263278	*ULBP2*	frameshift insertion	-	G	R	24	uncovered	yes	no
6	166721385	166721385	*PRR18*	frameshift insertion	-	C	S	82	uncovered	yes	no
7	2394615	2394615	*EIF3B*	frameshift insertion	-	C	G	20	uncovered	yes	no
7	6662529	6662529	*ZNF853*	frameshift insertion	-	G	P	636	uncovered	yes	no
7	27170236	27170236	*HOXA4*	frameshift insertion	-	C	G	39	uncovered	yes	no
7	44268438	44268438	*CAMK2B*	frameshift insertion	-	G	P	475	uncovered	yes	no
7	138916512	138916512	*UBN2*	frameshift insertion	-	C	E	94	uncovered	yes	no
7	140482927	140482927	*BRAF*	frameshift deletion	G	-	P	403	yes	no	yes
7	141170468	141170468	*LOC100507421*	frameshift deletion	G	-	W	256	uncovered	yes	no
8	11566247	11566249	*GATA4*	non-frameshift deletion	CCT	-		142	uncovered	yes	no

Chr	Start	End	Gene	Type	Ref nt	Observed nt	Ref aa	Pos aa	Called in WES?	Called in WGS?	COSMIC
8	38645217	38645217	*TACC1*	frameshift insertion	-	G	P	39	uncovered	yes	no
8	54793598	54793598	*RGS20*	frameshift insertion	-	G	E	8	uncovered	yes	no
8	68950480	68950480	*PREX2*	frameshift insertion	-	T	V	264	yes	no	yes
8	145535689	145535689	*HSF1*	frameshift deletion	C	-	P	301	uncovered	yes	no
9	13217182	13217182	*MPDZ*	frameshift insertion	-	T	L	400	uncovered	yes	no
9	125330322	125330324	*OR1L8*	non-frameshift deletion	CAG	-		145	yes	yes	no
9	136342177	136342177	*SLC2A6*	frameshift insertion	-	C	L	148	uncovered	yes	no
9	139259595	139259595	*CARD9*	frameshift deletion	G	-	H	478	uncovered	yes	no
9	139564768	139564768	*EGFL7*	frameshift deletion	C	-	A	186	uncovered	yes	no
10	21101841	21101841	*NEBL*	frameshift insertion	-	T	T	129	yes	no	no
10	23728536	23728536	*OTUD1*	frameshift insertion	-	G	T	50	uncovered	yes	no
10	88423518	88423518	*OPN4*	frameshift deletion	C	-	P	453	uncovered	yes	no
10	99400629	99400629	*PI4K2A*	frameshift insertion	-	C	S	44	uncovered	yes	no
10	101295273	101295273	*NKX2-3*	frameshift insertion	-	G	A	297	uncovered	yes	no
10	102746683	102746685	*MRPL43*	non-frameshift deletion	CTC	-		96	uncovered	yes	no
10	102987157	102987157	*LBX1*	frameshift deletion	G	-	P	239	uncovered	yes	no
10	103990420	103990420	*PITX3*	frameshift insertion	-	G	Y	254	uncovered	yes	no
10	104182716	104182716	*FBXL15*	frameshift insertion	-	G	A	290	uncovered	yes	no
10	121411254	121411254	*BAG3*	frameshift deletion	C	-	P	23	uncovered	yes	no
10	127462683	127462683	*MMP21*	frameshift deletion	G	-	P	138	uncovered	yes	no
11	2181203	2181203	*INS*	frameshift deletion	C	-	G	71	uncovered	yes	no
11	33721959	33721961	*C11orf91*	non-frameshift deletion	AGA	-		109	uncovered	yes	no
11	47202199	47202199	*PACSIN3*	frameshift deletion	A	-	F	85	uncovered	yes	no
11	58346929	58346929	*ZFP91*	frameshift insertion	-	C	A	59	uncovered	yes	no
11	65403681	65403681	*PCNXL3*	frameshift insertion	-	G	G	1832	uncovered	yes	no
11	66307267	66307267	*ZDHHC24*	frameshift insertion	-	AAGG	L	196	uncovered	yes	no
11	67262953	67262953	*PITPNM1*	frameshift deletion	G	-	P	812	uncovered	yes	no
11	71951160	71951160	*PHOX2A*	frameshift insertion	-	C	A	163	uncovered	yes	no
11	75379136	75379136	*MAP6*	frameshift insertion	-	C	R	93	uncovered	yes	no
11	87013374	87013374	*TMEM135*	frameshift deletion	T	-	S	174	yes	yes	no

Chr	Start	End	Gene	Type	Ref nt	Observed nt	Ref aa	Pos aa	Called in WES?	Called in WGS?	COSMIC
11	92087750	92087750	*FAT3*	frameshift insertion	-	A	A	824	yes	yes	no
11	93428779	93428779	*KIAA1731*	frameshift deletion	A	-	R	550	uncovered	yes	no
11	108256684	108256684	*C11orf65*	frameshift insertion	-	T	K	250	no	yes	no
11	122848460	122848460	*BSX*	frameshift insertion	-	C	A	200	uncovered	yes	no
12	19592890	19592890	*AEBP2*	frameshift insertion	-	C	S	86	uncovered	yes	no
12	48577960	48577960	*C12orf68*	frameshift deletion	C	-	P	19	uncovered	yes	no
12	50190327	50190327	*NCKAP5L*	frameshift deletion	G	-	P	439	uncovered	yes	no
12	54332886	54332886	*HOXC13*	frameshift insertion	-	C	A	66	uncovered	yes	yes
12	56827355	56827355	*TIMELESS*	frameshift insertion	-	A	L	111	yes	no	no
12	58149402	58149402	*MARCH6*	non-frameshift insertion	-	CGG	R	31	uncovered	yes	no
12	62954511	62954511	*MON2*	frameshift deletion	A	-	E	1217	yes	yes	no
12	65563984	65563984	*LEMD3*	frameshift insertion	-	G	A	203	uncovered	yes	no
12	80175596	80175596	*PPP1R12A*	frameshift insertion	-	T	R	898	uncovered	yes	no
12	106532389	106532389	*NUAK1*	frameshift insertion	-	G	L	15	uncovered	yes	no
12	110941673	110941673	*RAD9B*	frameshift deletion	A	-	K	37	no	yes	no
13	32885712	32885712	*ZAR1L*	frameshift deletion	G	-	P	117	uncovered	yes	no
13	110435612	110435612	*IRS2*	frameshift insertion	-	G	R	930	uncovered	yes	no
13	112722529	112722529	*SOX1*	frameshift insertion	-	C	G	186	uncovered	yes	no
14	61190643	61190643	*SIX4*	frameshift insertion	-	G	P	50	uncovered	yes	no
14	75230626	75230626	*YLPM1*	frameshift deletion	C	-	S	145	uncovered	yes	no
14	77744769	77744769	*POMT2*	frameshift insertion	-	C	G	705	uncovered	yes	no
14	101005508	101005508	*BEGAIN*	frameshift insertion	-	G	S	194	uncovered	yes	no
14	104638177	104638179	*KIF26A*	non-frameshift deletion	GGT	-		411	uncovered	yes	no
15	34444996	34444996	*C15orf29*	frameshift deletion	A	-	S	145	yes	yes	no
15	41687234	41687234	*NDUFAF1*	frameshift deletion	A	-	F	194	yes	yes	no
15	42174159	42174159	*SPTBN5*	frameshift insertion	-	C	A	776	uncovered	yes	no
15	63414257	63414257	*LACTB*	frameshift insertion	-	C	A	63	uncovered	yes	no
15	68119568	68119568	*SKOR1*	frameshift insertion	-	C	A	424	uncovered	yes	no
15	72612182	72612182	*CELF6*	frameshift insertion	-	G	A	12	uncovered	yes	no
15	73615063	73615063	*HCN4*	frameshift deletion	C	-	G	1124	uncovered	yes	no

Chr	Start	End	Gene	Type	Ref nt	Observed nt	Ref aa	Pos aa	Called in WES?	Called in WGS?	COSMIC
15	73615394	73615394	*HCN4*	frameshift deletion	C	-	A	1014	uncovered	yes	no
15	74726122	74726122	*SEMA7A*	frameshift insertion	-	C	G	46	uncovered	yes	no
15	75942736	75942736	*SNX33*	frameshift deletion	C	-	D	431	no	yes	no
16	1245562	1245562	*CACNA1H*	frameshift insertion	-	G	A	181	uncovered	yes	no
16	1401980	1401980	*GNPTG*	frameshift insertion	-	G	L	5	uncovered	yes	no
16	20810178	20810178	*ERI2*	frameshift deletion	T	-	N	315	uncovered	yes	no
16	20844363	20844363	*LOC81691*	frameshift deletion	T	-	L	434	yes	yes	no
16	27506189	27506189	*GTF3C1*	frameshift insertion	-	G	P	891	yes	no	no
16	30021389	30021389	*DOC2A*	frameshift insertion	-	CC	A	52	uncovered	yes	no
16	50187728	50187728	*PAPD5*	frameshift insertion	-	C	T	51	uncovered	yes	no
16	67695546	67695546	*PARD6A*	frameshift insertion	-	C	G	84	uncovered	yes	no
16	72993830	72993830	*ZFHX3*	frameshift insertion	-	C	S	72	uncovered	yes	yes
16	75682105	75682105	*TERF2IP*	frameshift insertion	-	C	T	109	uncovered	yes	no
16	87448918	87448918	*ZCCHC14*	frameshift insertion	-	C	A	343	yes	yes	no
16	88495435	88495435	*ZNF469*	frameshift insertion	-	G	G	519	uncovered	yes	no
16	89349641	89349641	*ANKRD11*	frameshift deletion	T	-	K	1103	yes	yes	no
16	89753118	89753120	*CDK10*	frameshift deletion	CAT	-		10	uncovered	yes	no
17	2298441	2298441	*MNT*	frameshift deletion	G	-	P	127	uncovered	yes	no
17	4693224	4693224	*GLTPD2*	frameshift deletion	C	-	A	170	uncovered	yes	no
17	10600660	10600660	*SCO1*	frameshift insertion	-	G	S	55	uncovered	yes	no
17	18087645	18087645	*ALKBH5*	frameshift insertion	-	C	A	30	uncovered	yes	no
17	20108263	20108263	*SPECC1*	frameshift deletion	A	-	K	220	no	yes	yes
17	46985900	46985900	*UBE2Z*	frameshift insertion	-	G	A	12	uncovered	yes	no
17	46986017	46986018	*UBE2Z*	frameshift deletion	CA	-		51	uncovered	yes	no
17	47297238	47297238	*ABI3*	frameshift insertion	-	G	P	183	uncovered	yes	no
17	48624635	48624635	*SPATA20*	frameshift insertion	-	G	A	22	uncovered	yes	no
17	76968105	76968105	*LGALS3BP*	frameshift deletion	C	-	G	437	yes	yes	no
17	79425361	79425361	*BAHCC1*	frameshift insertion	-	C	A	1683	uncovered	yes	no
17	79425400	79425400	*BAHCC1*	frameshift insertion	-	C	A	1696	uncovered	yes	no
19	577814	577814	*BSG*	frameshift deletion	G	-	V	36	uncovered	yes	no

Chr	Start	End	Gene	Type	Ref nt	Observed nt	Ref aa	Pos aa	Called in WES?	Called in WGS?	COSMIC
19	1465537	1465537	*APC2*	frameshift insertion	-	C	G	746	uncovered	yes	no
19	2251182	2251182	*AMH*	frameshift deletion	C	-	V	303	uncovered	yes	no
19	4175089	4175089	*SIRT6*	frameshift insertion	-	G	L	198	uncovered	yes	no
19	4543990	4543990	*SEMA6B*	frameshift insertion	-	G	A	764	uncovered	yes	no
19	10121036	10121036	*COL5A3*	frameshift insertion	-	G	Q	9	uncovered	yes	no
19	14201203	14201203	*SAMD1*	frameshift insertion	-	G	P	10	uncovered	yes	no
19	19006845	19006845	*CERS1*	frameshift insertion	-	C	P	13	uncovered	yes	no
19	19654762	19654762	*CILP2*	frameshift deletion	C	-	P	470	no	yes	no
19	36120066	36120066	*RBM42*	frameshift insertion	-	G	A	4	uncovered	yes	no
19	36223970	36223970	*MLL4*	frameshift insertion	-	T	V	2174	uncovered	yes	no
19	36359544	36359544	*APLP1*	frameshift insertion	-	C	G	2	uncovered	yes	no
19	39226910	39226910	*CAPN12*	frameshift insertion	-	C	L	475	uncovered	yes	no
19	41699266	41699266	*CYP2S1*	frameshift deletion	C	-	P	33	uncovered	yes	no
19	44118005	44118006	*SRRM5*	frameshift deletion	GA	-		578	uncovered	yes	no
19	45899436	45899436	*PPP1R13L*	frameshift deletion	C	-	G	298	uncovered	yes	no
19	48949367	48949367	*GRWD1*	frameshift insertion	-	G	P	35	uncovered	yes	no
19	50161631	50161631	*SCAF1*	frameshift insertion	-	G	P	1305	uncovered	yes	no
19	54659494	54659494	*LENG1*	frameshift insertion	-	C	R	254	uncovered	yes	no
19	54974667	54974667	*LENG9*	frameshift insertion	-	G	A	37	uncovered	yes	no
19	55998075	55998075	*NAT14*	frameshift insertion	-	C	V	125	uncovered	yes	no
19	56011445	56011445	*SSC5D*	frameshift deletion	C	-	S	656	uncovered	yes	no
20	278509	278509	*ZCCHC3*	frameshift insertion	-	G	R	94	uncovered	yes	no
20	5987071	5987071	*CRLS1*	frameshift insertion	-	G	L	60	uncovered	yes	no
20	21695239	21695239	*PAX1*	frameshift insertion	-	G	Q	468	uncovered	yes	no
20	35064745	35064745	*DLGAP4*	frameshift deletion	C	-	N	411	uncovered	yes	no
20	35414897	35414897	*KIAA0889*	frameshift deletion	G	-	P	1659	uncovered	yes	no
20	42143414	42143414	*L3MBTL1*	frameshift insertion	-	G	A	77	uncovered	yes	no
20	57429638	57429640	*GNAS*	non-frameshift deletion	GAT	-		377	uncovered	yes	yes
20	58514841	58514841	*PPP1R3D*	frameshift insertion	-	G	R	49	uncovered	yes	no
20	62200526	62200528	*PRIC285*	non-frameshift deletion	CCT	-		354	uncovered	yes	no

Chr	Start	End	Gene	Type	Ref nt	Observed nt	Ref aa	Pos aa	Called in WES?	Called in WGS?	COSMIC
21	34443272	34443272	*OLIG1*	frameshift insertion	-	G	G	240	uncovered	yes	no
21	46897723	46897723	*COL18A1*	frameshift deletion	C	-	G	535	uncovered	yes	no
22	17601541	17601541	*CECR6*	frameshift insertion	-	C	G	159	uncovered	yes	no
22	17601574	17601574	*CECR6*	frameshift insertion	-	C	G	148	uncovered	yes	no
22	19511720	19511720	*CLDN5*	frameshift deletion	C	-	G	105	uncovered	yes	no
22	22221701	22221701	*MAPK1*	frameshift insertion	-	C	G	10	uncovered	yes	yes
22	26902289	26902289	*TFIP11*	frameshift deletion	C	-	G	167	yes	yes	no
22	40058306	40058306	*CACNA1I*	frameshift insertion	-	C	A	1045	uncovered	yes	no
22	51133261	51133261	*SHANK3*	frameshift insertion	-	T	G	363	uncovered	yes	no
X	16965034	16965034	*REPS2*	frameshift insertion	-	GG	A	17	uncovered	yes	no
X	48814518	48814518	*OTUD5*	frameshift insertion	-	C	G	105	uncovered	yes	no
X	119048674	119048674	*AKAP14*	frameshift deletion	A	-	K	92	yes	no	no
X	128782641	128782641	*APLN*	frameshift insertion	-	C	R	66	uncovered	yes	no

Table S 14 Exonic, somatic, filtered small InDels in the MSI tumor sample of the first patient
The following abbreviations were used in the header: Chr = chromosome, Pos = position, Ref = reference, nt = nucleotide, aa = amino acid.

Chr	Start	End	Gene	Type	Ref nt	Observed nt	Ref aa	Pos aa	Called in exome?	Called in genome?	COSMIC
1	38274248	38274248	*C1orf122*	frameshift insertion	-	G	G	27	uncovered	yes	no
2	9533687	9533687	*ASAP2*	frameshift insertion	-	C	K	820	uncovered	yes	no
2	42275911	42275911	*PKDCC*	frameshift insertion	-	G	L	191	uncovered	yes	no
2	101869561	101869561	*C2orf29*	frameshift insertion	-	G	S	45	uncovered	yes	no
2	102003876	102003876	*CREG2*	frameshift insertion	-	G	R	15	uncovered	yes	no
2	175201193	175201193	*SP9*	frameshift insertion	-	G	A	127	uncovered	yes	no
2	176957800	176957801	*HOXD13*	frameshift deletion	CA	-		61	uncovered	yes	yes
2	230578923	230578923	*DNER*	frameshift insertion	-	G	A	73	uncovered	yes	no
3	57199225	57199227	*IL17RD*	non-frameshift deletion	GGA	-		30	uncovered	yes	no
3	118621756	118621756	*IGSF11*	frameshift deletion	T	-	I	303	no	yes	no
3	147128069	147128069	*ZIC1*	frameshift insertion	-	T	A	57	uncovered	yes	no
6	42893110	42893111	*PTCRA*	frameshift deletion	TG	-		179	uncovered	yes	no

Chr	Start	End	Gene	Type	Ref nt	Observed nt	Ref aa	Pos aa	Called in exome?	Called in genome?	COSMIC
6	110679218	110679218	*C6orf186*	frameshift deletion	G	-	G	86	uncovered	yes	no
6	166721600	166721600	*PRR18*	frameshift deletion	C	-	A	11	uncovered	yes	no
7	65447139	65447139	*GUSB*	frameshift deletion	G	-	A	11	uncovered	yes	no
7	155532530	155532531	*RBM33*	frameshift deletion	AG	-		620	uncovered	yes	no
8	25224393	25224398	*DOCK5*	non-frameshift deletion	ACAATT	-	na	1044	yes	yes	no
8	37756930	37756930	*RAB11FIP1*	frameshift insertion	-	C	G	10	uncovered	yes	no
8	145735004	145735004	*MFSD3*	frameshift insertion	-	C	G	96	uncovered	yes	no
10	48438611	48438611	*GDF10*	frameshift insertion	-	G	S	34	uncovered	yes	no
11	1092475	1092475	*MUC2*	frameshift insertion	-	AT	P	1432	uncovered	yes	no
11	64565113	64565113	*MAP4K2*	frameshift deletion	T	-	E	374	uncovered	yes	no
15	23891258	23891258	*MAGEL2*	frameshift insertion	-	G	T	544	uncovered	yes	no
15	68119471	68119471	*SKOR1*	frameshift insertion	-	G	G	391	uncovered	yes	no
16	29818333	29818333	*MAZ*	frameshift insertion	-	G	P	76	uncovered	yes	no
17	636375	636375	*FAM57A*	frameshift insertion	-	C	S	54	uncovered	yes	no
17	4462251	4462251	*GGT6*	frameshift insertion	-	G	A	149	uncovered	yes	no
19	40317538	40317538	*DYRK1B*	frameshift insertion	-	C	G	395	uncovered	yes	no
20	25062560	25062560	*VSX1*	frameshift insertion	-	C	P	58	uncovered	yes	no
20	57429638	57429640	*GNAS*	non-frameshift deletion	GAT	-		377	uncovered	yes	yes
20	62373884	62373884	*SLC2A4RG*	frameshift insertion	-	C	R	292	uncovered	yes	no
X	11682758	11682758	*ARHGAP6*	frameshift insertion	-	G	R	64	uncovered	yes	no

Table S 15 Exonic, somatic, filtered small InDels in the MSS tumor sample of the second patient
The following abbreviations were used in the header: Chr = chromosome, Pos = position, Ref = reference, nt = nucleotide, aa = amino acid.

6.2.1.4 Exonic gene conservation score

The ECS is a simplified indicator for clinically relevant exonic small variants. The score compares the non-synonymous mutation rate within the exonic regions of the gene of interest with the average non-synonymous mutation rate over all exonic regions of the genome. The ECS was calculated for the exonic regions of each gene using the coding variants of 1092 samples from the 1000 Genomes Project. To ensure comparability between genes, the score was normalized by the exon length. On average 10,755.21 non-synonymous SNVs affecting 5,912.16 genes were detected in each of the 1000 Genomes Project data sets. Out of these variants, 2,816.90 SNVs in 2,171.07 genes were predicted to be damaging to protein function. Additionally, on average 509.23 small InDels existed in 418.45 genes including 399.70 frameshift InDels in 338.59 genes. 18,209 genes were affected in at least one sample by one or more coding variations.

Each gene was on average affected in 285 samples by at least one exonic non-synonymous variant. Out of these samples, 101 harbored at least one SNV with damaging prediction and 16 samples one or more frameshift InDels (Figure S 11 A).

Across all genes the average ECS was 1.52 (Figure S 11 B), while the average ECS of genes out of the cancer gene census list from the COSMIC database [122] was 0.66. The score was applied as a simplified filter to interpret the variants from the two investigated clinical samples. An extremely low ECS was found for example in the genes: *PCGF2* (0.023), *ZIC1* (0.0), *VHLL* (0.044), *INHBA* (0.016), *MTA2* (0.011), *BRAF* (0.076), *GNB2L1* (0.015), *JAK2* (0.081), *CDC25B* (0.0459), and *EXT2* (0.099).

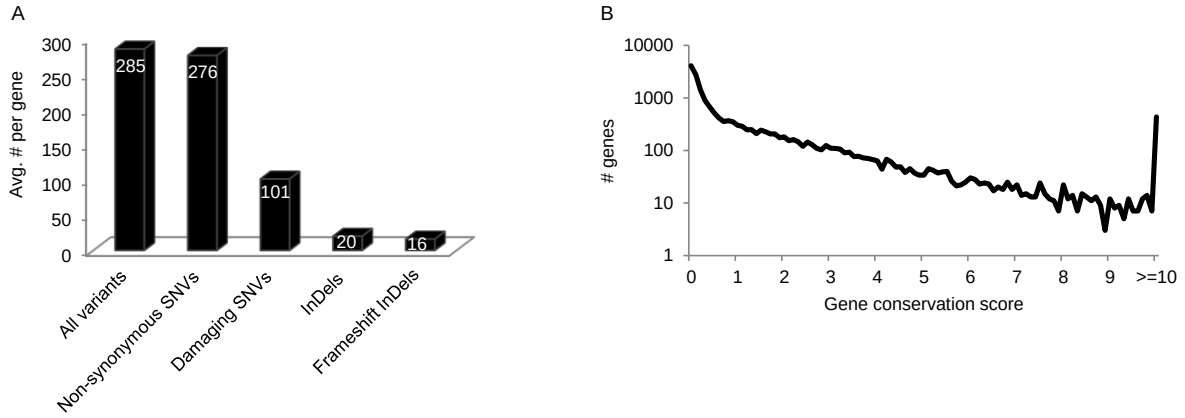


Figure S 11 Coding variant distribution in samples from the 1000 Genomes Project
(A) The barplot illustrates the average number of samples affected by the variant type listed on the y-axis. The damaging SNVs were included in the non-synonymous SNVs, which in turn were part of the total number of variants. The InDel count contained also the frameshift InDels and were part of the total number of variants. In some samples, a gene might contain an SNV as well as an InDel. In that case the gene counted for both categories. (B) In this figure, the distribution of the ECS values is displayed.

To demonstrate the relevance of the ECS for cancer development, the obtained scores were compared to (i) the associated cPI values and (ii) the number of samples harboring a mutation with predicted functional relevance in cancer (Figure S 12) for each gene in the COSMIC data

set. The first analysis demonstrated that a low ECS was associated with a negative cPI (potential tumor suppressor genes) or a positive cPI value (potential oncogenes), while genes with a neutral associated cPI value harbored a high ECS score (Figure S 12 A+B). This was also confirmed by a one-tailed Wilcoxon rank sum test comparing genes having an absolute cPI value of at least 0.4 with genes having an absolute cPI value smaller 0.4 (p = 0.0008). The cPI values separated by tumor suppressor genes and oncogenes were 0.2545 and 8.698E-05, respectively. However, the non-significant p-value is probably a consequence of too low power. In a second test, it was show that genes with low ECS harbored more often mutations predicted as cancer-driver mutations by FATHMM (p = 1.75E-07). The results are depicted in Figure S 12 C+D.

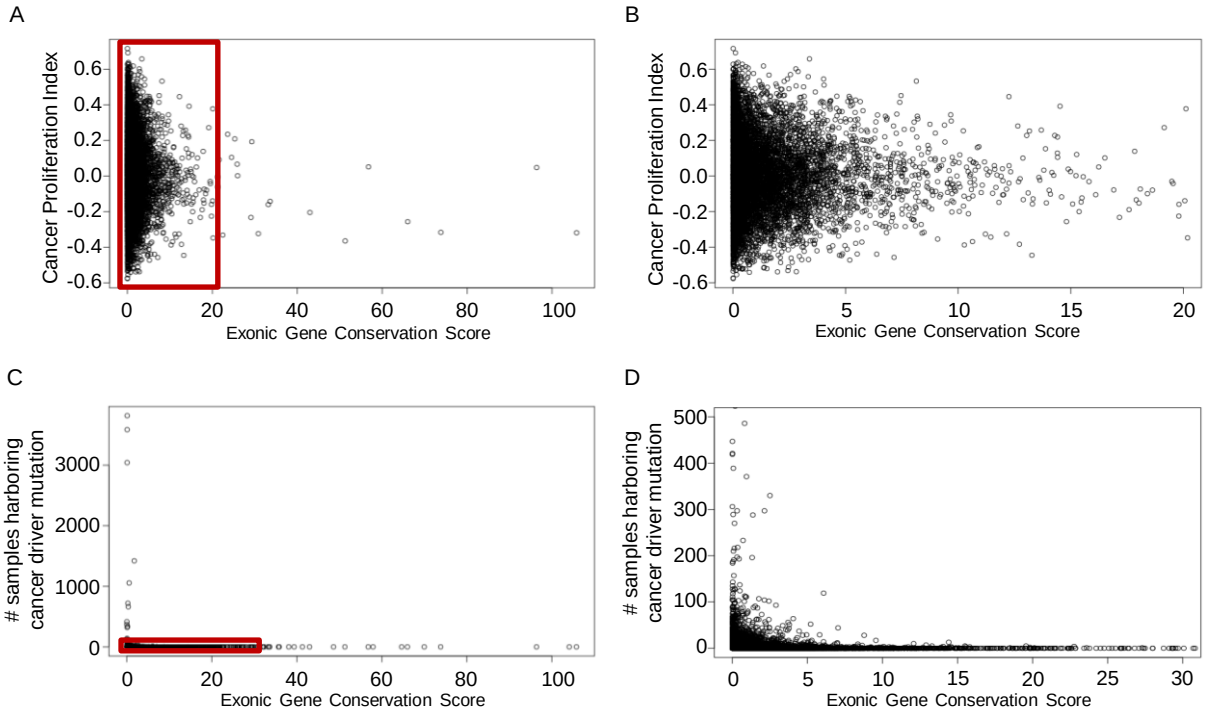


Figure S 12 Comparison of ECS with other cancer-relevant parameters
Relation between the novel developed ECS and known cancer associations: (A) Scatter plot comparing the ECS with the cPI, whereby tumor suppressors exhibit negative cPI values and oncogenes positive cPI values. (B) Enlargement of the small value region of figure part (A). (C) Scatter plot comparing the ECS with the number of samples having a mutation with an FATHMM cancer prediction in the COSMIC database. (D) Enlargement of the small value region of figure part (C).

6.2.1.5 High and low quality SNVs in GC samples

Chr	Pos	Gene	Effect	Ref nt	Observed nt	Ref aa	Observed aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	1000G freq	G-st	E-st	E-bs
1	8424848	*RERE*	nonsense	G	A	Q	X	500	0.17	0.74	0.9996	0.1114	0.1114	yes	yes	yes
5	31423048	*DROSHA*	nonsense	G	A	Q	X	1052	0.01			0.6416	0.6416	yes	yes	yes
7	150647363	*KCNH2*	non-syn	G	A	P	L	424	0.00	0.95	0.9987	0.3248	0.3248	yes	yes	yes
19	49713590	*TRPM4*	non-syn	C	T	R	C	941	0.02	1.00	0.9989	0.1508	0.1508	yes	yes	yes
21	43708011	*ABCG1*	non-syn	C	T	A	V	329	0.03	0.14	0.9968	0.0142	0.0142	yes	yes	yes

Table S 16 High quality somatic SNVs in the MSI tumor sample of the first patient
Somatic SNVs were called with Samtools and diBayes / Bioscope in the WES and with Samtools in the WGS data. None of the SNVs were annotated in the database dbSNP and all SNVs were supported by at least 20% of the reads with a high quality base at this position in both data set. Additionally, the program PolyPhen2 or the program Sift predicted a change of the protein function for all variants. The following abbreviations were used: Chr = chromosome, Pos = position, Ref = reference allele, nt = nucleotide, aa = amino acid, G-st = called in WGS data with Samtools, E-st = called in WES data with Samtools, E-bs = called in WES data with diBayes / Bioscope, 1000G freq = allele frequency in the 1000 Genomes project, non-syn = non-synonymous

Chr	Pos	Gene	Effect	Ref nt	Observed nt	Ref aa	Observed aa	Pos aa	Sift	PolyPhen	PhyloP	ECS	1000G freq	G-st	E-st	E-bs
14	100363521	*EML1*	nonsense	C	G	Y	X	239	0.00	0.74	0.9984	1.7618		yes	yes	yes

Table S 17 High quality somatic SNVs in the MSS tumor sample of the second patient
Somatic SNVs were called with Samtools and diBayes / Bioscope in the WES and with Samtools in the WGS data. None of the SNVs were annotated in the database dbSNP and all SNVs were supported by at least 20% of the reads with a high quality base at this position in both data set. Additionally, the program PolyPhen2 or the program Sift predicted a change of the protein function for all variants. The following abbreviations were used in the header: Chr = chromosome, Pos = position, Ref = reference allele, nt = nucleotide, aa = amino acid, G-st = called in WGS data with Samtools, E-st = called in WES data with Samtools, E-bs = called in WES data with diBayes / Bioscope.

			Allele counts WES Patient 1 tumor (MSI)						Allele counts WES Patient 1 non-tumor						Allele counts WGS Patient 1 tumor (MSI)						Allele counts WGS Patient 1 non-tumor					
Chr	**Position**	**Gene**	**Cov**	**#A**	**#C**	**#G**	**#T**	**%allele**	**Cov**	**#A**	**#C**	**#G**	**#T**	**%allele**	**Cov**	**#A**	**#C**	**#G**	**#T**	**%allele**	**Cov**	**#A**	**#C**	**#G**	**#T**	**%allele**
2	222428603	*EPHA4*	66	0	0	56	10	15.15	94	0	0	82	12	12.77	63	0	0	63	0	0	82	1	0	81	0	1.22
4	38800365	*TLR1*	41	8	33	0	0	19.51	33	5	28	0	0	15.15	56	0	56	0	0	0	107	1	106	0	0	0.93
7	120614861	*ING3*	28	0	22	6	0	21.43	30	2	26	2	0	6.67	67	0	67	0	0	0	115	0	115	0	0	0
9	36966598	*PAX5*	15	4	11	0	0	26.67	22	1	21	0	0	4.55	39	0	39	0	0	0	57	0	57	0	0	0
11	47296590	*MADD*	43	0	1	9	33	20.93	42	0	0	1	41	2.38	42	0	0	0	42	0	53	0	2	0	51	3.77
12	54369186	*HOXC11*	12	0	8	4	0	33.33	21	0	19	2	0	9.52	36	0	36	0	0	0	84	0	84	0	0	0
14	70990319	*ADAM20*	49	6	43	0	0	12.24	47	3	43	0	1	6.38	47	0	47	0	0	0	85	0	85	0	0	0
17	56439928	*RNF43*	13	9	0	4	0	30.77	16	15	0	1	0	6.25	46	43	2	1	0	4.35	54	45	4	5	0	9.26

Table S 18 Allele calls of failed validations (MSI tumor)
All SNVs with failed validations using Pyromark sequencing are listed in the table. All variants were called in the MSI WES data, but not in the WGS data. The table shows the mutant allele fractions based on all sequencing data sets. Chr = chromosome, Cov = coverage.

			Allele counts WES Patient 2 tumor (MSS)						Allele counts WES Patient 2 non-tumor						Allele counts WGS Patient 2 tumor (MSS)						Allele counts WGS Patient 2 non-tumor					
Chr	**Position**	**Gene**	**Cov**	**#A**	**#C**	**#G**	**#T**	**%allele**	**Cov**	**#A**	**#C**	**#G**	**#T**	**%allele**	**Cov**	**#A**	**#C**	**#G**	**#T**	**%allele**	**Cov**	**#A**	**#C**	**#G**	**#T**	**%allele**
2	222428603	*EPHA4*	98	0	0	82	16	16.33	70	0	1	62	7	10	29	0	0	29	0	0	43	1	0	42	0	2.33
4	38800365	*TLR1*	29	4	25	0	0	13.79	32	2	30	0	0	6.25	28	0	28	0	0	0	57	0	56	0	1	1.75
7	120614861	*ING3*	25	0	16	9	0	36	31	0	28	3	0	9.68	29	0	29	0	0	0	44	0	42	1	1	2.27
9	36966598	*PAX5*	19	5	14	0	0	26.32	11	0	11	0	0	0	12	0	12	0	0	0	37	0	37	0	0	0
12	54369186	*HOXC11*	19	0	16	3	0	15.79	23	0	22	1	0	4.35	28	1	27	0	0	3.57	24	0	24	0	0	0
14	70990319	*ADAM20*	52	11	39	0	2	21.15	54	5	46	1	2	9.26	41	0	41	0	0	0	55	0	55	0	0	0
17	56439928	*RNF43*	26	19	0	7	0	26.92	14	14	0	0	0	0	16	14	0	2	0	12.5	37	31	4	2	0	10.81

Table S 19 Allele calls of failed validations (MSS tumor)
All SNVs with failed validations using Pyromark sequencing are listed in the table. All variants were called in the MSI WES data, but not in the WGS data. The table shows the mutant allele fractions based on all sequencing data sets. Chr = chromosome, Cov = coverage.

6.2.1.6 Structural variants

Chromosome	Range start	Range end	Insert length	Closest gene	Distance to closest gene
chr2	92103056	92103061	6	*ACTR3BP2*	26098
chr7	57688147	57688147	20	*L37717*	10499
chr7	57760881	57760881	5	*L37717*	62077
chr7	57786314	57786314	11	*L37717*	87510
chr7	141743389	141743389	18	*MGAM*	intronic
chr9	40407198	40407198	11	*CNTNAP3B*	91283
chr9	41897881	41897886	10	*BC019880* / *GLIDR*	50628
chr9	42084498	42084498	7	*LOC554249*	64085
chr9	44460316	44460321	9	*LOC103908605*	57056
chr9	45607139	45607139	7	*LOC100132167*	43576
chr9	46900844	46900849	10	*LOC103908605*	55685
chr16	33109610	33109613	31	*uc021thc.2* / abParts	intronic
chr20	26255314	26255315	6	*LOC284801*	65445

Table S 20 Putative somatic insertion positions (MSI tumor)
Exclusively positions with adjacent gene closer than 100 kb are reported. It was not possible to decide the exact insertion position for all contigs. Thus, several insertion positions might be mentioned for the same contig.

Chromosome	Start	End	Affected genes
chr1	142555082	142563172	
chr1	144597345	147731206	
chr4	71878175	71955584	*DCK* (part, 5 exons)
chr5	143406614	143409417	
chr5	147553039	147554779	*SPINK14* (complete, intronic/exonic (1 exon)
chr6	131884839	131884921	
chr6	148123470	148123785	
chr7	70420969	70438886	
chr7	82946477	82947144	
chr7	129066570	129091085	*AHCYL2* (part, 1 exon), *STRIP2* / *FAM40B* (part, 3 exons)
chr9	201453	68434529	*LOC642236* (part, 4 exons) + 268 genes completely included in inversion
chr13	63605566	63605666	
chr14	91825622	91856306	*CCDC88C* (complete, intronic; part, exonic (1 exon) → depending on isoform)
chr16	33919757	33982747	
chr16	46415696	46429995	
chr17	36350387	36406170	*LOC440434* (part, 11 exons)
chr18	66398498	66399176	*CCDC102B* (complete, intronic)

Table S 21 Somatic inversions in the MSI tumor of the first patient

Chromosome	Start	End	Affected genes
chr1	33107241	33108379	*ZBTB8OS* (complete, intronic)
chr1	46026619	46056013	*AKR1A1* (part, 8 exons), *NASP* (part, 1 exon)
chr1	58581661	58592142	*DAB1* (complete, intronic)
chr1	78351482	78352083	
chr1	88336591	88354625	
chr1	90638539	90680287	
chr1	91648865	91648917	
chr1	119569249	119600947	*WARS2* (part, 4 exons)
chr1	121484951	121485030	
chr1	247966170	247972189	
chr3	29048876	29048961	
chr3	37553372	37553456	*ITGA9* (complete, intronic)
chr3	83003318	83018450	
chr3	102543384	102544570	
chr3	106918402	106919564	*LOC100302640* (complete, intronic)
chr4	66275067	66348582	*EPHA5* (complete, intronic /exonic (2 exons))
chr4	79031136	79035783	*FRAS1* (complete, intronic)
chr4	84104479	84105808	
chr4	156815407	156815478	
chr5	43213865	43213953	*NIM1* (complete, intronic)
chr5	114736695	114747608	
chr5	147553039	147554616	*SPINK14* (complete, intronic / exonic (1 exon))
chr6	57403534	57410135	*PRIM2* (complete, intronic)
chr6	73042386	73047307	*RIMS1* (complete, intronic/exonic(1 exon))
chr6	74193716	74193781	*MTO1* (complete, intronic)

Chromosome	Start	End	Affected genes
chr6	88577129	88577896	
chr6	97175866	97175945	
chr7	25050949	25051015	
chr7	25648669	25691137	
chr7	49185769	49185822	
chr7	67761987	67762040	
chr7	73531412	73534012	*LIMK1* (complete, intronic)
chr7	83184322	83191781	*SEMA3E* (complete, intronic)
chr7	97067522	97073554	
chr8	92808962	92809039	
chr9	21731902	21780626	
chr9	22288707	22295759	
chr9	84948003	84948066	
chr9	103507288	103527689	
chr9	105073669	105080146	
chr10	4073418	4073485	
chr10	78801102	78882773	*KCNMA1* (complete, intronic/exonic (7-9 exons))
chr10	87963306	87967672	*GRID1* (complete, intronic / exonic (1 exon))
chr10	92094490	92126417	
chr11	47505799	47505871	*CELF1* (complete, intronic)
chr11	48755851	48766522	
chr11	51271771	51277705	
chr11	55233143	55244176	
chr11	72702282	72702339	*FCHSD2* (complete, intronic)
chr11	75447499	75449820	
chr11	97390551	97504019	
chr12	22591827	22593955	
chr12	25327274	25327342	*CASC1* (complete, intronic)
chr12	25718424	25749850	*IFLTD1* (complete, intronic)
chr12	26100409	26102521	
chr12	30161961	30163876	
chr12	55149052	55149137	
chr12	66321678	66321734	*HMGA2* (complete, intronic)
chr12	71251513	71251579	*PTPRR* (complete, intronic)
chr12	78394308	78394758	*NAV3* (complete, intronic)
chr12	79984196	79984275	
chr13	28815776	28815852	*PAN3* (complete, intronic)
chr13	56178670	56186577	
chr13	58064798	58064892	
chr13	66014522	66032936	
chr13	69647410	69670752	
chr13	85871603	85876732	
chr14	19041247	19041828	
chr14	37771227	37771306	*MIPOL1* (complete, intronic)
chr14	41376039	41376752	
chr14	41665929	41669668	

Chromosome	Start	End	Affected genes
chr14	49818005	49826727	
chr14	87568801	87568879	
chr15	24591229	24596136	
chr15	88012976	88028424	
chr15	96601282	96601375	
chr16	46404113	46428659	
chr16	71355217	71355279	
chr16	85148340	85149026	
chr17	78727180	78727267	*RPTOR* (complete, intronic)
chr18	18517757	18519508	
chr18	32588783	32588873	*MAPRE2* (complete, intronic)
chr18	66572493	66578708	*CCDC102B* (complete, intronic)
chr21	27252354	27252444	
chr21	38652658	38684076	
chrX	41896340	41897093	
chrX	55861181	55872362	
chrX	127890752	127890831	
chrX	133192855	133217454	

Table S 22 Somatic inversions in the MSS tumor of the second patient

Region	First breakpoint	Second breakpoint	Affected genes
1	chr1:38432684	chr6:28863596	*SF3A3*
1	chr1:38432879	chr6:28863596	*SF3A3*
2	chr1:90541646	chr2:214698198	
3	chr1:168025482	chr19:24033171	*DCAF6*
4	chr2:114173377	chr9:68434306	
5	chr2:133012697	chr12:20704282	*ANKRD30BL*
6	chr3:87409970	chr7:78953928	
6	chr3:87410199	chr7:78953928	
7	chr3:196625679	chr10:42596810	*SENP5*
7	chr3:196625679	chr10:42394405	*SENP5*
8	chr4:63927764	chr6:101513828	
8	chr4:63927927	chr6:101513828	
9	chr4:68264961	chr10:42409187	
9	chr4:68264961	chr10:42400514	
10	chr5:26793705	chr16:20578017	
10	chr5:26793705	chr16:20577768	
11	chr5:58988375	chr7:80801401	*PDE4D*
11	chr5:58988375	chr7:80802030	*PDE4D*
12	chr5:68499750	chrX:53055087	*CENPH*
12	chr5:68499750	chrX:53054134	*CENPH*
12	chr5:68500027	chrX:53054134	*CENPH*
13	chr6:24684231	chr22:32928263	*ACOT13*
14	chr6:58779261	chr11:51579574	
15	chr6:144120812	chr7:153001941	*PHACTR2*
16	chr7:81789072	chr10:60902712	*CACNA2D1*

Region	First breakpoint	Second breakpoint	Affected genes
16	chr7:81789232	chr10:60902404	*CACNA2D1*
17	chr7:87255610	chrX:97409801	*ABCB1*
17	chr7:87255610	chrX:97409606	*ABCB1*
18	chr10:68672793	chr16:59976458	*CTNNA3*
19	chr10:73519977	chr11:115199242	*CDH23,C10orf54*
20	chr12:23543029	chr14:88593567	
21	chr12:56989998	chr15:39994382	*BAZ2A*

Table S 23 Somatic interchromosomal translocations in the MSI tumor of the first patient

Region	First breakpoint	Second breakpoint	Affected genes
1	chr1:797220	chr8:245677	
2	chr1:17009765	chrX:20144024	
3	chr1:24999672	chr8:128100486	*SRRM1*
4	chr1:121215193	chr5:49771048	
5	chr1:143415010	chr4:49575973	
6	chr1:143486506	chr21:9987496	
6	chr1:143495641	chr21:9996427	
7	chr1:162753282	chr10:38637976	
8	chr3:612427	chr21:9539857	
8	chr3:624196	chr21:9527955	
8	chr3:630907	chr21:9521308	
8	chr3:636568	chr21:9515812	
8	chr3:656525	chr21:9495627	
9	chr3:75690142	chr4:190971339	
10	chr3:75989431	chr20:26208555	
10	chr3:75998372	chr20:26199800	
11	chr4:33859950	chr20:26128229	
12	chr4:190808236	chr9:68502759	
13	chr5:21474750	chr6:58272103	*GUSBP1*
14	chr5:79074210	chr12:25779850	*CMYA5*
14	chr5:79074210	chr12:25780263	*CMYA5*
15	chr5:112669294	chr6:40177372	*MCC*
16	chr6:3635462	chr9:21980977	
17	chr6:20112057	chr9:117857675	*MBOAT1*
18	chr7:35237763	chr12:64074427	
19	chr7:57624938	chr20:26129727	
19	chr7:57649520	chr20:26105996	
20	chr7:61867700	chr16:32493228	
21	chr7:111053279	chr12:108203016	*IMMP2L*
22	chr7:147472362	chr17:35072033	*CNTNAP2*
23	chr7:151957736	chr21:11045053	*MLL3*
23	chr7:151963767	chr21:11051127	*MLL3*
23	chr7:151963912	chr21:11051265	*MLL3*
23	chr7:151964511	chr21:11051789	*MLL3*
23	chr7:151972641	chr21:11059983	*MLL3*
23	chr7:151977173	chr21:11064581	*MLL3*

23	chr7:151981196	chr21:11068284	*MLL3*
23	chr7:151986563	chr21:11073827	*MLL3*
24	chr8:8638645	chr14:84865478	
25	chr8:15289716	chr13:74313860	
26	chr8:52730378	chr11:38812394	*PCMTD1*
27	chr9:37594147	chr10:123120352	
27	chr9:37594235	chr10:123120352	
28	chr9:68427074	chr20:29634618	
28	chr9:68427675	chr20:29633953	
28	chr9:68433406	chr20:29628176	*LOC642236*
29	chr9:79186892	chr12:127650468	
30	chr9:117857835	chr20:37093067	*TNC*
31	chr10:38974883	chr22:16950560	
32	chr12:23543032	chr14:88593566	
33	chr13:19481577	chr18:14644739	
34	chr14:19023273	chr21:14350483	
35	chr16:33975243	chr21:10702394	
35	chr16:33975351	chr21:10702582	
35	chr16:33984668	chr21:10711887	
36	chr16:46496441	chr21:9454722	
37	chr17:21548864	chr20:26079922	
47	chr17:21549513	chr20:26079255	
38	chr17:45264766	chr22:20062765	*CDC27*
39	chr19:19632268	chr22:16347238	*NDUFA13*
40	chr19:49810907	chrX:144519280	*SLC6A16*
41	chr20:11281494	chrX:81651230	
42	chr21:11183706	chr22:18886976	

Table S 24 Somatic interchromosomal translocations in the MSS tumor of the second patient

Chromosome	First deletion start	Last deletion end	Gene	# deletions	Mean deletion length
chr5	19847034	19867791	*CDH18*	2	23.00
chr5	21527908	21569501	*GUSBP1*	9	11.44
chr5	21855682	22715926	*CDH12*	4	10.50
chr5	40794615	40794979	*PRKAA1*	1	363.00
chr5	52172144	52232105	*ITGA1*	4	103.00
chr5	58277277	59541862	*PDE4D*	9	16.56
chr5	70853481	70853549	*BDP1*	1	67.00
chr5	75435663	75561971	*SV2C*	3	47.33
chr5	77004046	77062619	*TBCA*	2	160.50
chr5	78277726	78278046	*ARSB*	1	319.00
chr5	80039371	80039681	*MSH3*	1	309.00
chr5	80838449	81008438	*SSBP2*	2	140.50
chr5	88032012	88032343	*MEF2C*	1	330.00
chr5	89918620	90351220	*GPR98*	4	10.25
chr5	93757792	93916621	*KIAA0825*	2	174.00
chr5	101796603	101796650	*SLCO6A1*	1	46.00

Chromosome	First deletion start	Last deletion end	Gene	# deletions	Mean deletion length
chr5	110584044	110709132	*CAMK4*	4	85.50
chr5	115781960	115894458	*SEMA6A*	2	163.50
chr5	122701699	122727303	*CEP120*	3	25.67
chr5	136604692	136783317	*SPOCK1*	2	27.50
chr5	137022562	137023887	*KLHL3*	1	1324.00
chr5	147462155	147462472	*SPINK5*	1	316.00
chr5	149746191	149746518	*TCOF1*	1	326.00
chr5	152986428	152986521	*GRIA1*	1	92.00
chr5	154315917	154316003	*GEMIN5*	1	85.00
chr5	158592397	158592743	*RNF145*	1	345.00
chr5	170092722	170158907	*KCNIP1*	2	50.00
chr6	38498278	38498416	*BTBD9*	1	137.00
chr6	38720386	38932857	*DNAH8*	4	93.25
chr6	41953350	41995188	*CCND3*	2	162.50
chr6	45063196	45180261	*SUPT3H*	2	25.00
chr6	45927255	45967786	*CLIC5*	4	88.75
chr6	51736109	51904144	*PKHD1*	2	351.00
chr6	55337265	55408298	*HMGCLL1*	4	81.75
chr6	57207690	57489304	*PRIM2*	48	252.25
chr6	62410046	62835931	*KHDRBS2*	5	11.40
chr6	64794896	66290505	*EYS*	14	164.50
chr6	69648332	70043874	*BAI3*	5	19.40
chr6	74421253	74501003	*CD109*	3	107.00
chr6	76716172	76746584	*IMPG1*	2	50.50
chr6	89449444	89449487	*RNGTT*	1	42.00
chr6	91237575	91237897	*MAP3K7*	1	321.00
chr6	102030791	102357886	*GRIK2*	4	13.25
chr6	105567837	105568173	*BVES*	1	335.00
chr6	106966857	106986043	*AIM1*	4	86.50
chr6	108031342	108032403	*SCML4*	1	1060.00
chr6	116630476	116751133	*DSE*	2	163.50
chr6	123624457	123897619	*TRDN*	5	10.80
chr6	124234027	125119226	*NKAIN2*	6	34.17
chr6	134811194	134817663	*LOC154092*	2	168.50
chr6	137313693	137314187	*NHEG1*	1	493.00
chr6	143138010	143161521	*HIVEP2*	3	37.67
chr6	153029971	153033790	*MYCT1*	1	3818.00
chr6	154446915	154447238	*OPRM1*	1	322.00
chr6	157957840	157958148	*ZDHHC14*	1	307.00
chr6	158548266	158549067	*SERAC1*	1	800.00
chr6	161795638	163140940	*PARK2*	6	18.33
chr6	169905557	170038446	*WDR27*	2	210.00
chr7	69094157	70093358	*AUTS2*	4	25.75
chr7	70852212	71166844	*WBSCR17*	4	19.25
chr7	71398080	71828232	*CALN1*	7	13.43
chr7	77692552	78955527	*MAGI2*	15	44.00

Chromosome	First deletion start	Last deletion end	Gene	# deletions	Mean deletion length
chr7	81358337	81380822	*HGF*	2	141.50
chr7	81590035	81904509	*CACNA2D1*	4	13.50
chr7	83019539	83251598	*SEMA3E*	5	15.80
chr7	83653800	83724687	*SEMA3A*	4	110.75
chr7	86836614	86836924	*C7orf23*	1	309.00
chr7	93169987	93170105	*CALCR*	1	117.00
chr7	101521036	101911193	*CUX1*	3	16.33
chr7	102488704	102607726	*FBXL13*	3	117.67
chr7	103231015	103614941	*RELN*	6	62.83
chr7	104187384	104347403	*LHFPL3*	3	115.33
chr7	105733990	105734329	*SYPL1*	1	338.00
chr7	110536533	110543010	*IMMP2L*	2	256.50
chr7	113867592	114154152	*FOXP2*	5	14.80
chr7	120881096	120881435	*C7orf58*	1	338.00
chr7	121963925	122346024	*CADPS2*	7	52.86
chr7	126301958	126775701	*GRM8*	3	36.00
chr7	131813284	132279786	*PLXNA4*	4	13.25
chr7	134349999	134350423	*BPGM*	1	423.00
chr7	134475804	134490463	*CALD1*	2	22.00
chr7	136739075	136845672	*LOC349160*	3	105.33
chr7	144382749	144382898	*TPK1*	1	148.00
chr7	146174662	148076324	*CNTNAP2*	8	468.00
chr7	151397836	151516844	*PRKAG2*	4	11.00
chr7	152099668	152109219	*MLL3*	16	15.69
chr7	157870222	158266488	*PTPRN2*	7	17.29
chr7	158563013	158574681	*ESYT2*	2	87.00

Table S 25 Somatic, intragenic large deletions detected in the MSI tumor of the first patient

Chromosome	Deletion start	Deletion end	Gene	# deletions	Mean deletion length
chr8	9636709	9638499	*TNKS*	1	1789.00
chr11	61107785	61108125	*DAK*	1	339.00
chr12	23741655	24460697	*SOX5*	3	100.67
chr12	129088970	129089092	*TMEM132C*	1	121.00

Table S 26 Somatic, intragenic large deletions detected in the MSS tumor of the second patient

6.2.1.7 Pathway analyses

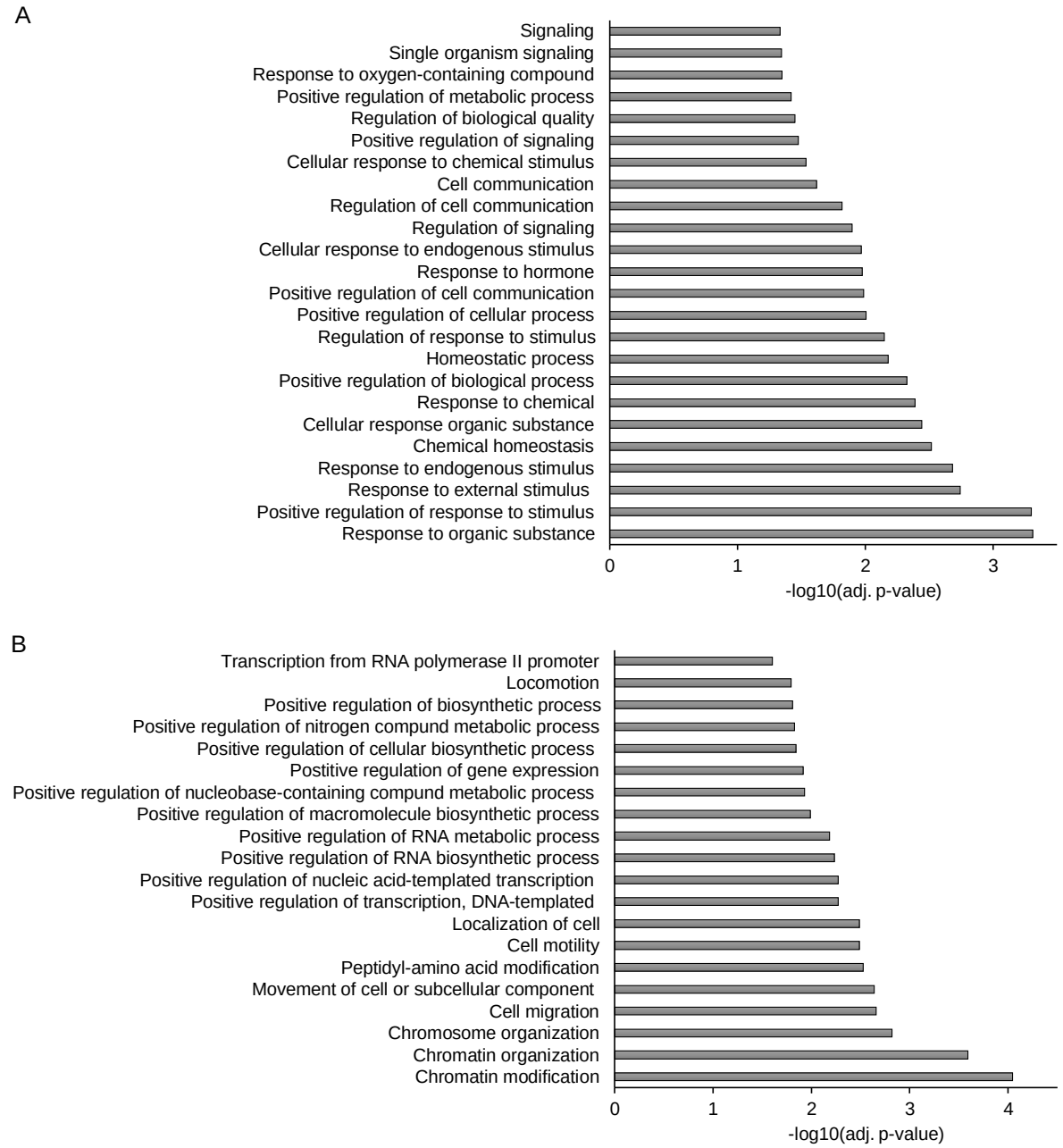


Figure S 13 Enriched processes of the two subnetworks identified in the protein-protein interaction network based on genes affected by at least one SNV with predicted damaging effect on protein function in the first patient (MSI tumor)
(A) Larger subnetwork. (B) Smaller subnetwork.

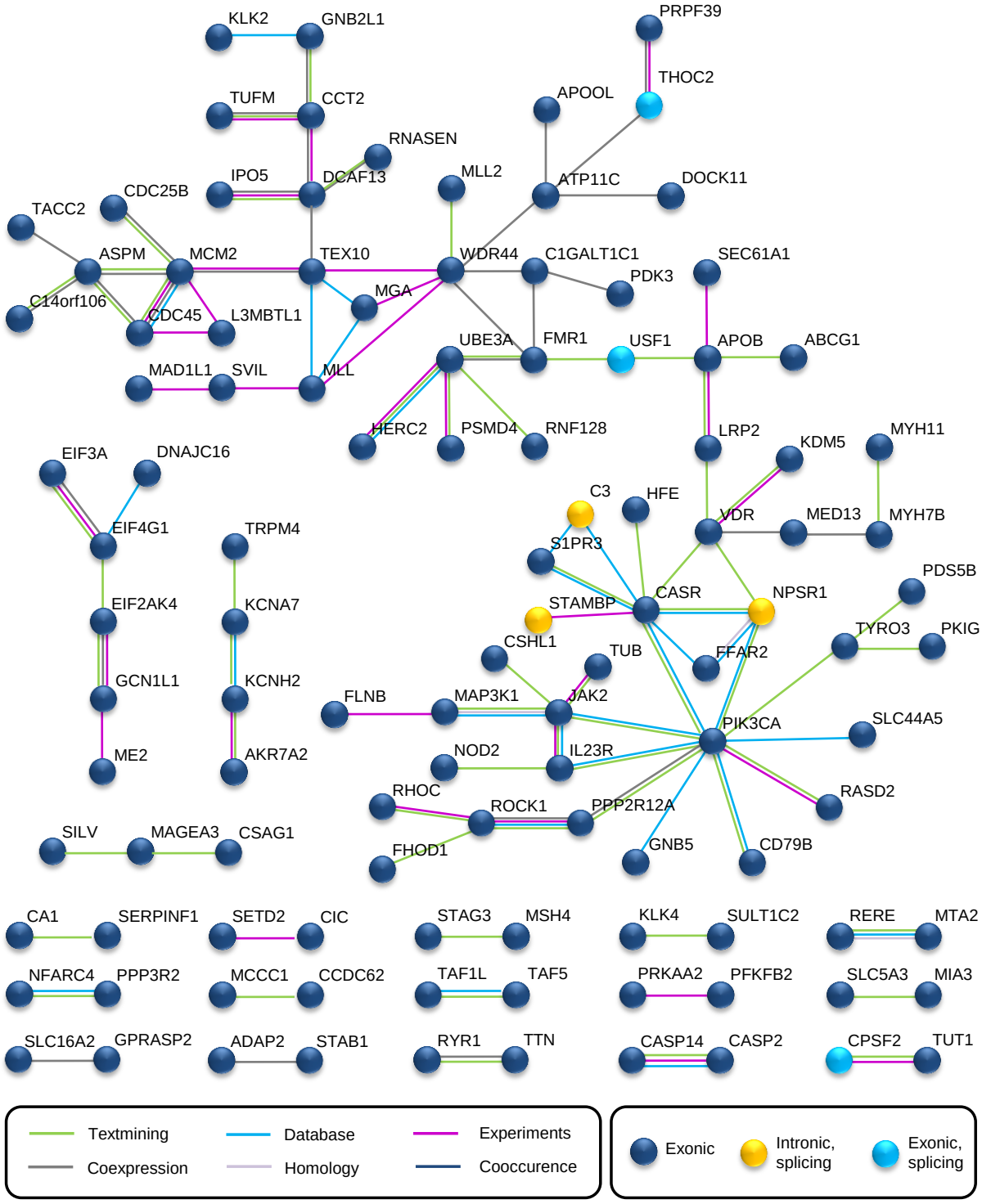


Figure S 14 Protein-protein interaction network for the MSI tumor based on all non-synonymous SNVs
The links were based on information from the String database. The line color indicates the source of the connection. Proteins with an SNV in the exonic region of the corresponding gene were marked in dark blue. If the SNV was in the exonic splice site part, the node is colored bright blue and if the variant was in the intronic splice site part, the node is yellow. Candidates without a connection to the network or any other candidate were excluded from the figure.

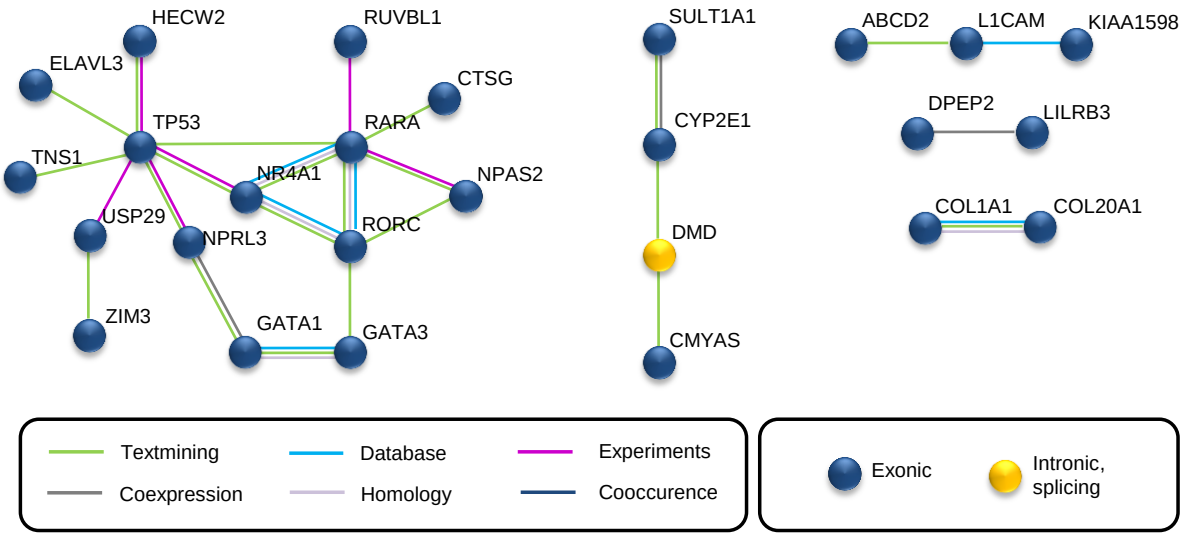


Figure S 15 Protein-protein interaction network for the MSS tumor based on all non-synonymous SNVs
The links were based on information from the String database. The line color indicates the source of the connection. Proteins with an SNV in the exonic region of the corresponding gene were marked in dark blue. If the SNV was in the exonic splice site part, the node is colored bright blue and if the variant was in the intronic splice site part, the node is yellow. Candidates without a connection to the network or any other candidate were excluded from the figure.

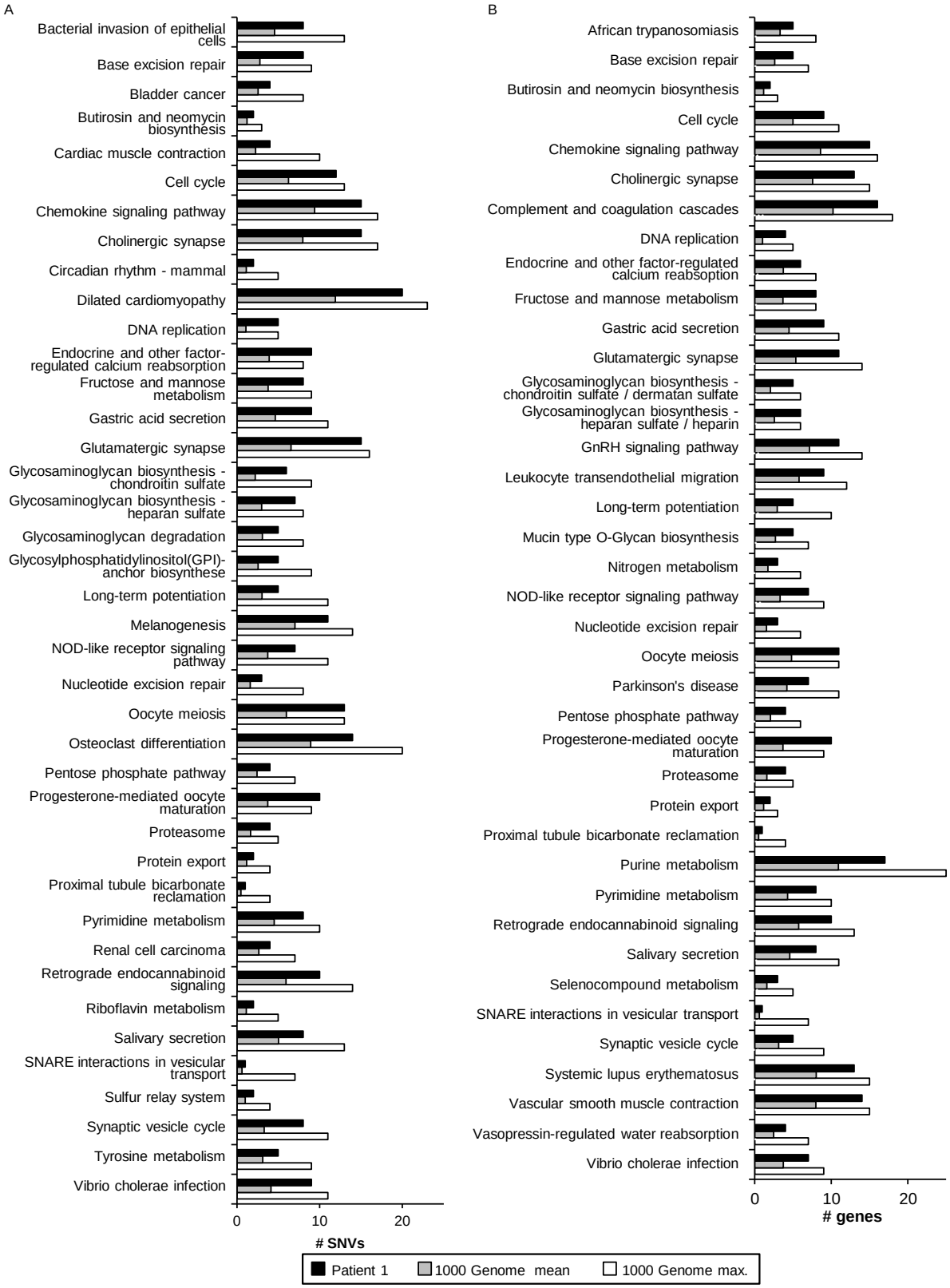


Figure S 16 Enriched KEGG pathways in patient 1 (MSI tumor)
All Pathways, which were at least 1.5 times more often affected in the tumor sample than the average value of all individuals in the 1000 Genomes Project, were included in the figure. (A) Analysis based on number of genes affected by at least one SNV with predicted damaging effect on protein function. (B) Analysis based on number of SNVs with predicted damaging effect on protein function.

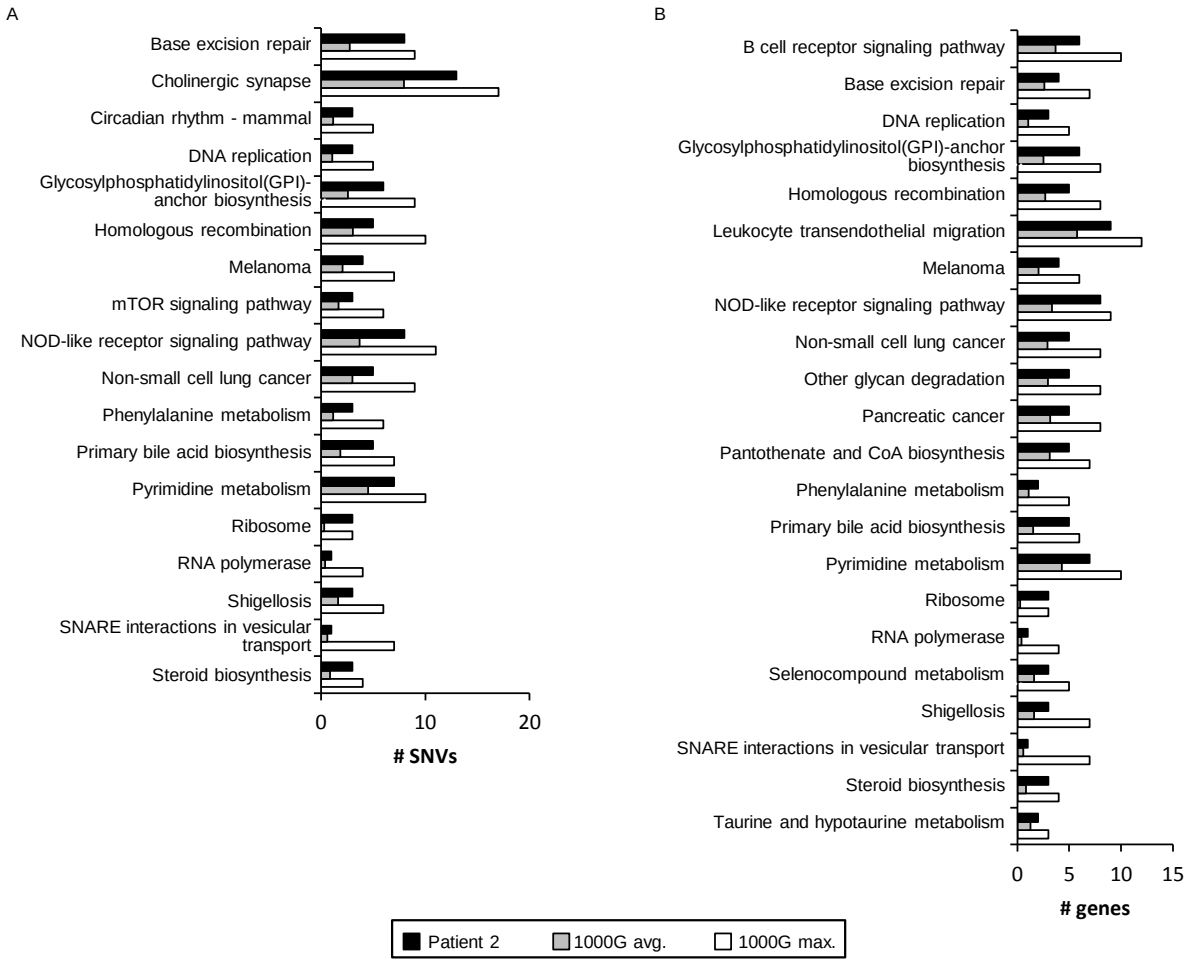


Figure S 17 Enriched KEGG pathways in patient 2 (MSS tumor)
All Pathways, which were at least 1.5 times more often affected in the tumor sample than the average value of all individuals of the 1000 Genomes Project, were included in the figure. (A) Analysis based on number of genes affected by at least one SNV with predicted damaging effect on protein function. (B) Analysis based on number of SNVs with predicted damaging effect on protein function.

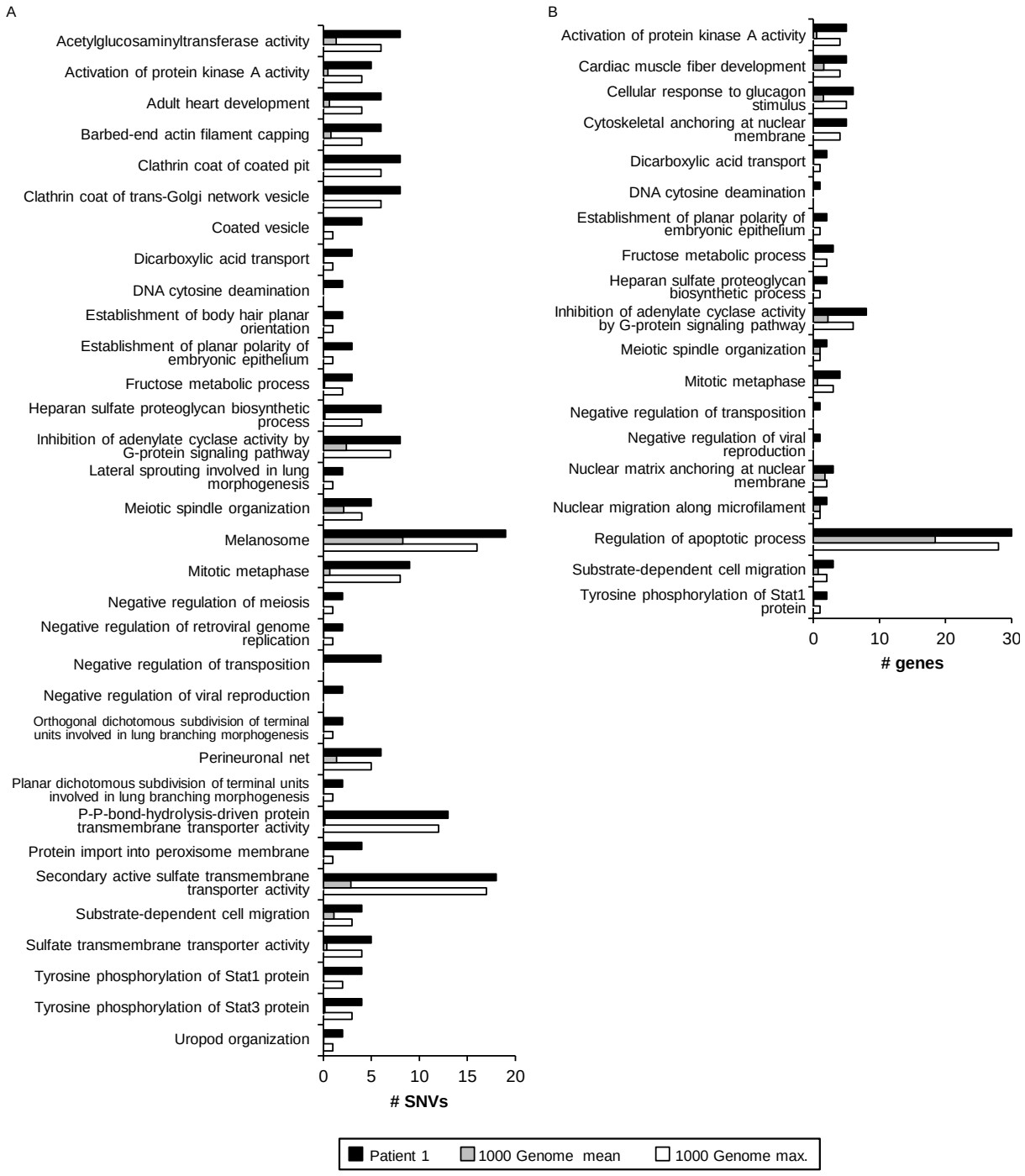


Figure S 18 Enriched GO terms in patient 1 (MSI tumor)
GO terms, which were more often affected in the tumor sample than in all individuals of the 1000 Genomes Project, were included in the figure. (A) Analysis based on number of genes affected by at least one SNV with predicted damaging effect on protein function. (B) Analysis based on number of SNVs with predicted damaging effect on protein function.

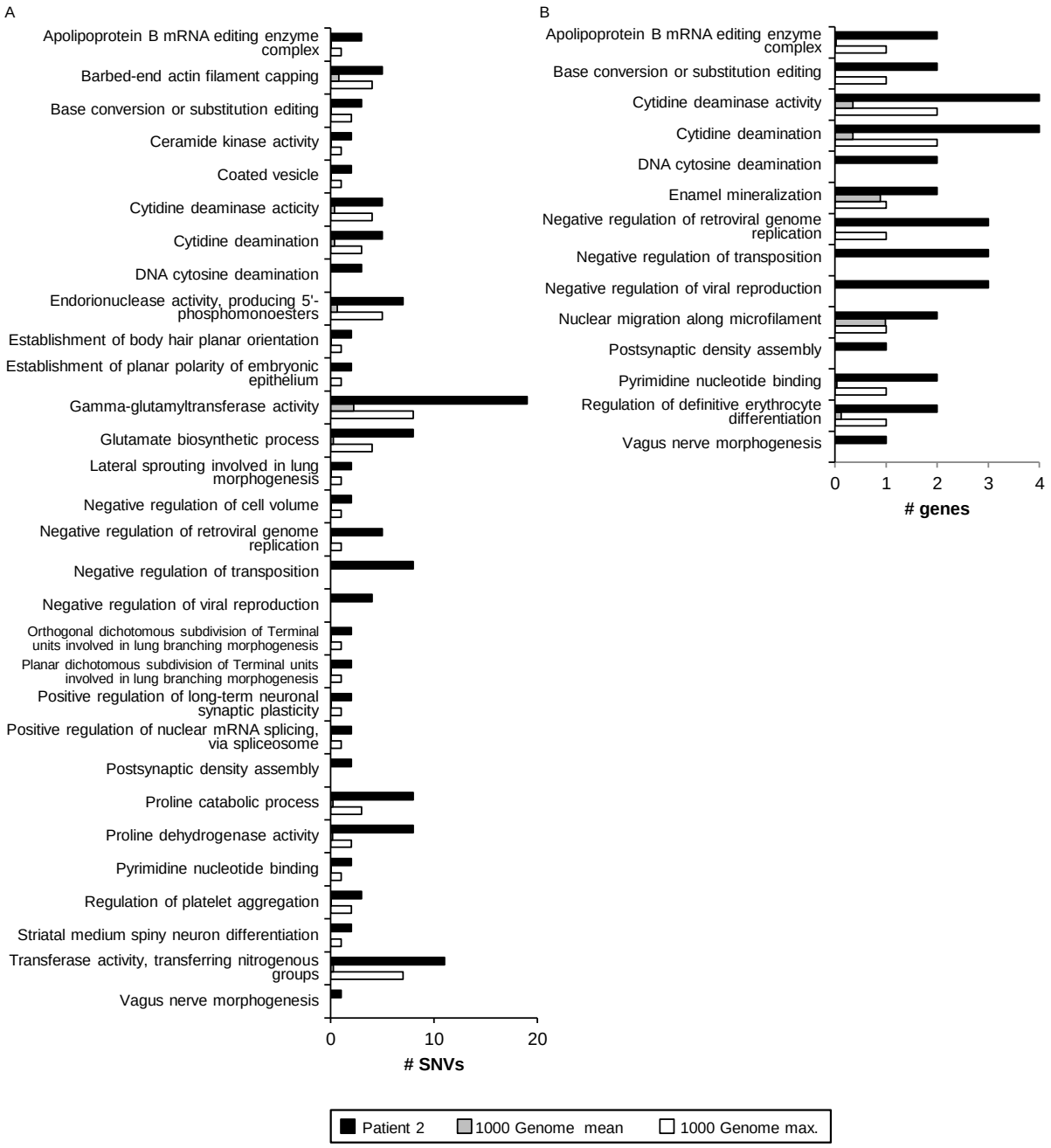


Figure S 19 Enriched GO terms in patient 2 (MSS tumor)
GO terms, which were more often affected in the tumor sample than in all individuals of the 1000 Genomes Project, were included in the figure. (A) Analysis based on number of genes affected by at least one SNV with predicted damaging effect on protein function. (B) Analysis based on number of SNVs with predicted damaging effect on protein function.

6.2.2 Murine AOM/DSS-triggered colorectal cancer

6.2.2.1 Phenotype caused by AOM/DSS treatment

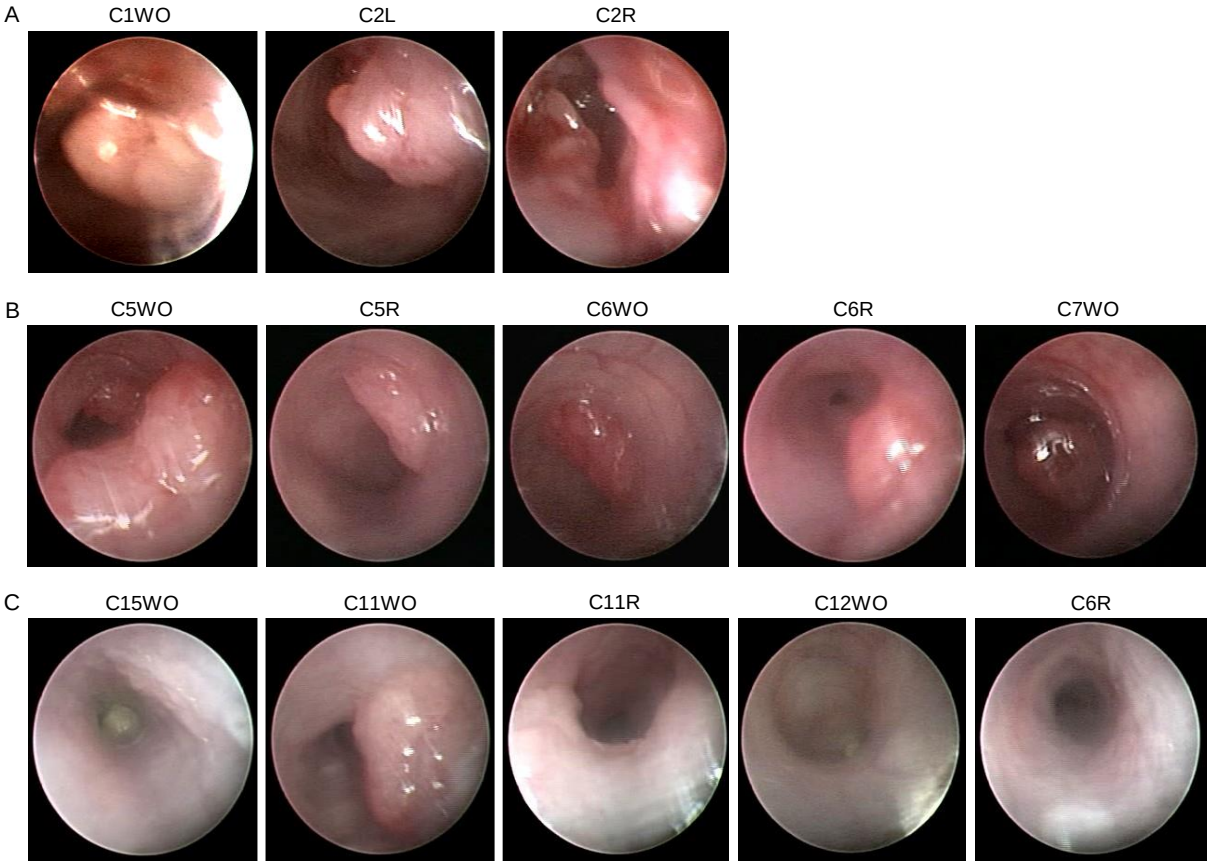


Figure S 20 View on distal tumors by coloscopy of AOM/DSS-treated mice at day of sacrifice
(A) AOM/DSS treatment set 1 (high DSS dose). (B) AOM/DSS treatment set 2 (medium DSS dose). (C) AOM/DSS treatment set 3 (low DSS dose).

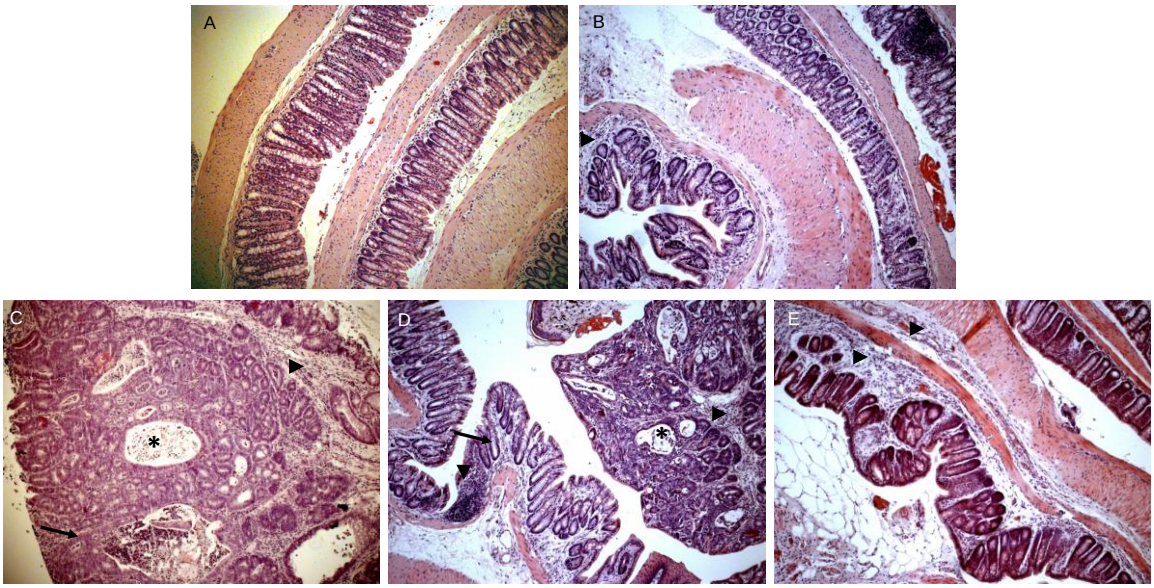


Figure S 21 Microscopic features of all treatment types
H&E staining of the colon. The arrows indicate crypt elongation, the arrow-heads infiltration of immune cells and the asterisks crypt abscesses. (A) Sample C4R (control, set 1). (B) Sample C8WO (DSS, set 2 (medium DSS dose)). (C) Sample C2L (AOM/DSS, set 1 (high DSS dose)). (D) Sample C5WO (AOM/DSS, set 2 (medium DSS dose)). (E) C15WO (AOM/DSS, set 3 (low DSS dose)).

6.2.2.2 WES sequencing results

Sample	Treatment	Set	Run / lane	Library batch	# raw reads	# filtered reads	# mapped (% mapped)	# duplicates (% duplicates)	Avg. coverage (target region)	% bp on target	Avg. insert size	Avg. mapping quality	Avg. base quality (filtered)	Avg. read length (filtered)
C1WO_AN	AOM/DSS proximal	1	1/2	1	102,792,060	91,318,654	90,418,058 (99.0%)	11,605,076 (12.8%)	46.5	31.3	132.3	51.2	36.3	99.1
C1WO_T1	AOM/DSS tumor	1	1/1	1	126,406,776	109,209,690	108,030,771 (98.9%)	15,003,277 (13.9%)	48.3	27.4	113.1	50.9	36.1	99.0
C1WO_T3	AOM/DSS tumor	1	1/1	1	119,002,468	100,983,716	99,942,471 (99.0%)	14,225,311 (14.2%)	50.1	31.1	123.6	50.3	36.1	98.9
C2L_AN	AOM/DSS proximal	1	1/3	1	99,865,546	87,017,316	86,444,450 (99.3%)	12,522,840 (14.5%)	43.2	31.1	119.0	50.8	36.3	99.3
C2L_T1	AOM/DSS tumor	1	1/1	1	113,964,146	102,018,680	101,059,622 (99.1%)	14,622,246 (14.5%)	44.6	27.3	130.2	50.5	36.2	99.0
C2L_T2	AOM/DSS tumor	1	1/1	1	128,290,592	109,951,316	108,828,918 (99.0%)	14,340,075 (13.2%)	52.1	29.2	114.4	50.4	36.2	99.1
C2R_AN	AOM/DSS proximal	1	1/3	1	118,880,754	106,922,814	106,331,573 (99.4%)	11,113,050 (10.5%)	57.3	31.8	123.9	51.2	36.3	99.3
C2R_T1	AOM/DSS tumor	1	1/3	1	129,628,180	114,776,976	114,011,783 (99.3%)	10,096,680 (8.9%)	63.8	32.2	116.1	51.6	36.3	99.2
C2R_T6	AOM/DSS tumor	1	1/2	1	110,764,602	99,835,144	99,061,643 (99.2%)	13,197,120 (13.3%)	51.4	31.5	115.1	51.4	36.4	99.4
C3R_D	DSS	1	1/2	1	108,435,456	97,219,092	96,529,057 (99.3%)	16,760,210 (17.4%)	43.9	29.1	123.7	51.2	36.4	99.3
C4R_C	Control	1	1/3	1	120,654,112	106,682,288	105,699,003 (99.1%)	8,726,263 (8.3%)	60.0	32.6	127.6	51.9	36.2	99.1
C4WO_C	Control	1	1/2	1	145,811,550	128,599,356	127,271,940 (99.0%)	14,490,515 (11.4%)	67.8	31.7	114.6	51.7	36.4	99.3
C5R_AN	AOM/DSS proximal	2	2/2	2	99,810,950	91,775,340	90,760,818 (98.9%)	18,236,511 (20.1%)	37.3	27.4	150.8	50.1	36.4	99.2
C5R_T1	AOM/DSS tumor	2	2/3	2	100,646,634	93,155,330	91,699,512 (98.4%)	15,301,477 (16.7%)	41.1	28.8	151.9	49.2	36.5	99.2
C5R_T2	AOM/DSS tumor	2	1/4	3	103,307,388	90,381,360	87,411,579 (96.7%)	14,957,570 (17.1%)	28.7	21.1	137.9	49.1	36.3	99.1
C5WO_AN	AOM/DSS proximal	2	2/4	2	145,986,308	133,616,536	132,405,825 (99.1%)	13,555,668 (10.2%)	61.5	27.5	134.0	50.7	36.5	99.4
C5WO_T1	AOM/DSS tumor	2	2/1	2	82,969,912	75,316,084	73,837,342 (98.0%)	13,901,194 (18.8%)	26.8	23.7	151.8	49.1	36.3	99.1
C5WO_T5	AOM/DSS tumor	2	1/6	3	103,816,236	91,511,408	90,030,793 (98.4%)	23,361,938 (26.0%)	32.7	26.0	130.2	50.0	36.3	99.2
C6R_AN	AOM/DSS proximal	2	1/4	3	124,560,478	111,572,464	109,545,250 (98.2%)	15,334,156 (14.0%)	42.3	23.9	139.5	49.5	36.4	99.3
C6R_T1	AOM/DSS tumor	2	2/1	2	127,920,242	114,556,136	112,813,468 (98.5%)	16,107,408 (14.3%)	42.5	23.4	157.8	49.3	36.2	98.9

Sample	Treatment	Set	Run / lane	Library batch	# raw reads	# filtered reads	# mapped (% mapped)	# duplicates (% duplicates)	Avg. coverage (target region)	% bp on target	Avg. insert size	Avg. mapping quality	Avg. base quality (filtered)	Avg. read length (filtered)
C6R_T2	AOM/DSS tumor	2	1/5	3	113,195,124	99,772,642	98,821,079 (99.0%)	21,202,218 (21.5%)	44.0	29.9	136.6	50.8	36.3	99.2
C6WO_AN	AOM/DSS proximal	2	2/6	2	128,427,628	116,777,736	115,680,131 (99.1%)	11,688,647 (10.1%)	62.8	31.8	110.1	51.4	36.2	99.1
C6WO_T1	AOM/DSS tumor	2	2/2	2	100,804,548	92,575,292	91,403,399 (98.7%)	18,273,082 (20.0%)	38.2	27.8	155.3	49.9	36.3	99.1
C6WO_T2	AOM/DSS tumor	2	1/4	3	123,818,964	109,805,538	108,176,830 (98.5%)	13,810,036 (12.8%)	41.5	23.2	140.8	50.3	36.4	99.2
C7WO_AN	AOM/DSS proximal	2	2/3	2	91,209,834	84,347,280	83,274,199 (98.7%)	19,425,907 (23.3%)	32.3	27.1	153.0	49.3	36.5	99.3
C7WO_T1	AOM/DSS tumor	2	2/4	2	93,680,372	84,920,038	83,520,510 (98.3%)	8,739,849 (10.5%)	37.9	26.9	163.6	49.7	36.2	98.9
C7WO_T3	AOM/DSS tumor	2	1/5	3	109,759,536	96,021,520	94,967,642 (98.9%)	18,299,150 (19.3%)	44.3	30.5	141.9	50.6	36.2	99.1
C8R_D	DSS	2	2/5	2	88,780,402	80,524,854	78,964,493 (98.1%)	10,015,232 (12.7%)	37.8	28.9	107.7	51.0	36.2	99.0
C8WO_D	DSS	2	2/2	2	98,210,336	90,458,644	89,737,464 (99.2%)	21,876,094 (24.4%)	36.1	28.2	144.0	50.8	36.4	99.3
C9WO_D	DSS	2	1/4	3	113,455,370	101,399,544	100,090,526 (98.7%)	17,168,776 (17.2%)	33.7	21.5	130.0	49.7	36.5	99.4
C10R_C	Control	2	1/5	3	128,424,668	112,287,052	110,825,736 (98.7%)	21,946,858 (19.8%)	50.4	30.4	139.5	50.5	36.3	99.1
C10WO_C	Control	2	2/2	2	135,120,018	124,903,160	123,666,158 (99.0%)	20,309,392 (16.4%)	54.0	27.9	146.9	50.1	36.4	99.3
C11R_AN	AOM/DSS proximal	3	2/7	2	117,890,186	107,058,076	106,037,556 (99.0%)	10,246,160 (9.7%)	59.6	32.9	110.9	51.2	36.2	99.0
C11R_T1	AOM/DSS tumor	3	2/6	2	98,594,558	90,011,126	88,836,404 (98.7%)	13,001,184 (14.6%)	44.9	31.3	114.3	50.4	36.2	99.2
C11WO_AN	AOM/DSS proximal	3	2/3	2	130,765,832	120,902,560	120,034,219 (99.3%)	17,376,514 (14.5%)	56.4	29.1	136.0	51.1	36.6	99.4
C11WO_T1	AOM/DSS tumor	3	2/1	2	107,721,670	97,600,676	95,108,746 (97.4%)	15,453,849 (16.3%)	37.3	24.9	135.8	47.5	36.4	99.3
C12R_AN	AOM/DSS proximal	3	1/5	3	117,475,034	103,974,956	103,051,684 (99.1%)	20,518,864 (19.9%)	47.2	30.3	137.3	51.0	36.3	99.2
C12R_T1	AOM/DSS tumor	3	2/5	2	119,661,324	109,124,302	107,634,849 (98.6%)	11,381,757 (10.6%)	51.3	28.2	113.9	49.4	36.3	99.1
C12WO_AN	AOM/DSS proximal	3	2/7	2	123,715,202	112,061,616	111,026,435 (99.1%)	10,760,263 (9.7%)	60.4	31.8	104.6	51.3	36.2	99.1
C12WO_T2	AOM/DSS tumor	3	2/6	2	116,034,126	105,478,786	104,107,461 (98.7%)	11,675,908 (11.2%)	53.6	30.7	118.0	50.8	36.1	99.0
C15R_AN	AOM/DSS proximal	3	2/4	2	93,765,874	85,450,174	84,262,691 (98.6%)	10,789,755 (12.8%)	36.0	26.1	158.9	49.9	36.3	99.1
C15R_T1	AOM/DSS tumor	3	1/6	3	117,511,638	103,096,988	101,953,300 (98.9%)	20,618,603 (20.2%)	41.6	27.1	137.9	50.5	36.3	99.1

Sample	Treatment	Set	Run / lane	Library batch	# raw reads	# filtered reads	# mapped (% mapped)	# duplicates (% duplicates)	Avg. coverage (target region)	% bp on target	Avg. insert size	Avg. mapping quality	Avg. base quality (filtered)	Avg. read length (filtered)
C15R_T2	AOM/DSS tumor	3	2/4	2	93,581,696	85,332,634	84,083,454 (98.5%)	8,886,850 (10.6%)	38.7	27.3	160.0	49.8	36.3	99.0
C15WO_AN	AOM/DSS proximal	3	1/6	3	133,705,724	117,536,650	116,059,291 (98.7%)	17,504,046 (15.1%)	53.6	28.8	125.1	50.6	36.3	99.3
C15WO_T1	AOM/DSS tumor	3	1/6	3	114,324,694	102,725,216	101,370,756 (98.7%)	20,932,547 (20.7%)	40.0	26.3	114.8	50.4	36.3	99.4
C16R_AN	AOM/DSS proximal	3	2/7	2	123,250,190	110,614,286	109,609,389 (99.1%)	9,929,615 (9.1%)	59.5	31.6	110.1	51.6	36.1	99.0
C16R_T1	AOM/DSS tumor	3	2/7	2	105,005,866	95,621,626	94,300,686 (98.6%)	12,594,290 (13.4%)	49.8	32.3	112.8	50.2	36.2	99.1
C18R_D	DSS	3	2/3	2	90,534,656	83,557,082	82,413,613 (98.6%)	11,166,910 (13.6%)	38.9	29.1	171.6	51.0	36.3	98.9
C19R_D	DSS	3	2/5	2	130,404,734	118,637,064	116,809,477 (98.5%)	10,739,390 (9.2%)	58.2	29.0	108.8	50.3	36.3	99.1
C19WO_D	DSS	3	2/6	2	119,894,542	108,766,176	107,923,663 (99.2%)	10,874,520 (10.1%)	61.7	33.7	120.9	51.2	36.1	99.0
C20R_C	Control	3	2/1	2	128,903,986	116,591,324	114,475,038 (98.2%)	18,658,582 (16.3%)	41.8	23.4	153.3	49.1	36.3	99.1
C21WO_C	Control	3	2/5	2	122,450,302	110,864,528	109,951,418 (99.2%)	9,277,680 (8.4%)	53.2	27.9	119.1	51.0	36.2	98.9

Table S 27 Sequencing and mapping results based on the WES within the AOM/DSS mouse experiment

The first part of each sample name refers to the mouse ID, while the second part refers to the sample type (AN = formerly inflamed tissue from the proximal colon part from an AOM/DSS-treated mouse, Tx = Tumor tissue from an AOM/DSS-treated mouse, C = proximal colon tissue from an untreated control mouse, D = formerly inflamed tissue from the proximal colon part from a DSS-treated mouse

6.2.2.3 Sample distribution (WES data)

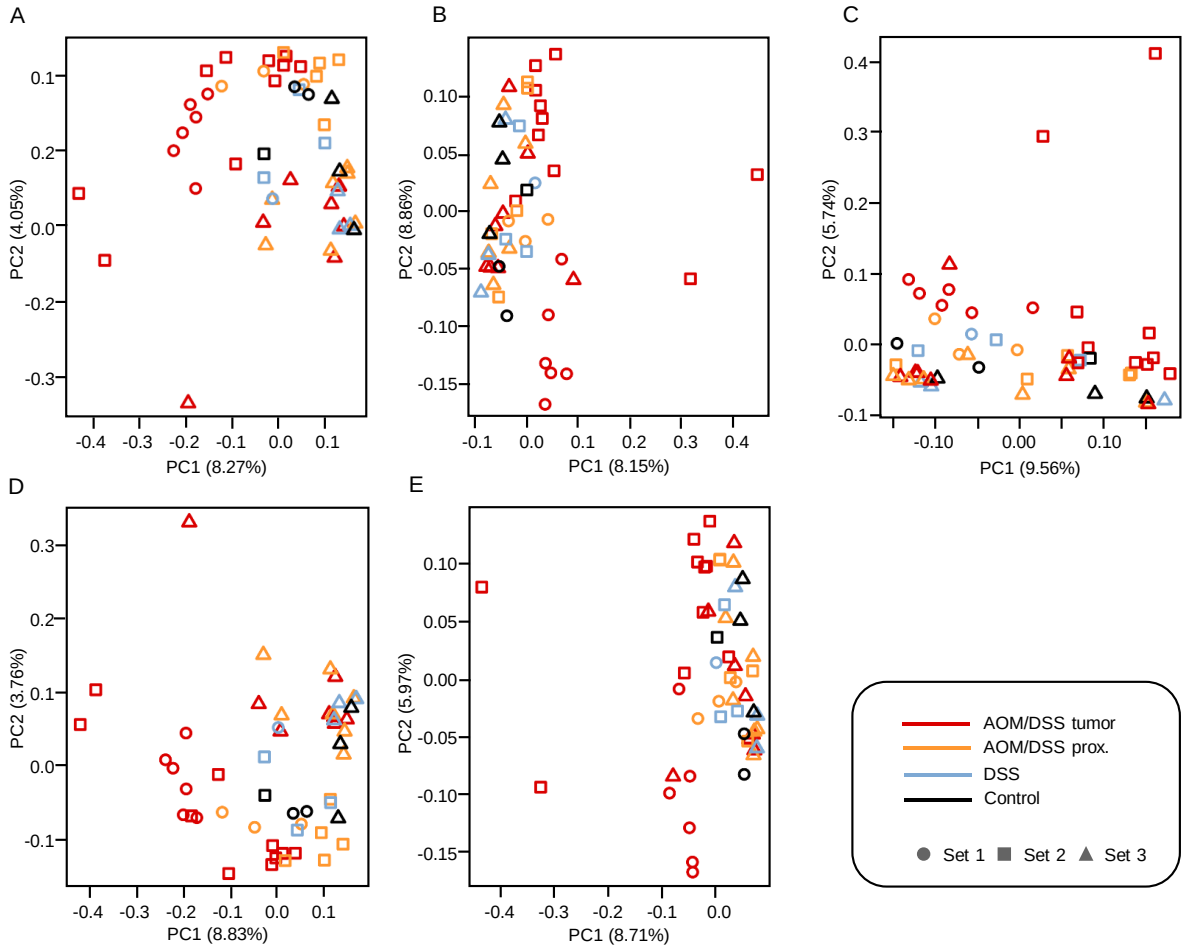


Figure S 22 PCoAs on variant and gene level
(A) PCoA based on genes affected by non-synonymous variants. (B) PCoA based on number of non-synonymous variants per gene. (C) PCoA based on exonic, non-synonymous variants. (D) PCoA based on genes affected by a variant (synonymous or non-synonymous). (E) PCoA based on number of synonymous and non-synonymous variants per gene. (A) - (D) The set indicates the DSS dose.

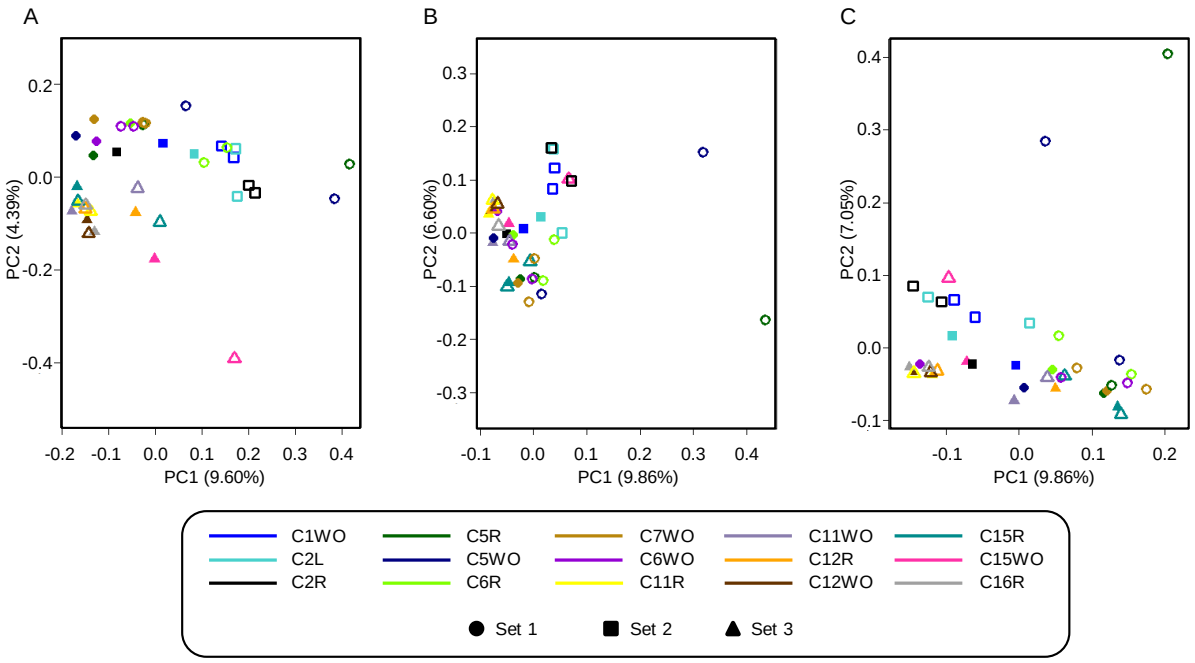


Figure S 23 PCoAs of samples from AOM/DSS-treated mice
In the plots, the color indicates the mouse ID, while the shape refers to the treatment set. The set is based on the DSS treatment conditions. Filled symbols indicate proximal, non-tumor samples from AOM/DSS-treated mice and unfilled ones refer to tumor samples. (A) PCoA based on genes affected by at least one SNV or InDel. (B) PCoA based on number of variants per gene. (C) PCoA based on exonic SNVs and InDels.

A

		Affected genes		# variants per gene	
Set	Sample	Same mouse	All tumors	Same mouse	All tumors
1	C1WO_T1	0.788	0.788	0.516	0.516
	C1WO_T3	0.788	0.782	0.516	0.510
	C2L_T1	0.797	0.777	0.546	0.513
	C2L_T2	0.797	0.792	0.546	0.528
	C2R_T1	0.806	0.806	0.560	0.534
	C2R_T6	0.806	0.795	0.560	0.540
2	C5R_T1	0.852	0.722	0.615	0.432
	C5R_T2	0.852	0.861	0.615	0.640
	C5WO_T1	0.855	0.767	0.608	0.474
	C5WO_T5	0.855	0.851	0.608	0.608
	C6R_T1	0.785	0.769	0.517	0.483
	C6R_T2	0.785	0.785	0.517	0.508
	C6WO_T1	0.638	0.708	0.377	0.424
	C6WO_T2	0.638	0.723	0.377	0.432
	C7WO_T1	0.644	0.739	0.391	0.453
	C7WO_T3	0.644	0.735	0.391	0.436
3	C15R_T1	0.656	0.767	0.388	0.474
	C15R_T2	0.656	0.697	0.388	0.392

B

		SNVs / InDels		Affected genes		# variants per gene	
Set	Sample	Same mouse	All tumors	Same mouse	All tumors	Same mouse	All tumors
1	C1WO_T1	0.552	0.567	0.772	0.772	0.526	0.526
	C1WO_T3	0.552	0.552	0.772	0.771	0.526	0.519
	C2L_T1	0.583	0.564	0.779	0.768	0.555	0.537
	C2L_T2	0.583	0.570	0.779	0.774	0.555	0.534
	C2R_T1	0.584	0.581	0.798	0.795	0.573	0.546
	C2R_T6	0.584	0.585	0.798	0.785	0.573	0.542
2	C5R_T1	0.651	0.500	0.845	0.724	0.633	0.453
	C5R_T2	0.651	0.671	0.845	0.860	0.633	0.654
	C5WO_T1	0.652	0.549	0.849	0.754	0.626	0.492
	C5WO_T5	0.652	0.640	0.849	0.842	0.626	0.618
	C6R_T1	0.555	0.528	0.760	0.754	0.509	0.491
	C6R_T2	0.555	0.549	0.760	0.761	0.509	0.509
	C6WO_T1	0.453	0.485	0.606	0.697	0.386	0.433
	C6WO_T2	0.453	0.490	0.606	0.717	0.386	0.449
	C7WO_T1	0.467	0.523	0.625	0.730	0.388	0.456
	C7WO_T3	0.467	0.491	0.625	0.711	0.388	0.441
3	C15R_T1	0.458	0.528	0.654	0.760	0.394	0.491
	C15R_T2	0.458	0.458	0.654	0.663	0.394	0.399

Figure S 24 Distance comparison between tumor samples
Comparison between the distance to the tumor from the same mouse and the median distance to all other tumors. For all paired columns, the black background highlights the larger diversity, a grey color indicates the same value. The set is based on the DSS treatment conditions. (A) The distance calculation on gene level is based on all exonic variants. In the first column pair, genes affected by at least one exonic variant were used as input, while in the second one, the number of variants per gene were used. (B) The distance calculation is based on all non-synonymous SNVs and InDels. The input for the first column pairs are variants, for the second pair, genes affected by at least one variant, and for the third one, the number of variants per gene.

6.2.2.4 Strain specific differences

Many SNVs occurred in several samples including the controls. It was therefore assumed that these variants were due to a discrepancy in host genetic makeup: C57BL/6N mice were used in the experiment, but the sequences were mapped against a C57BL/6J reference, because it existed no publicly available annotation for C57BL/6N. Differences between the two mouse strains reflect mutational patterns caused by evolutionary influences. In order to investigate this in more detail, all base substitutions called in at least 50% of the non-tumor samples were extracted. This resulted in 115,556 SNVs positions, whereof 2.2% were exonic. Based on a 5% allele support threshold, 101,740 variants were called heterozygous, of which 210 showed two non-reference alleles. For the last-mentioned variants, just the allele with the highest read support was used for the following analyses.

After normalization for the length of the enriched regions and after exclusion of random and additional contigs / scaffolds, the highest variability was detected on chromosomes M (mitochondrial DNA), X, Y, and 17. The largest degree of conservation was observed on chromosomes 9, 14, 18, and 19 (Figure S 25).

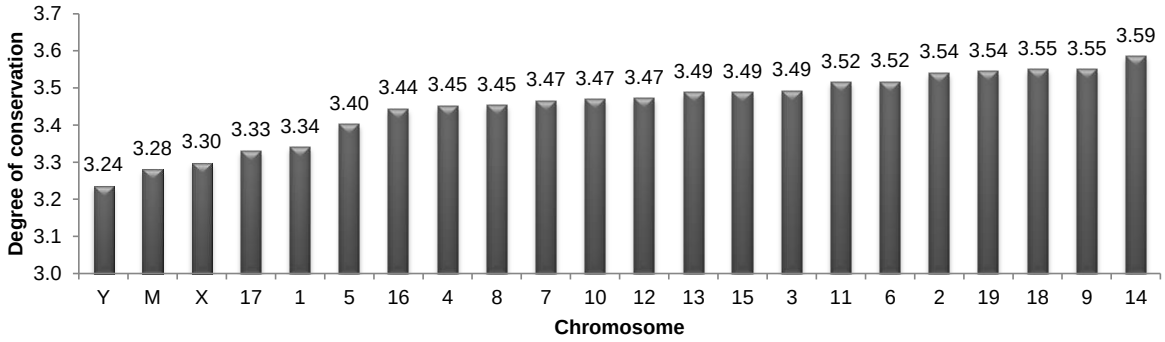


Figure S 25 Conservation degree for each chromosome based on strain dependent SNVs
The degree of conservation was defined as log10(<target region length> / <# SNVs in target region>). In the figure, the chromosomes are in ascending order according to their conservation levels. All unplaced and unlocalized chromosomes were excluded.

Using rainfall plots, it could be shown that strain specific SNVs form hypervariable regions (Figure S 26 A). To investigate the size and density of these clusters more detailed, the sliding window method with three different start parameters was used. First, it was searched for small hotspots with a high SNV density. Therefore, eight or more substitutions within 160 bp frame were required. This resulted in 2,680 regions with an average length of 227 bp and a mean of 16 SNVs. Next, larger clusters were analyzed with at least 50 variants in a 5,000 bp window. Using these options, 381 hotspots with on average 161 variants in regions with a mean size of 8,502 bp were detected. Last, parameters to find medium size clusters with a high variant density were used. Therefore, at least 50 SNVs in 500 bp regions were required. Even with this more stringent strategy, 60 regions with an average size of 658 bp and a mean number of 74 SNVs were identified. This clearly demonstrates that the strain-associated SNVs were enriched in specific regions.

Next, SNVs including the flanking bases were investigated (Figure S 26 B+C). Most of the differences between the strains C57BL/6N and C57BL/6J were C>T substitutions in a GpCpG context followed by the opposite T>C variants without any 5' or 3' preference. These variants were followed by C>A substitutions with a 5' G and the opposite T>G in a GpTpG context.

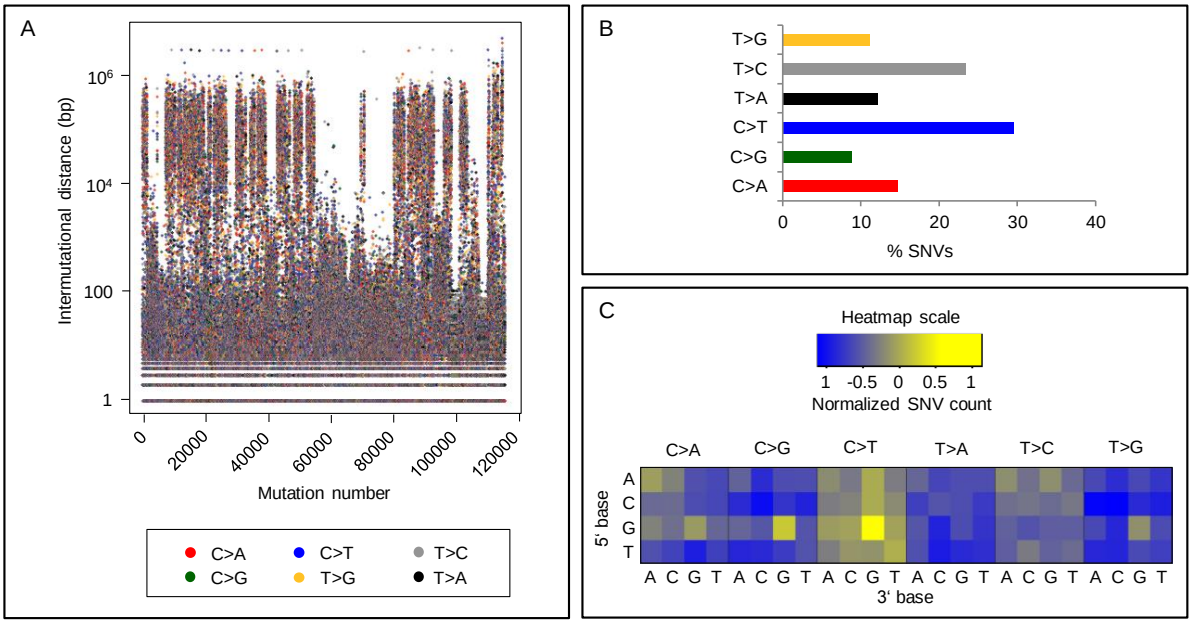


Figure S 26 Strain specific SNV patterns
All figures are based on SNVs, which were called in at least 50% of the non-tumor samples. (A) Rainfall plot showing SNV hotspots. SNVs were ordered by genomic position on the x-axis. The genomic distances between adjacent SNVs are plotted on the y-axis. The colors of the dots describe the SNV type. (B) SNV distribution of the six SNV types. (C) Heatmap showing the log-transformed normalized counts of the base context for each substitution type. The 5‘ base is shown on the vertical axis and the 3‘ base on the horizontal axis.

6.2.2.5 Variants and mutated genes in murine AOM/DSS-triggered colorectal cancer

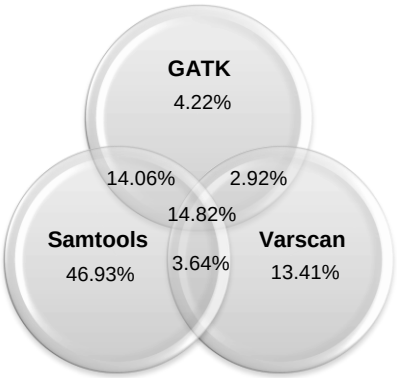


Figure S 27 SNV caller comparison
The figure shows the overlap of SNVs called with GATK, Samtools or Varscan, respectively.

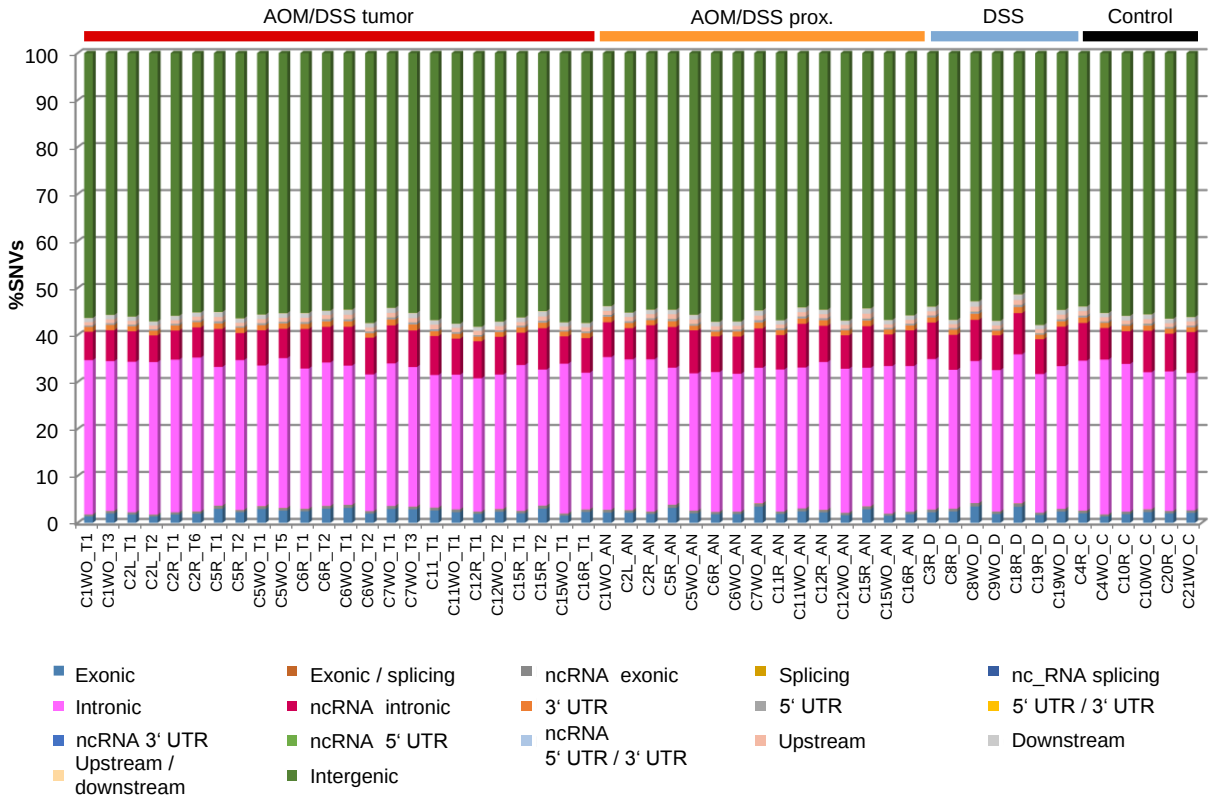


Figure S 28 SNV region distribution based on all called SNVs
No significant differences were observed between the different sample groups.

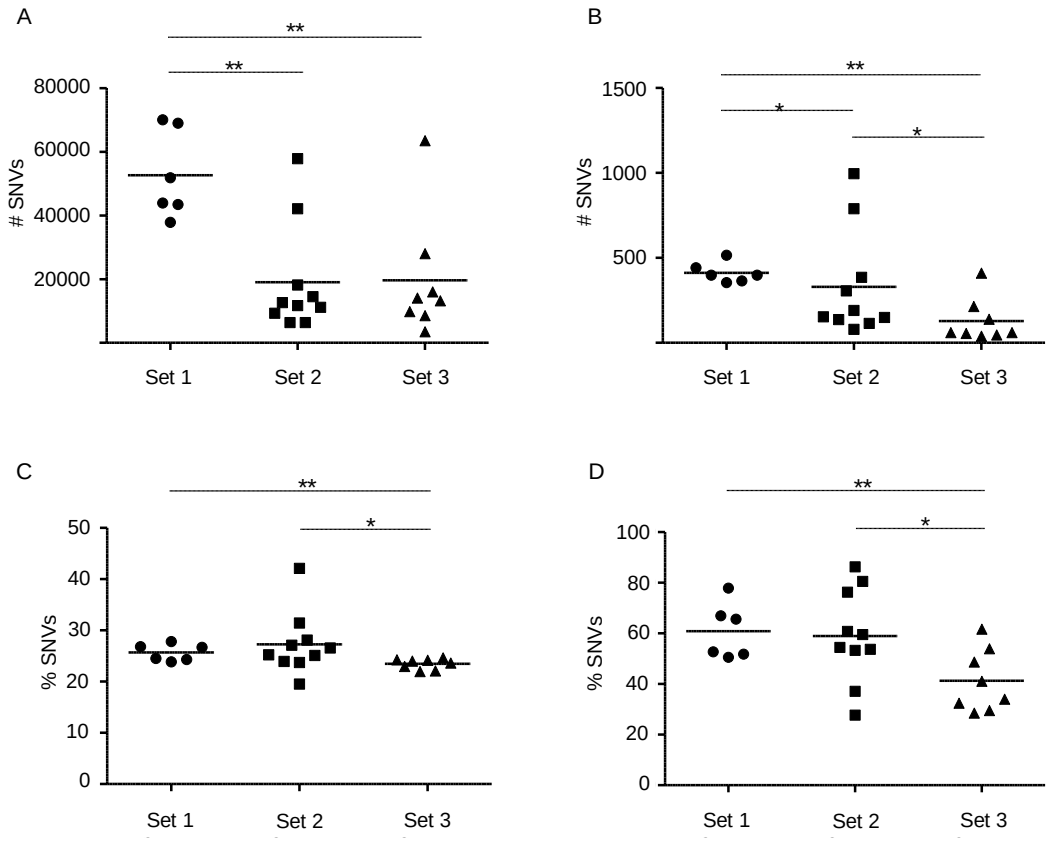


Figure S 29 SNV count comparison between tumor samples of the three DSS treatment sets
(A) Number of all somatic SNVs. (B) Number of all exonic, somatic SNVs. (C) Relative proportion of somatic C>T substitutions. (D) Relative proportion of somatic exonic C>T variants.

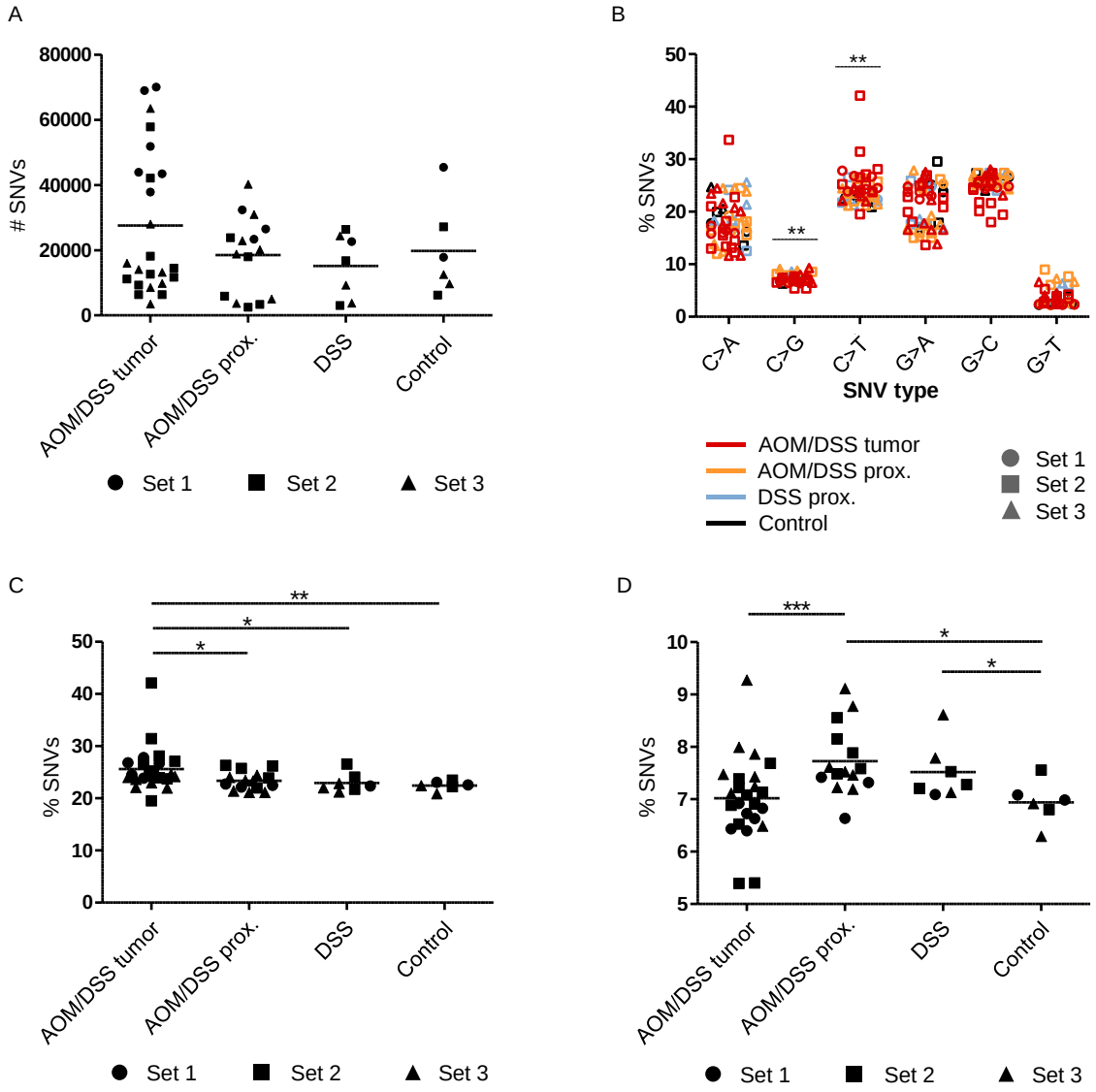


Figure S 30 SNV distribution based on all somatic SNVs
The set is based on the DSS treatment conditions. (A) Number of SNVs. (B) SNV type distribution. Significance was proofed with Kruskal-Wallis test. (C) Relative proportion of C>T base substitutions. (D) Relative proportion of C>G base substitutions.

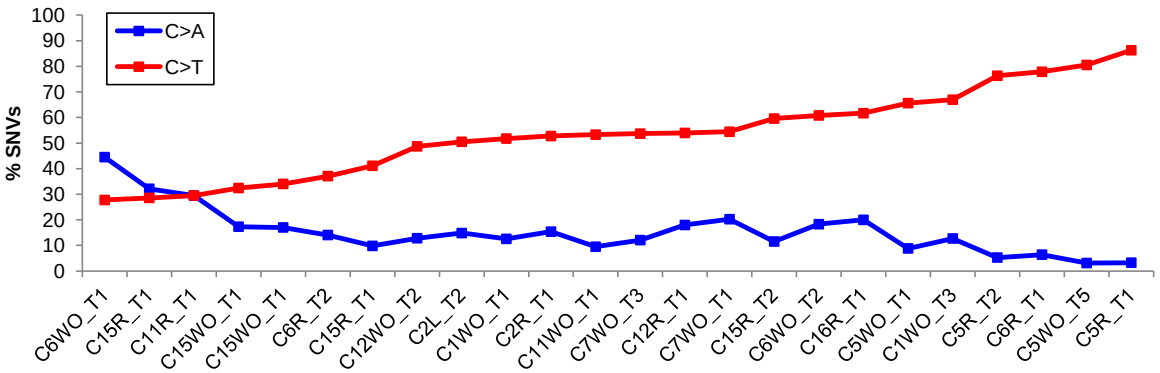


Figure S 31 Comparison between C>A and C>T substitutions
The percentage corresponds to the relative proportion of the SNV type in the total number of exonic somatic SNVs. The samples are ordered by the percentage of C>T SNVs.

SNV type	Treatment 1	Treatment 2	P-value
C>T	AOM/DSS tumor	AOM/DSS proximal	0.0075
C>T	AOM/DSS tumor	DSS	0.0124
C>T	AOM/DSS tumor	Control	0.0028
C>G	AOM/DSS tumor	AOM/DSS proximal	<0.00001
C>G	AOM/DSS tumor	DSS	0.0345
C>G	AOM/DSS proximal	Control	0.0057
C>G	DSS	Control	0.0111
T>A	AOM/DSS tumor	Control	0.0024
T>C	AOM/DSS tumor	AOM/DSS proximal	0.0013

Table S 28 Pairwise comparisons for each SNV type (somatic)
All comparisons with significant p-value are included in the table. The test was performed with a one-tailed Wilcoxon Rank-Sum test (Wilcoxon Signed-Rank for AOM/DSS tumor vs. AOM/DSS proximal).

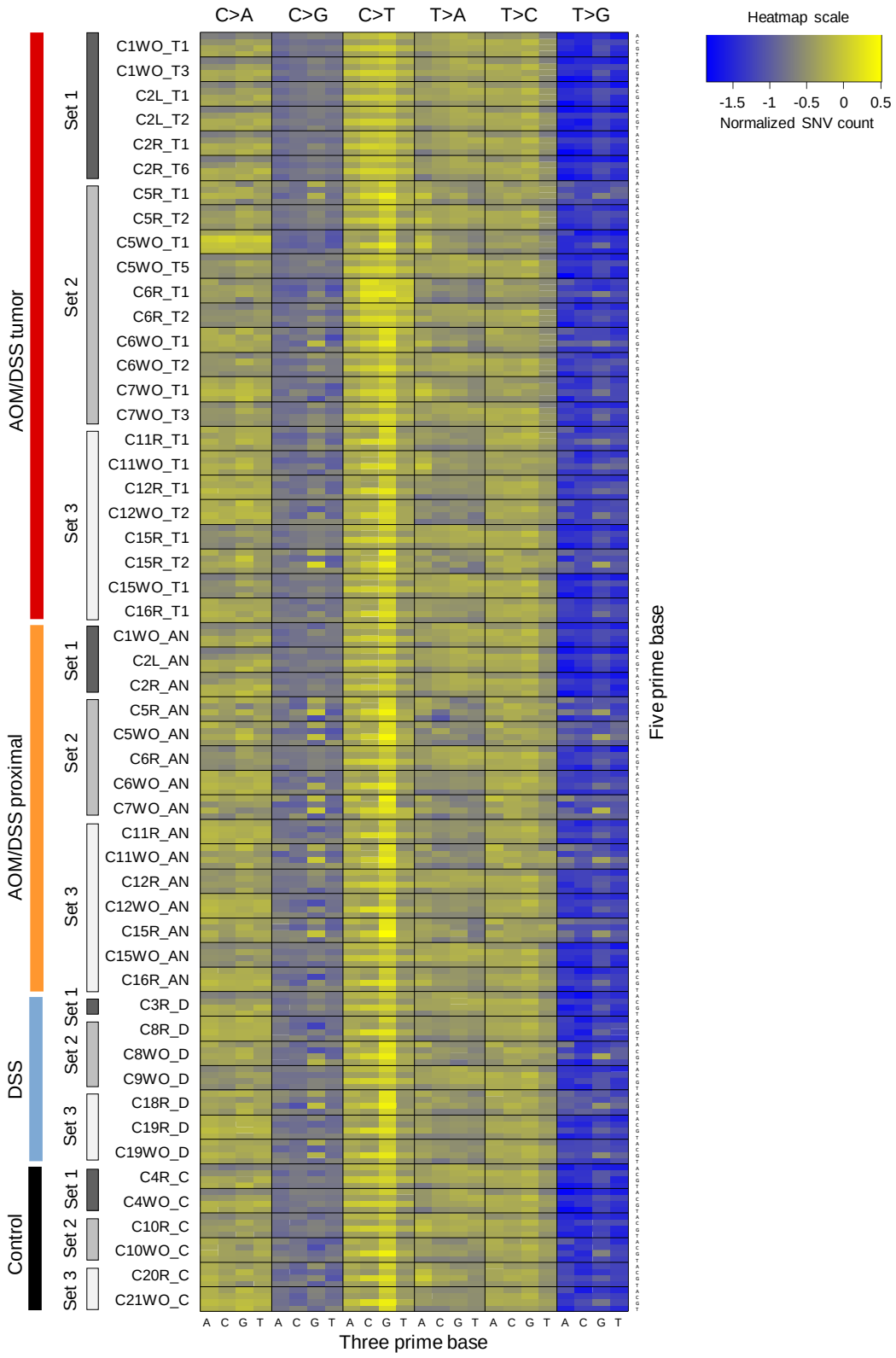


Figure S 32 Heatmap based on number of SNVs including base context
Heatmap shows the log-transformed normalized number of each substitution type including the base context for all samples. The base, which is five prime located of the SNV, is shown on the vertical axis, while the 3‘ base is on the horizontal axis. No clear clustering of the samples was detectable. The set is based on the DSS treatment conditions.

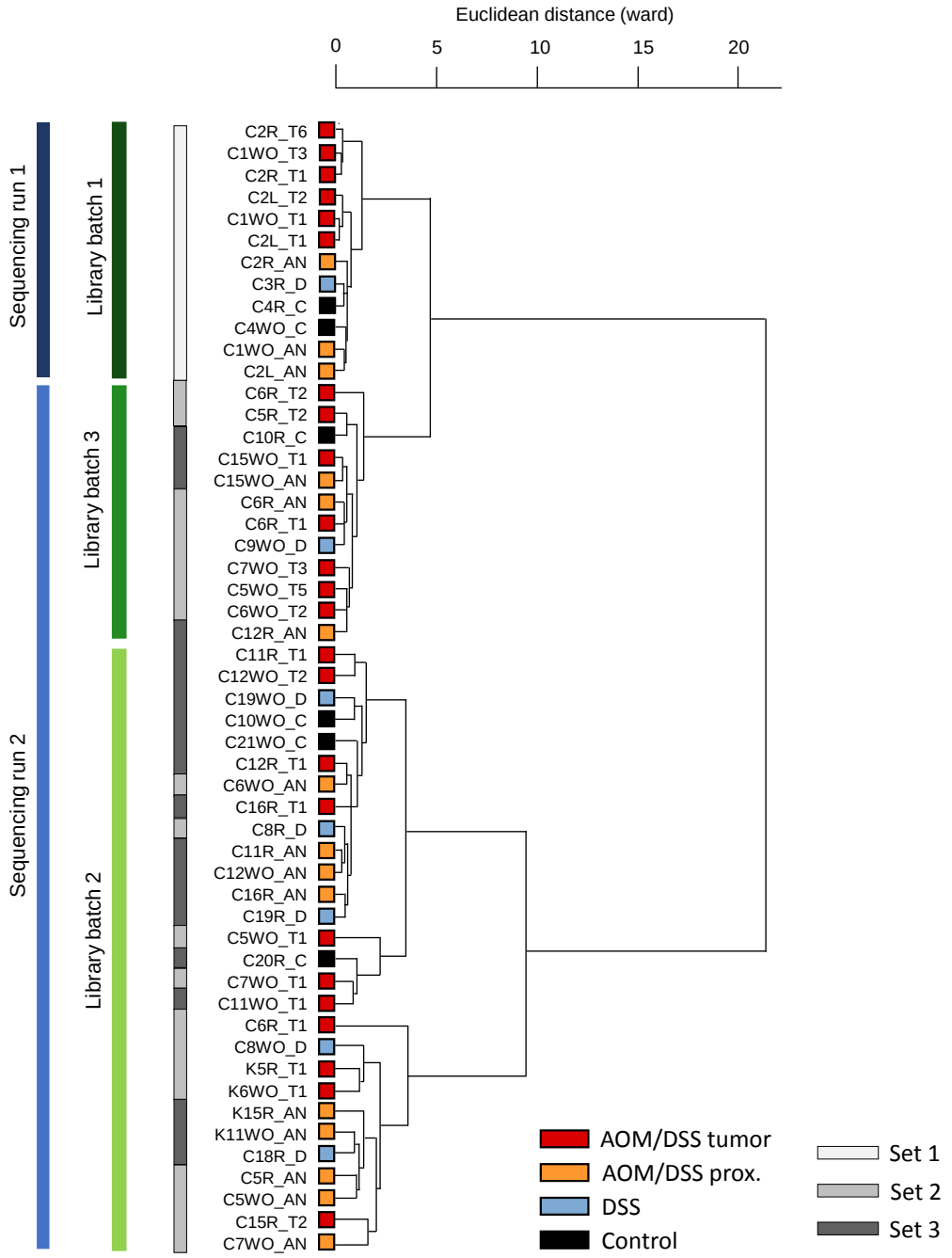


Figure S 33 Hierarchical clustering based on all SNVs including base context
The clustering was performed with Euclidean distance and the ward method. Although a strong technical bias was observed, a distribution of the samples according to the treatment was detectable for the first treatment set (library batch 1 / sequencing run 1). In the two other treatment sets, no clustering was visible.

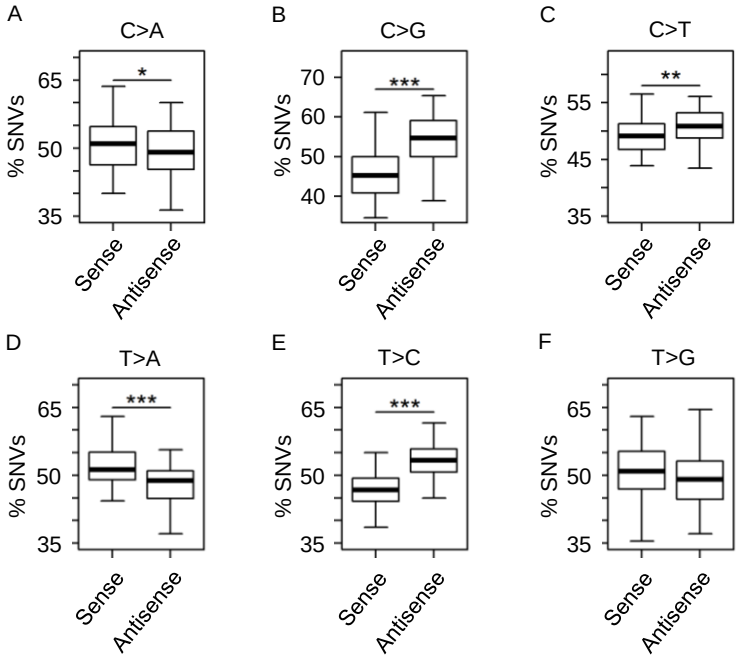


Figure S 34 Comparison between all SNVs (germline + somatic) on sense and antisense strand for all SNV types

The analysis was based on all exonic SNVs called in one of the investigated samples (tumor or non-tumor tissue samples from AOM/DSS-treated mice, tissue samples from DSS-treated mice or control mice). The following substitutions were on the sense strand: (A) C>A substitutions. (B) C>G substitutions. (C) C>T substitutions. (D) T>A substitutions. (E) T>C substitutions. (F) T>G substitutions.

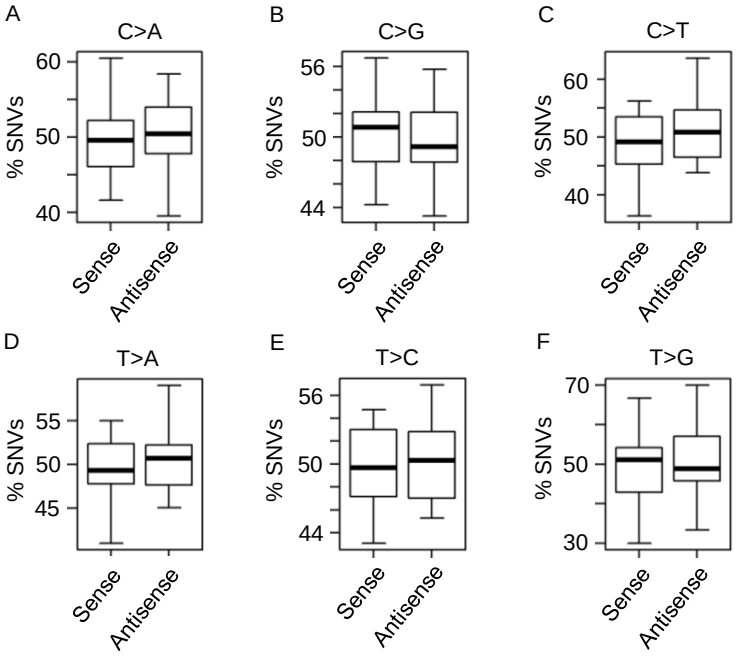


Figure S 35 Comparison between somatic SNVs on sense and antisense strand for all SNV types

The analysis was based on all somatic exonic SNVs of the tumor samples. The following substitutions were on the sense strand: (A) C>A substitutions. (B) C>G substitutions. (C) C>T substitutions. (D) T>A substitutions. (E) T>C substitutions. (F) T>G substitutions.

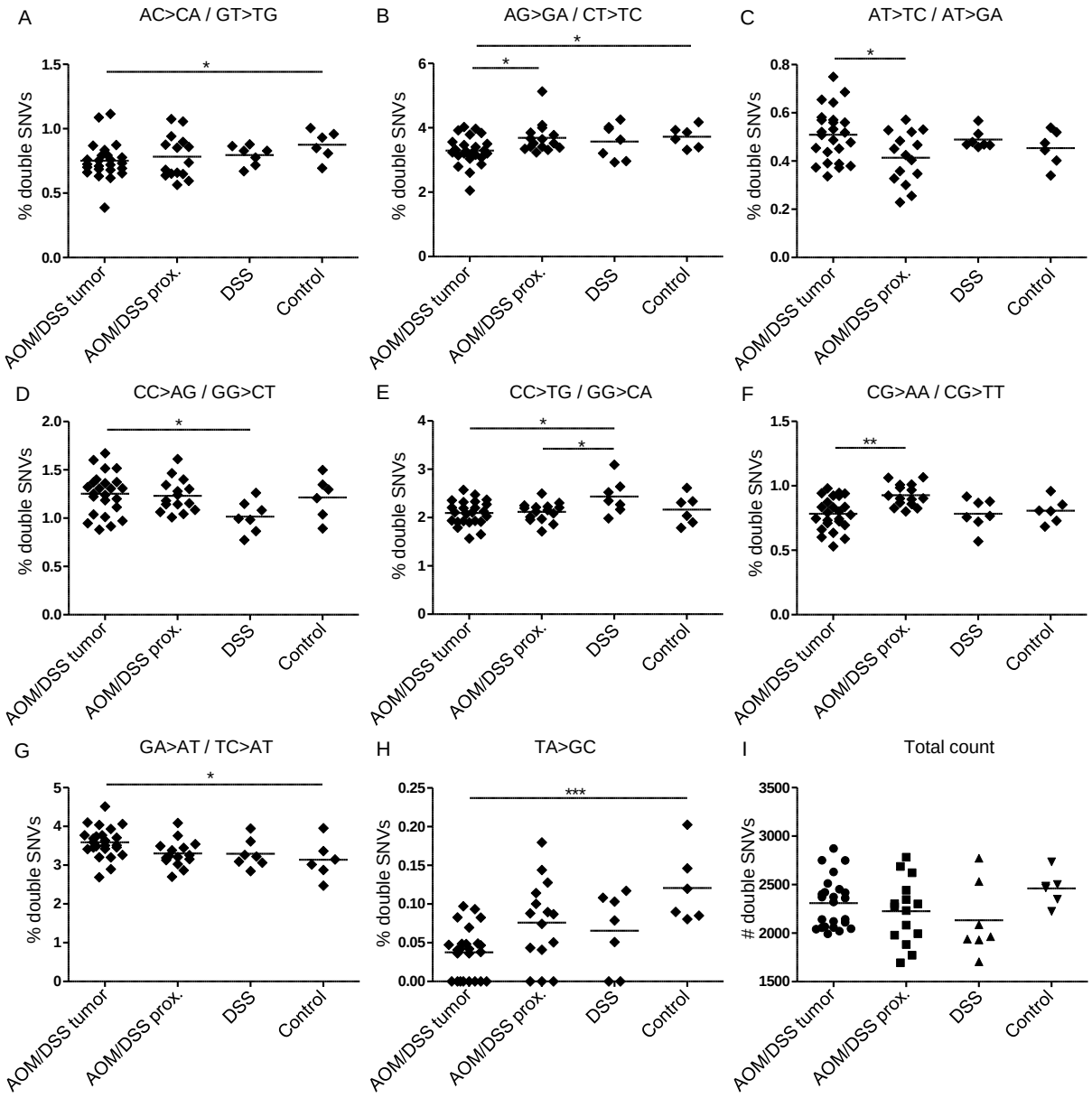
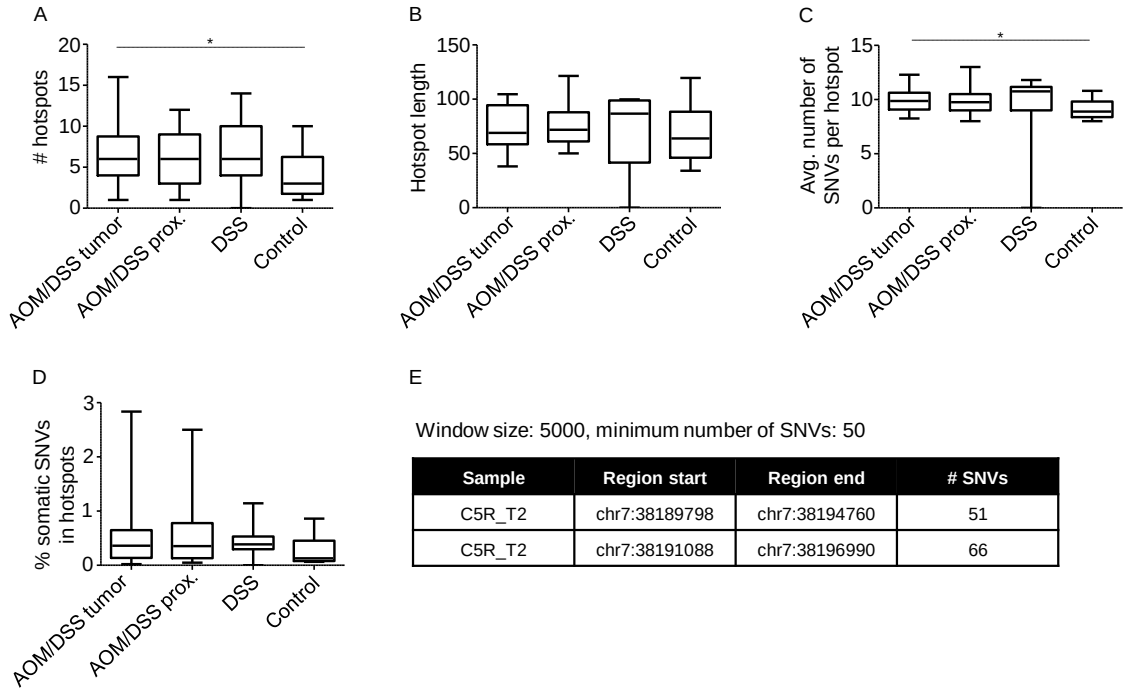


Figure S 36 Double SNVs
The figures show the proportion of each double SNV type within each sample type.



E

Window size: 5000, minimum number of SNVs: 50

Sample	Region start	Region end	# SNVs
C5R_T2	chr7:38189798	chr7:38194760	51
C5R_T2	chr7:38191088	chr7:38196990	66

Figure S 37 SNV hotspot detection with sliding window method
(A)-(D) Figures show the results of the hotspot comparison between the treatment groups based on an initial window size of 160 bp and an initial SNV count of eight. (A) Number of hotspots. (B) Hotspot lengths. (C) Average number of SNVs per hotspot. (D) Relative proportion of somatic SNVs located in hotspots. (E) Description of two hotspots located close together in the tumor sample C5R_T2 detected with an initial window size of 5000 bp and an initial SNV count of 50.

A

FPKM class	Lower threshold	Upper threshold
1	-	0.01
2	0.01	1
3	1	2
4	2	5
5	5	10
6	10	-

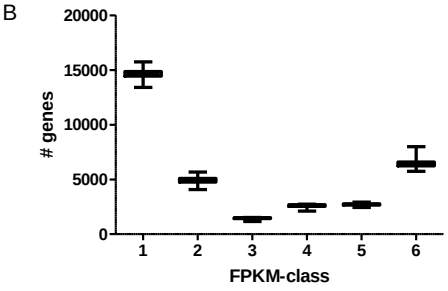


Figure S 38 FPKM class description
(A) FPKM thresholds for each class. (B) Number of genes per FPKM class across all samples.

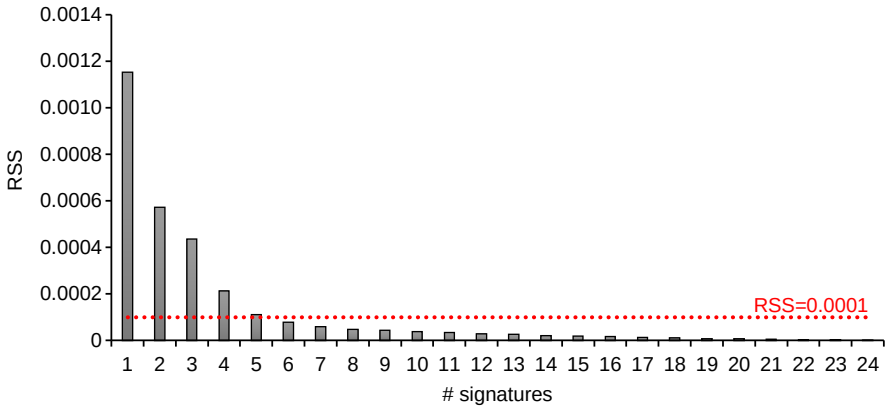


Figure S 39 RSS value distribution based on increasing number of contributing signatures
The optimal number of signatures that should be used in the SNV type estimation model was decided with the RSS value distribution. To achieve a model accuracy of 99% corresponding to an RSS value below 0.0001, six contributing signatures were necessary.

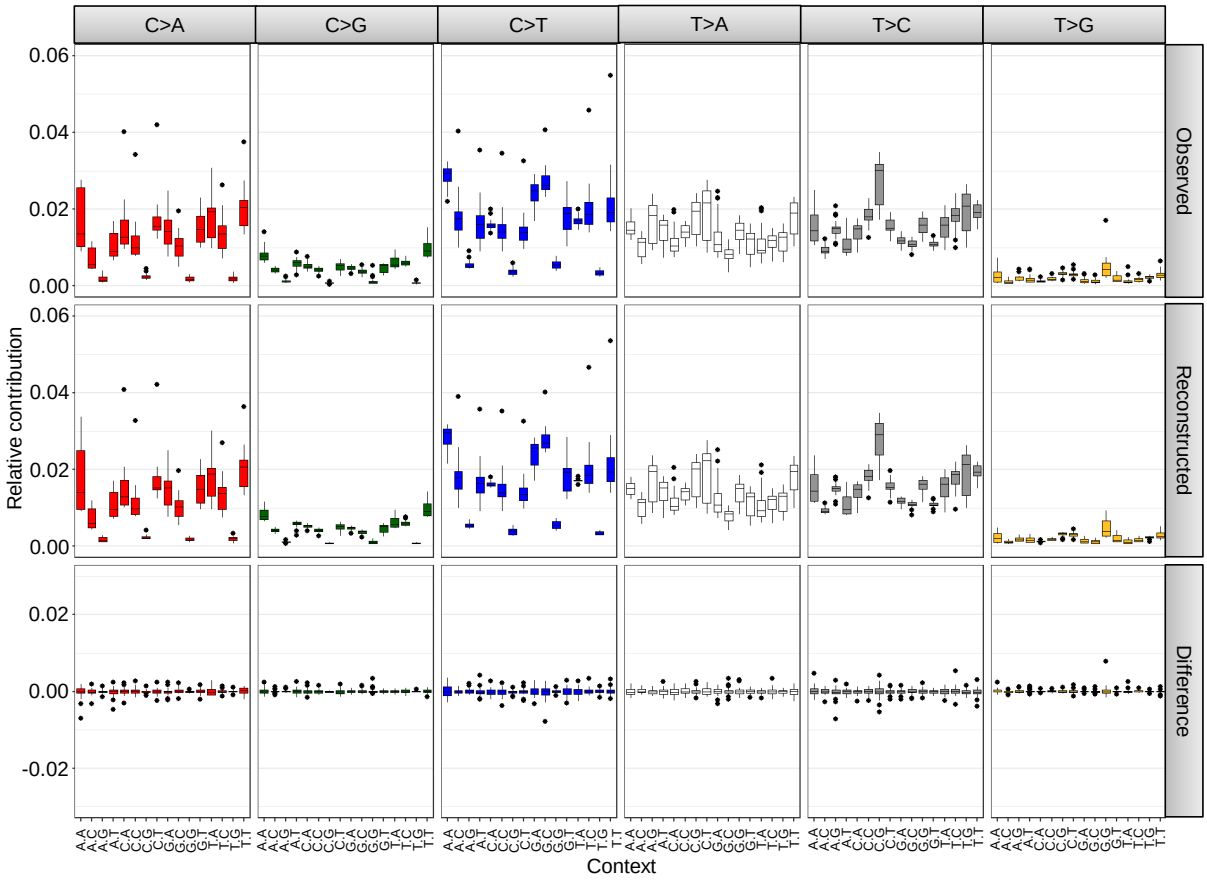


Figure S 40 Comparison between observed and reconstructed SNV patterns based on novel defined SNV signatures

Mouse						Human									# affected samples				Method	
Position	SNV type	PhyloP	Nt change	Aa change	Gene	Position	Gene	SNV type	Aa change	Sift score	Sift prediction	PolyPhen2 score	PolyPhen2 prediction	PhyloP score (rescaled by dbNSFP)	PhyloP prediction	AOM/DSS tumor	AOM/DSS proximal	DSS	Control	Method 1	Method 2
chr1:34288006	ns	6.177	T20351A, T14309A	I6784N, I4770N	*Dst*	chr6:56346861	*DST*	ns	I4550N	NA	NA	NA	NA	NA	NA	3	1	0	0	yes	no
chr1:110908073	ns	1.263	G1438A	A480T	*Cdh19*	chr18:64197096	*CDH19*	ns	A482T	0.85	T	0.01	B	0.978	C	3	0	0	0	yes	no
chr2:27997596	ns	3.815	C2878T	P960S	*Col5a1*	chr9:137688727 rs141301771	*COL5A1*	ns	P960S	0.99	D	0.70	NA	0.997	C	3	0	0	0	yes	no
chr2:73631801	stop	3.847	T657A, T282A	C219X, C94X	*Chn1*	chr2:175689217	*CHN1*	stop		NA	NA	NA	NA	NA	NA	3	0	0	0	yes	yes
chr2:74659083	ns	5.59-0	G337A	D113N	*Evx2*	chr2:176948168	*EVX2*	ns		0.99	D	0.78	P	0.999	C	3	0	0	0	yes	yes
chr2:125247758	stop	2.585	C11A	S4X	*Dut*	chr15:48624459	*DUT*	ns	S4Y	NA	NA	NA	NA	NA	NA	3	0	0	0	yes	no
chr2:131936461	ns	3.480	T32A	L11H	*Prnp*	chr20:4679898	*PRNP*	ns		1.00	D	0.29	P	0.997	C	3	1	0	0	yes	yes
chr2:154632289	ns	6.947	G1033A	E345K	*Zfp341*	chr20:32349678	*ZNF341*	ns		1.00	D	1.00	D	1.000	C	3	0	0	0	yes	yes
chr2:174299663	ns	0.586	C1802T	S601F	*Gnas*	chr20:57429846	*GNAS*	stop		NA	NA	NA	NA	NA	NA	3	0	0	0	yes	yes
chr3:88113945	ns	1.030	G4001C	G1334A	*Iqgap3*	chr1:156502877	*IQGAP3*	ns	G1333A	0.02	T	0.30	P	0.970	C	4	2	0	0	yes	yes
chr4:81386477	ns	2.497	T185C	V62A	*Mpdz*	chr9:13224581	*MPDZ*	ns	V62A							3	2	1	0	yes	no
chr4:95996765	ns	6.321	A481T	S161C	*Hook1*	chr1:60302551	*HOOK1*	ns	S161C	0.96	D	0.61	P	0.999	C	3	0	0	0	yes	yes
chr4:117887184	ns	0.847	G53A	C18Y	*Atp6v0b*	chr1:44440765	*ATP6V0B*	ns	C18Y	0.98	D	0.04	B	0.976	C	3	0	0	0	yes	yes
chr4:129601102	stop	0.530	C136T	Q46X	*Tmem234*	chr1:32687500	*C1orf91, TMEM234*	stop	Q46X	0.90	NA	0.73	NA	0.999	C	3	1	0	0	yes	yes
chr4:156331188	ns	0.607	A7C	K3Q	*Vmn2r-ps159*	no liftover possible										8	4	1	0	yes	no
chr5:102963505	ns	7.358	T1072C, T958C	Y358H, Y320H	*Mapk10*	chr4:86985457	*MAPK10*	ns	Y106H	1.00	D	1.00	D	0.999	C	3	0	0	0	yes	yes
chr5:107723792	ns	5.923	A47G, A245G	H16R, H82R	*Gfi1*	chr1:92948998	*GFI1*	ns	H16R	1.00	D	1.00	D	0.996	C	5	1	1	0	yes	yes
chr5:111419864	ns	5.002	C1699T	H567Y	*Mn1*	chr22:28194794	*MN1*	ns	H580Y	0.94	NA	1.00	D	0.999	C	3	1	0	0	yes	yes
chr5:134409235	ns	4.283	T497A	V166E	*Gtf2ird1*	chr7:73932544	*GTF2IRD1, RBAP2*	ns	V166E, V198E	1.00	D	0.97	D	0.996	C	3	1	1	0	yes	yes
chr5:134673121	ns	3.290	A86T, A182T	H29L, H61L	*Limk1*	chr7:73510981	*LIMK1*	ns	H27L	0.99	D	0.99	D	0.998	C	3	1	1	0	yes	yes

Mouse						Human										# affected samples				Method	
Position	SNV type	PhyloP	Nt change	Aa change	Gene	Position	Gene	SNV type	Aa change	Sift score	Sift prediction	PolyPhen2 score	PolyPhen2 prediction	PhyloP score (rescaled by dbNSFP)	PhyloP prediction	AOM/DSS tumor	AOM/DSS proximal	DSS	Control	Method 1	Method 2
chr5:151575684	stop	1.966	G1314A	W438X	*Vmn2r18*	no liftover possible										3	0	0	0	yes	no
chr6:56780361	ns	7.333	C389T	A130V	*Kbtbd2*	chr7:32910440	*KBTBD2*	ns	A130V	0.49	T	0.26	P	1.000	C	3	0	0	0	yes	yes
chr6:114131434	ns	6.677	A158G	E53G	*Slc6a11*	chr3:10858123	*SLC6A11*	ns	E58G	1.00	D	1.00	D	0.991	C	3	1	0	0	yes	yes
chr7:4119721	ns	1.564	G67A	D23N	*Ttyh1*	chr19:54926793	*TTYH1*	ns	D23N	0.21	T	0.00	B	0.994	C	3	0	0	0	yes	no
chr7:16810900	ns	1.499	T1036G	S346A	*Fkrp*	chr19:47259743	*FKRP*	ns	S346A	0.98	D	0.98	D	0.996	C	3	1	0	0	yes	yes
chr7:67662416	ns	0.400	T70A	S24T	*Ttc23*	chr15:99768848	*TTC23*	ns	H24N	1.00	D	0.00	B	0.958	C	3	0	0	0	yes	yes
chr8:13143130	ns	6.809	A1064T, A2027T	E355V, E676V	*Cul4a*	chr13:113909435	*CUL4A*	ns	E355V, E676V	1.00	D	0.97	D	0.998	C	3	0	1	0	yes	yes
chr8:46210849	ns	7.078	C74G	A25G	*Slc25a4*	chr4:186064600	*SLC25A4*	ns	A25G	0.97	D	0.25	P	0.999	C	3	3	2	0	yes	yes
chr8:72176337	stop	0.029	T429G	Y143X	*Rab8a*	chr19:16238850	*RAB8A*	stop	Y143X	0.83	NA	0.59	NA	0.119	N	3	0	0	0	yes	yes
chr8:125464212	ns	6.143	G3038A	R1013Q	*Sipa1l2*	chr1:232596693	*SIPA1L2*	ns	R86H, R1012H	1.00	D	1.00	D	0.999	C	3	0	0	0	yes	yes
chr9:54957703	ns	6.873	G700A	E234K	*Psma4*	chr15:78841200	*PSMA4*	ns	E210K, E163K	0.49	T	0.04	B	1.000	C	3	0	0	0	yes	no
chr9:120950603	ns	7.457	A95T	D32V	*Ctnnb1*	chr3:41266098 rs121913396	*CTNNB1*	ns	D32V	1.00	D	1.00	D	0.999	C	4	0	0	0	yes	no
chr9:120950609	ns	7.446	G101A	G34E	*Ctnnb1*	chr3:41266104 rs28931589	*CTNNB1*	ns	G34E	1.00	D	1.00	D	1.000	C	8	0	0	0	yes	no
chr10:5644006	ns	0.697	C358T	P120S	*Vip*	chr6:153077288	*VIP*	ns	P119S	0.26	T	0.06	B	0.938	N	3	0	0	0	yes	no
chr10:80393944	ns	6.189	A626G, A530G	N209S, N177S	*Mbd3*	chr19:1581142	*MBD3*	ns	N209S	0.99	D	1.00	D	0.996	C	4	0	0	0	yes	yes
chr10:128922672	ns	4.258	T116A	V39E	*Bloc1s1*	chr12:56110771	*BLOC1S1*	ns	V39E, V67E	0.99	D	0.42	P	0.999	C	4	0	0	0	yes	yes
chr11:69103193	ns	2.726	G1324A, G1264A	A442T, A422T	*Per1*	chr17:8051056	*PER1*	ns	A442T	0.88	T	0.02	B	0.999	C	3	0	0	0	yes	no
chr11:70229108	ns	5.950	A52C	T18P	*Bcl6b*	chr17:6927042	*BCL6B*	ns	T18P	0.78	T	0.01	B	0.999	C	8	9	3	0	yes	no
chr11:83529163	ns	3.078	G181A	A61T	*Ccl5*	chr17:34205539	*CCL5*	ns	A61T	0.99	D	0.87	D	0.999	C	3	0	0	0	yes	yes

Mouse						Human									# affected samples				Method	
Position	SNV type	PhyloP	Nt change	Aa change	Gene	Position	Gene	SNV type	Aa change	Sift score	Sift prediction	PolyPhen2 score	PolyPhen2 prediction	PhyloP score (rescaled by dbNSFP)	PhyloP prediction	AOM/DSS tumor	AOM/DSS proximal	DSS	Control	Method 1	Method 2
chr13:67621393	ns	0.184	C716T	A239V	*A530054K11Rik*	chr19:22157042 rs377709801	*ZNF208*	ns	A265V							4	1	0	0	yes	no
chr14:54617795	ns	5.349	A197T	K66M	*Psmb5*	chr14:23503894	*PSMB5*	ns	K66M	0.88	NA	0.61	P	0.997	C	3	0	0	0	yes	yes
chr15:10575303	ns	0.257	G1655A	R552Q	*Rai14*	chr5:34823602	*RAI14*	syn	K523K							4	2	0	0	yes	no
chr15:89315104	ns	2.485	G13C	A5P	*Sbf1*	chr22:50913257	*SBF1*	ns	A5P	0.99	D	0.96	D	0.991	C	15	2	3	0	yes	yes
chr15:101369849	ns	5.802	A206G	D69G	*Krt80*	chr12:52585481	*KRT80*	ns	D69G	0.99	D	0.97	D	0.998	C	3	1	1	0	yes	yes
chr16:97748032	ns	3.379	A823T	T275S	*Ripk4*	chr21:43166782	*RIPK4*	ns	T275S	0.53	T	0.00	B	0.965	C	3	2	0	0	yes	no
chr17:5890126	ns	2.934	C151T	L51F	*Snx9*	chr6:158294185	*SNX9*	syn	L51L							3	1	0	0	yes	no
chr17:48432513	ns	1.549	C97A	L33M	*Apobec2*	chr6:41021183	*APOBEC2*	ns	L33M	0.75	T	0.00	B	0.981	C	3	0	0	0	yes	no
chr17:94876634	ns	0.267	A712T	T238S	*LOC100044193*	no liftover possible										9	8	3	0	yes	no
chr18:44812159	ns	4.049	C5T	A2V	*Mcc*	chr5:112824104	*MCC*	ns	A3V	0.78	NA	0.41	NA	0.975	C	3	0	1	0	yes	no
chr19:53310901	ns	4.827	A202G	T68A	*Mxi1*	chr10:111967768	*MXI1*	ns	T68A	0.97	D	0.99	D	0.997	C	3	0	0	0	yes	yes
chrX:38563683	ns	7.295	T539C, T761C	L180S, L254S	*Cul4b*	chrX:119693958	*CUL4B*	ns	L179S	1.00	D	0.97	D	0.998	C	4	0	0	0	yes	yes
chrX:56789960	ns	2.956	G415C, G463C, G205C	G139R, G155R, G69R	*Fhl1*	chrX:135290034	*FHL1*	ns	G139R, G155R	0.85	T	0.63	P	0.999	C	3	0	0	0	yes	yes
chrX:106107280	ns	3.179	C3184T, C3181T	R1062C, R1061C	*Atp7a*	chrX:77286994	*ATP7A*	ns	H1070Y	0.97	D	0.16	P	0.958	C	3	0	0	0	yes	yes

Table S 29 Novel SNVs

Method 1 includes all SNVs, which were not annotated in dbSNP, missense, and produce either a stop codon or were at a conserved position with positive PhyloP score. Method 2 includes all missense SNVs, for which the corresponding human position is conserved, not annotated in dbSNP, and one of the damaging prediction tools (SIFT or PolyPhen2) preestimated an influence on the protein function. For both methods, the SNVs were shared by at least three tumor samples and did not exist in any of the controls. The SNV type 'ns' stands for non-synonymous and 'stop' for stopgain.

Position	Ref.	Observ.	Type	Gene	Aa pos	# AOM/DSS tumor	# AOM/DSS proximal	# DSS	# control
chr1:125572486-125572486	T	-	fs deletion	Slc35f5	F246fs	9	5	0	0
chr1:191354614-191354618	AAGAA	-	fs deletion	Ppp2r5a	p.380_381del	3	1	0	0
chr2:32619664-32619664	T	-	fs deletion	St6galnac6	V378fs	6	2	1	0
chr2:91931373-91931373	T	-	fs deletion	Mdk	K25fs	3	3	3	0
chr3:108825895-108825896	AA	-	fs deletion	Stxbp3a	123_123del	4	2	1	0
chr5:151648294-151648294	T	-	fs deletion	Rfc3	I82fs	6	3	1	0
chr8:109683676-109683676	T	-	fs deletion	2400003C14Rik	T30fs	3	2	3	0
chr11:85822035-85822035	A	-	fs deletion	Bcas3	G56fs	3	2	3	0
chr12:59069203-59069203	T	-	fs deletion	Pnn	P170fs	3	0	0	0
chr14:44717786-44717786	-	A	fs insertion	Gm8267	L199fs	4	0	0	0
chr17:21229254-21229254	A	-	fs deletion	Vmn1r234	L143fs	6	4	1	0
chr17:94747584-94747584	-	A	fs insertion	Mettl4	L142fs	3	2	2	0

Table S 30 Novel frameshift InDels

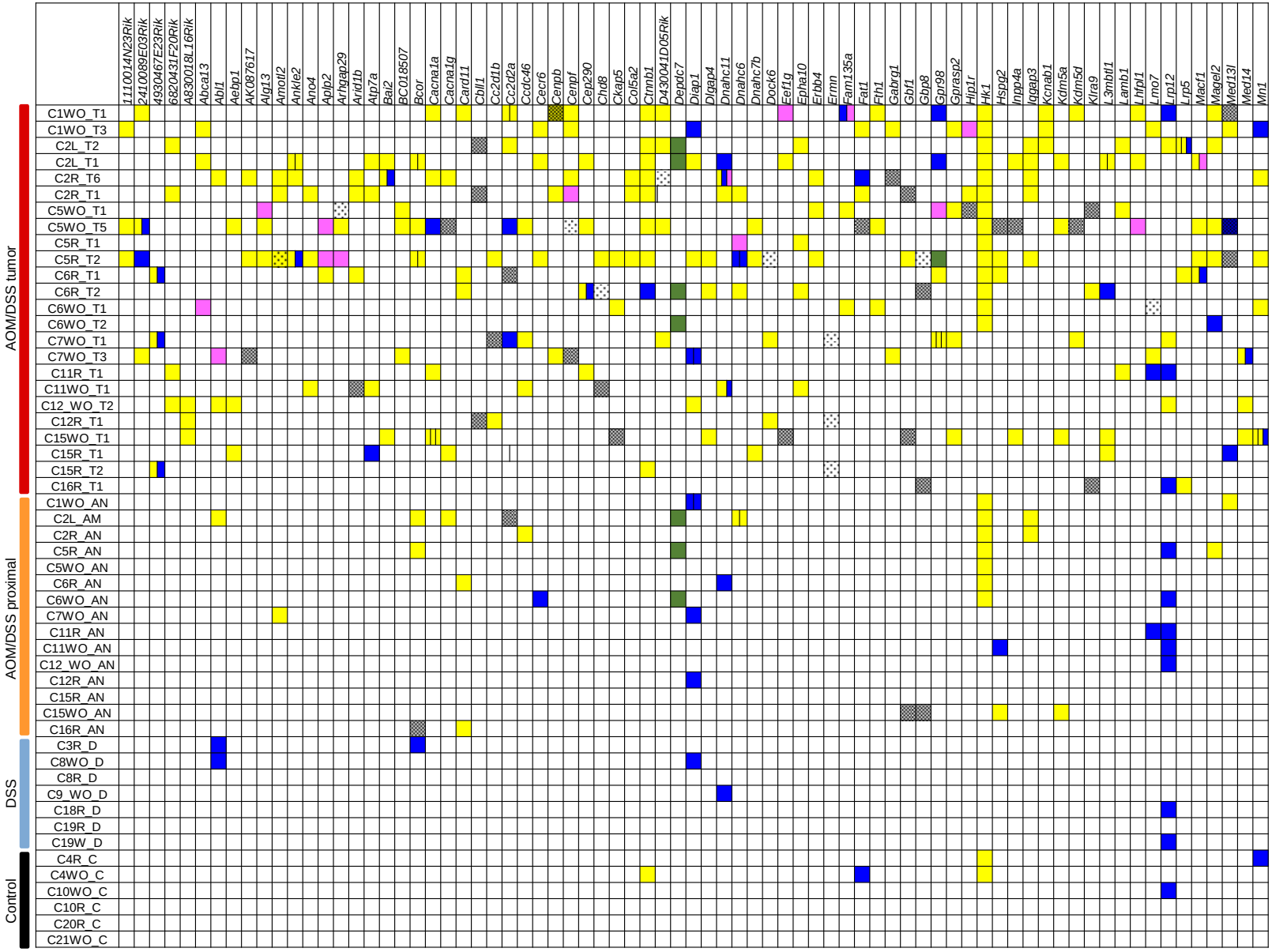
In the table, all InDels were listed, which were not annotated in dbSNP, did not exist in any of the controls, were supported by at least 5% of the reads in at least three tumor samples, and were supported by three or more reads in at least one tumor sample. Ref. = reference allele, Observ. = observed allele.

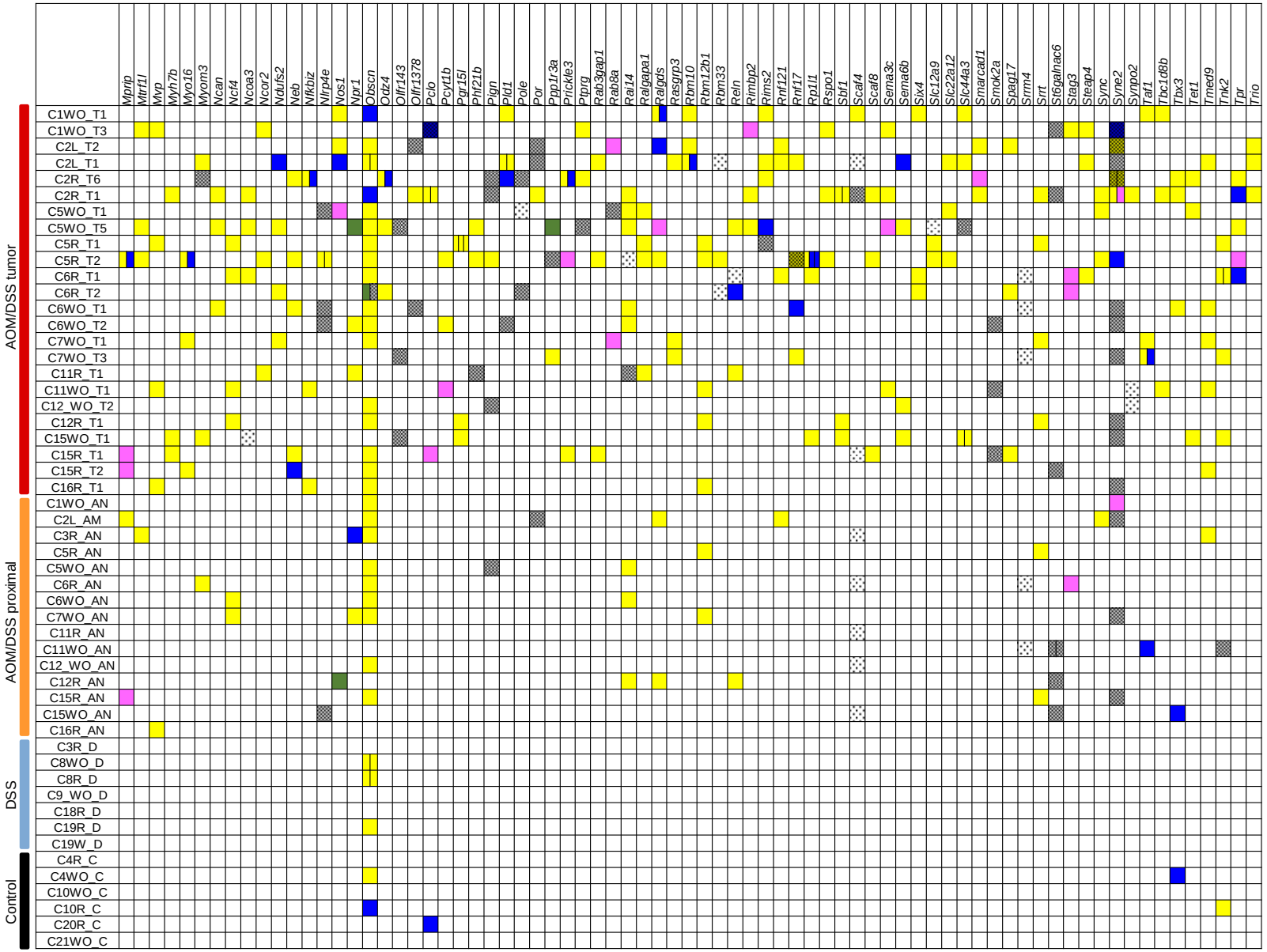
Position	Gene	Primary site (site subtype)
chr2:120720255	*PTPN4*	Endometrium, thyroid
chr2:231256910	*SP140L*	Central nervous system
chr3:149677870	*RNF13*	Lung
chr3:41266098	*CTNNB1*	Central nervous system (brain, brainstem, cerebellum), cervix, endometrium, large intestine, liver, lung, ovary, pancreas, pituitary (craniopharyngeal duct), prostate, skin (arm, forearm), soft tissue (fibrous tissue and uncertain origin), stomach, testis, urinary tract (bladder)
chr3:41266101	*CTNNB1*	Biliary tract (galbladder), bone (mandible, maxilla), breast, central nervous system (brain, brainstem, cerebellum, medulla, posterior fossa), endometrium, haematopoietic and lymphoid tissue (lymph node), kidney, large intestine (right, left, colon, rectum), liver, lung, esophagus, ovary, pancreas, parathyroid, pituitary (craniopharyngeal duct), prostate, skin (arm, back, eye, face, forearm, neck, scalp), soft tissue (fibrous tissue and uncertain origin), stomach, testis, thyroid, urinary tract (bladder)
chr3:41266104	*CTNNB1*	Biliary tract (gallbladder, bone (maxilla), breast, central nervous system (brain, cerebellum, parietal lobe, supratentorial), endometrium, large intestine (caecum, colon, rectum, right), liver, lung, ovary, pancreas, pituitary (craniopharyngeal duct), prostate, skin (face, upper leg), stomach, testis
chr4:46067487	*GABRG1*	Lung
chr4:5667334	*EVC2*	Liver
chr5:134870962	*NEUROG1*	Liver
chr8:116632282	*TRPS1*	Urinary tract (bladder)
chr9:133747588	*ABL1*	Large intestine (caecum)
chr10:28228927	*ARMC4*	Lung
chr11:22707267	*GAS2*	Ovary
chr16:58616700	*CNOT1*	Lung
chr17:41477150	*ARL4D*	Ovary
chr17:7576889	*TP53*	Lung, urinary tract
chr20:40709572	*PTPRT*	Bone
chr20:6064774	*FERMT1*	Endometrium
chr2:165365288-165365288	*GRB14*	Breast, liver, large intestine (colon, caecum), ovary
chr4:46060358-46060358	*GABRG1*	Large intestine (colon)
chr4:56336954-56336954	*CLOCK*	Large intestine (colon, caecum, rectum), lung, ovary, pancreas
chr10:98336475-98336475	*TM9SF3*	Large intestine (colon, caecum), lung
chr14:21859176-21859176	*CHD8*	Large intestine (caecum, colon)
chr14:53540568-53540568	*DDHD1*	Large intestine (colon)
chr16:3817721-3817721	*CREBBP*	Large intestine (colon), stomach
chr19:56370249-56370249	*NLRP4*	Esophagus, large intestine (colon)
chr22:50279276-50279276	*ZBED4*	Stomach, large intestine (colon)

Table S 31 COSMIC variants
The table includes all variants, which were predicted as cancer-promoting / driver mutations by the FATHMM algorithm. The upper part of the table describes SNVs, the lower part InDels.

Position	dbSNP	Gene	Variant clinical significance	Variant disease name
chr2:31620584	rs72549369	*XDH*	Pathogenic	Deficiency of xanthine oxidase
chr2:234681059	rs34993780	*UGT1A5*, *UGT1A9*, *UGT1A3*, *UGT1A6*, *UGT1A4*, *UGT1A1*, *UGT1A8*, *UGT1A10*, *UGT1A7*	Pathogenic	Lucey-Driscoll syndrome, Crigler-Najjar syndrome
chr3:41266097	rs28931588	*CTNNB1*	Other	Hepatoblastoma, pilomatricoma
chr4:47593286	rs112565536	*ATP10D*	Untested	Malignant melanoma
chr8:121290770	rs267601752	*COL14A1*	Untested	Malignant melanoma
chr9:34649422	rs367543265	*GALT*	Pathogenic	Deficiency of UDPglucose-hexose-1-phosphate uridylyltransferase
chr9:34649485	rs111033802	*GALT*	Pathogenic	Deficiency of UDPglucose-hexose-1-phosphate uridylyltransferase
chr11:5246839	rs33964352	*HBB*	Other	Hemoglobin kochi, hemoglobin mito
chr11:68204455	rs28939709	*LRP5*	Pathogenic	Exudative vitreoretinopathy 4
chr11:119170256	rs374672276	*CBL*	Unknown	All highly penetrant
chr12:52885521	rs267607465	*KRT6A*	Untested	Not provided
chr16:2135341	rs45517357	*TSC2*	Untested	Tuberous sclerosis syndrome
chr16:4242230	rs267604545	*SRL*	Untested	Malignant melanoma
chr17:7915920	rs56130505	*GUCY2D*	Untested	Not provided
chr19:30193878	rs398122409	*C19orf12*	Pathogenic	Neurodegeneration with brain iron accumulation 4
chr20:33583264	rs267605899	*MYH7B*	Untested	Malignant melanoma
chrX:80185786	rs267606520	intergenic	Untested	Malignant melanoma
chrX:153993753	rs121912292	*DKC1*	Pathogenic	Dyskeratosis congenita X-linked

Table S 32 Clinvar variants
The table includes all SNVs, which were called in at least one tumor sample and were annotated in the Clinvar database.


1110014N23Rik
2410089E03Rik
4930467E23Rik
6820431F20Rik
A830018L16Rik
Abca13
Abl1
Aebp1
AK087617
Alg13
Amotl2
Ankle2
Ano4
Aplp2
Arhgap29
Arid1b
Atp7a
Bai2
BC018507
Bcor
Cacna1a
Cacna1g
Card11
Cbll1
Cc2d1b
Cc2d2a
Ccdc46
Cecr6
Cenpb
Cenpf
Cep290
Chd8
Ckap5
Col5a2
Ctnnb1
D430041D05Rik
Depdc7
Diap1
Dlgap4
Dnahc11
Dnahc6
Dnahc7b
Dock6
Eef1g
Epha10
Erbb4
Ermn
Fam135a
Fat1
Fth1
Gabrg1
Gbf1
Gbp8
Gpr98
Gprasp2
Hip1r
Hk1
Hspg2
Inpp4a
Iqgap3
Kcnab1
Kdm5a
Kdm5d
Klra9
L3mbtl1
Lamb1
Lhfpl1
Lmo7
Lrp12
Lrp5
Macf1
Magel2
Med13l
Med14
Mn1
C1WO_T1
C1WO_T3
C2L_T2
C2L_T1
C2R_T6
C2R_T1
C5WO_T1
C5WO_T5
C5R_T1
C5R_T2
C6R_T1
C6R_T2
C6WO_T1
C6WO_T2
C7WO_T1
C7WO_T3
C11R_T1
C11WO_T1
C12_WO_T2
C12R_T1
C15WO_T1
C15R_T1
C15R_T2
C16R_T1
C1WO_AN
C2L_AM
C2R_AN
C5R_AN
C5WO_AN
C6R_AN
C6WO_AN
C7WO_AN
C11R_AN
C11WO_AN
C12_WO_AN
C12R_AN
C15R_AN
C15WO_AN
C16R_AN
C3R_D
C8WO_D
C8R_D
C9_WO_D
C18R_D
C19R_D
C19W_D
C4R_C
C4WO_C
C10WO_C
C10R_C
C20R_C
C21WO_C
AOM/DSS tumor
AOM/DSS proximal
DSS
Control


Mprip
Mtrf1l
Myp
Myh7b
Myo16
Myom3
Ncan
Ncf4
Ncoa3
Ncor2
Ndufs2
Neb
Nfkbiz
Nlrp4e
Nos1
Npr1
Obscn
Odz4
Olfr143
Olfr1378
Pclo
Pcyt1b
Pgr15l
Phf21b
Pign
Pld1
Pole
Por
Ppp1r3a
Prickle3
Ptprg
Rab39qap1
Rab8a
Rai14
Ralgapa1
Ralgds
Rasgrp3
Rbm10
Rbm12b1
Rbm33
Reln
Rimbp2
Rims2
Rnf121
Rnf17
Rp1l1
Rspo1
Sbf1
Scaf4
Scaf8
Sema3c
Sema6b
Six4
Slc12a9
Slc22a12
Slc44a3
Smarcad1
Smok2a
Spag17
Srrm4
Srrt
St6galnac6
Stag3
Steap4
Sync
Syne2
Synpo2
Taf1
Tbc1d8b
Tbx3
Tet1
Tmed9
Tnk2
Tpr
Trio
AOM/DSS tumor
AOM/DSS proximal
DSS
Control
C1WO_T1
C1WO_T3
C2L_T2
C2L_T1
C2R_T6
C2R_T1
C5WO_T1
C5WO_T5
C5R_T1
C5R_T2
C6R_T1
C6R_T2
C6WO_T1
C6WO_T2
C7WO_T1
C7WO_T3
C11R_T1
C11WO_T1
C12_WO_T2
C12R_T1
C15WO_T1
C15R_T1
C15R_T2
C16R_T1
C1WO_AN
C2L_AM
C3R_AN
C5R_AN
C5WO_AN
C6R_AN
C6WO_AN
C7WO_AN
C11R_AN
C11WO_AN
C12_WO_AN
C12R_AN
C15R_AN
C15WO_AN
C16R_AN
C3R_D
C8WO_D
C8R_D
C9_WO_D
C18R_D
C19R_D
C19W_D
C4R_C
C4WO_C
C10WO_C
C10R_C
C20R_C
C21WO_C

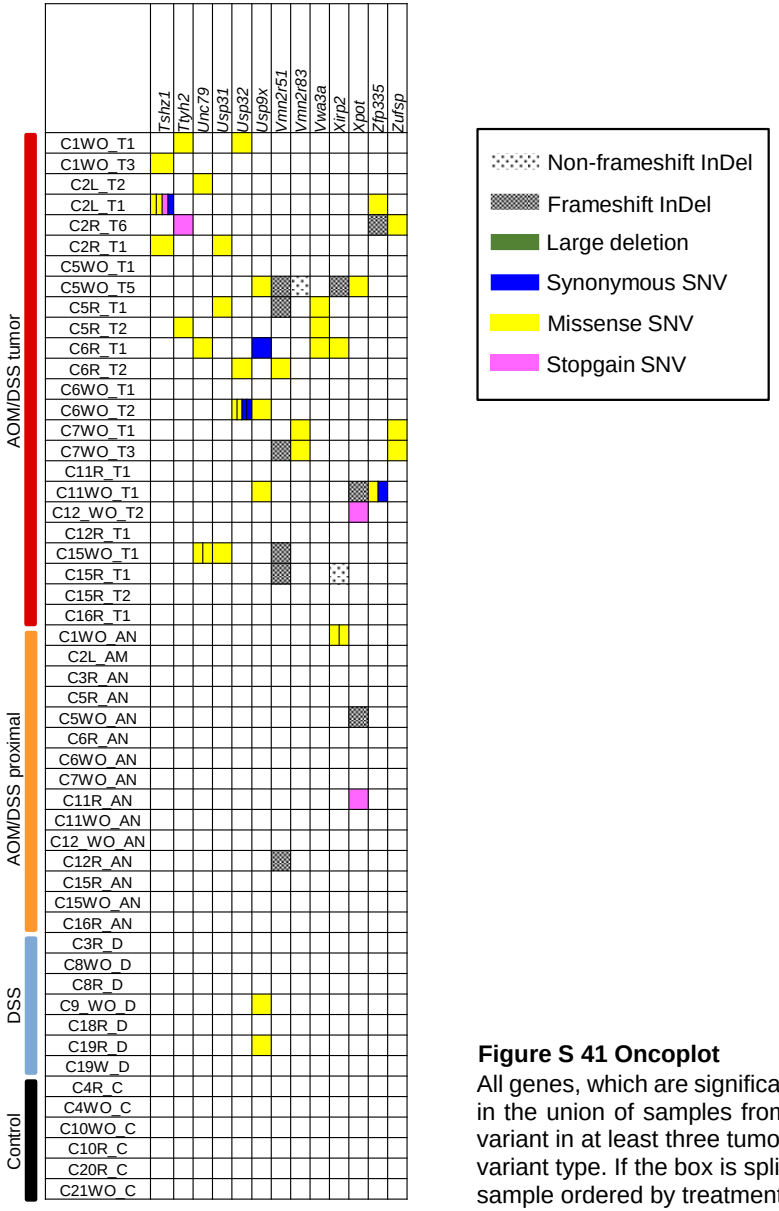


Figure S 41 Oncoplot

All genes, which are significantly more often affected by at least one non-synonymous variant in the tumor samples than in the union of samples from control and DSS treated mice, are shown in oncoplot. Furthermore all genes having a variant in at least three tumors and in none of the controls are included in the plot. The colors of the boxes indicate the variant type. If the box is split, the appropriate number of variants were called in that sample. Each row represents one sample ordered by treatment and DSS dose set. Each column stands for one gene. They are listed alphabetically.

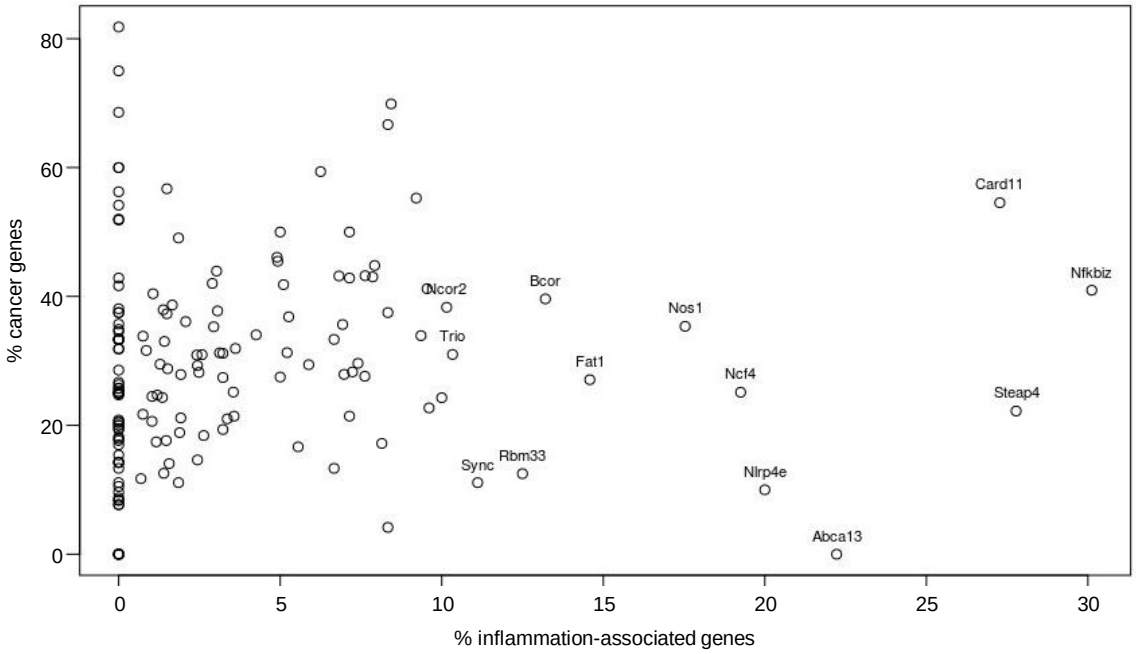


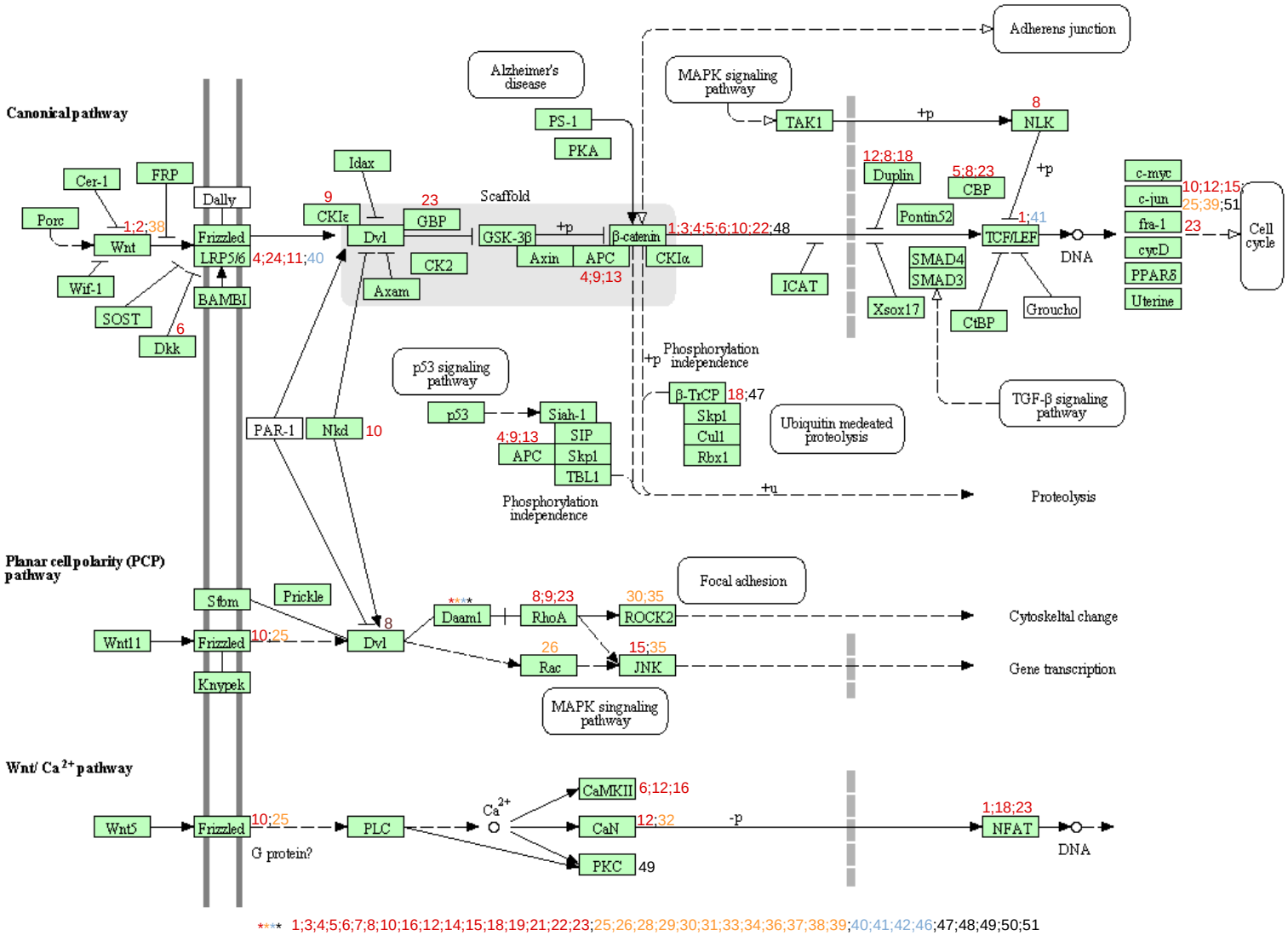
Figure S 42 Percentage of inflammation- or cancer-associated genes of directly connected genes
Nfκbiζ is especially promising due to the high percentage of connected inflammation-associated genes. While 30.1% of the directly connected genes of *Nfκbiζ* were involved in inflammatory processes, the mean for all other affected genes was only 3.7%. The percentage of directly connected, cancer-associated genes was for *Nfκbiζ* 41.0% compared with an average of 29.5%.

6.2.2.6 Mutated processes in AOM/DSS-triggered colorectal cancer

ID	Term	Group	P-value
GO:0003231	Cardiac ventricle development	2	0.049
GO:2000008	Regulation of protein localization to cell surface	3	0.039
GO:0051588	Regulation of neurotransmitter transport	4	0.031
GO:0061550	Cranial ganglion development	5	0.029
GO:0034720	Histone H3-K4 demethylation	6	0.039
GO:0005488	Binding	7	0.00011
GO:0005085	Guanyl-nucleotide exchange factor activity	7	0.0028
GO:0016192	Vesicle-mediated transport	8	0.0098
GO:0051650	Establishment of vesicle localization	8	0.017
GO:0060627	Regulation of vesicle-mediated transport	8	0.017
GO:0017156	Calcium ion regulated exocytosis	8	0.022
GO:0017157	Regulation of exocytosis	8	0.031
GO:0017158	Regulation of calcium ion-dependent exocytosis	8	0.034
GO:0051641	Cellular localization	8	0.035
GO:0016079	Synaptic vesicle exocytosis	8	0.039
GO:0099536	Synaptic signaling	10	0.018
GO:0070988	Demethylation	12	0.023
GO:0140034	Methylation-dependent protein binding	14	0.023
GO:0003958	NADPH-hemoprotein reductase activity	15	0.035
GO:0042733	Embryonic digit morphogenesis	16	0.037
GO:1904491	Protein localization to ciliary transition zone	17	0.022
GO:0035418	Protein localization to synapse	18	0.036
GO:0000988	Transcription factor activity, protein binding	20	0.028

ID	Term	Group	P-value
GO:0044093	Positive regulation of molecular function	21	0.022
GO:0050790	Regulation of catalytic activity	21	0.047
GO:0015677	Copper ion import	22	0.022
GO:0048200	Golgi transport vesicle coating	23	0.0066
GO:0035964	COPI-coated vesicle budding	23	0.013
GO:0032453	Histone demethylase activity (H3-K4 specific)	24	0.039
GO:0071840	Cellular component organization or biogenesis	25	0.0017
GO:0051128	Regulation of cellular component organization	25	0.0053
GO:0007275	Multicellular organism development	25	0.0056
GO:0022008	Neurogenesis	25	0.008
GO:0060788	Ectodermal placode formation	25	0.012
GO:0032502	Developmental process	25	0.013
GO:0043412	Macromolecule modification	25	0.017
GO:0036211	Protein modification process	25	0.017
GO:0006464	Cellular protein modification process	25	0.017
GO:0008347	Glial cell migration	25	0.018
GO:0048468	Cell development	25	0.019
GO:0021795	Cerebral cortex cell migration	25	0.021
GO:0061196	Fungiform papilla development	25	0.022
GO:0007389	Pattern specification process	25	0.026
GO:0070925	Organelle assembly	25	0.035
GO:0043587	Tongue morphogenesis	25	0.035
GO:0031060	Regulation of histone methylation	25	0.037
GO:0061180	Mammary gland epithelium development	25	0.046
GO:0035791	Platelet-derived growth factor receptor-beta signaling pathway	26	0.047
GO:0007264	Small GTPase mediated signal transduction	27	0.034
GO:0065008	Regulation of biological quality	28	0.047
GO:0006007	Glucose catabolic process	31	0.039

Table S 33 GO terms enriched for genes with somatic mutation in tumor samples
Enrichment analysis was performed with all genes, which were more often affected by a non-synonymous SNVs or InDel in the tumors than in the pool of DSS and control samples [genes shown in the oncoplot (Figure S 41)]. Exclusively GO terms of the most general hierarchy level are reported.



**** 1;3;4;5;6;7;8;10;16;12;14;15;18;19;21;22;23;25;26;28;29;30;31;33;34;36;37;38;39;40;41;42;46;47;48;49;50;51

	Sample ID	Sample name
AOM/DSS tumor	1	C1WO_T1
	2	C1WO_T3
	3	C2L_T1
	4	C2L_T2
	5	C2R_T1
	6	C2R_T6
	7	C5R_T1
	8	C5R_T2
	9	C5WO_T1
	10	C5WO_T5
	11	C6R_T1
	12	C6R_T2
	13	C6WO_T1
	14	C6WO_T2
	15	C7WO_T1
	16	C7WO_T3
	17	C11R_T1
	18	C11WO_T1
	19	C12R_T1
	20	C12WO_T2
	21	C15R_T1
	22	C15R_T2
	23	C15WO_T1
	24	C16R_T1
AOM/DSS proximal	25	C1WO_AN
	26	C2L_AN
	27	C2R_AN
	28	C5R_AN
	29	C5WO_AN
	30	C6R_AN
	31	C6WO_AN
	32	C7WO_AN
	33	C11R_AN
	34	C11WO_AN
	35	C12R_AN
	36	C12WO_AN
	37	C15R_AN
	38	C15WO_AN
	39	C16R_AN
DSS	40	C3R_D
	41	C8R_D
	42	C8WO_D
	43	C9WO_D
	44	C18R_D
	45	C19R_D
	46	C19WO_D
Control	47	C4R_C
	48	C4WO_C
	49	C10R_C
	50	C10WO_C
	51	C20R_C
	52	C21WO_C

Figure S 43 Genes with somatic mutation in the Wnt signaling pathway
The figure was received from the KEGG database (https://www.genome.jp/kegg) and modified afterwards. If a gene was affected by at least one non-synonymous SNV or InDel in one of the samples, the sample ID was added to the gene label, respectively. The color of the sample ID indicates the treatment group. The assignment of sample IDs to sample names is shown on the right site. The samples C11R_T1, C12WO_T2, C2R_AN, C9WO_D, C18R_D, C19R_D, and C21WO_C were not affected by an SNV or InDel in any of the genes in the Wnt signaling pathway.

6.2.2.7 Transcriptome sequencing results

Sample	Treatment	Set	Run / lane / library	# raw reads	# filtered reads	Mean read length (filtered)	Mean base quality (filtered)	# mapped (%mapped)	% MQ=50	# duplicates (%duplicates)	% rRNA	% exonic (CDS/UTR)	% intronic	% intergenic	% s/as noise	% expressed genes
C1WO_AN	AOM/DSS proximal	1	2 / 1 / 1	70,801,908	63,707,926	98.4	36.1	43,582,691 (68.4)	87.5	20,885,821 (47.9)	32.8	27.8	26.1	13.3	0.9	56.0
C1WO_T2	AOM/DSS tumor	1	1 / 7 / 6	210,448,636	194,958,544	99.6	36.8	178,645,411 (91.6)	89.2	52,617,792 (29.5)	12.4	39.6	34.5	13.5	1.0	62.3
C2L_AN	AOM/DSS proximal	1	2 / 1 / 1	82,688,268	75,114,240	98.7	36.2	59,686,556 (79.5)	89.4	24,077,959 (40.3)	26.6	32.6	29.1	11.7	0.9	58.5
C2L_T1	AOM/DSS tumor	1	2 / 1 / 1	74,402,814	67,864,962	98.7	36.1	59,831,318 (88.2)	85.9	25,957,970 (43.4)	12.8	39.1	31.9	16.2	0.9	57.6
C2R_AN	AOM/DSS proximal	1	2 / 1 / 1	79,735,438	72,706,958	98.8	36.2	62,308,673 (85.7)	88.5	27,908,964 (44.8)	27.3	31.4	28.5	12.8	1.0	58.0
C2R_T5	AOM/DSS tumor	1	2 / 1 / 1	81,382,856	74,149,416	98.7	36.2	61,155,739 (82.5)	88.2	20,607,548 (33.7)	17.4	34.9	33.6	14.1	1.0	59.3
C3R_D	DSS	1	2 / 1 / 1	64,522,126	58,762,276	98.5	36.0	45,927,918 (78.2)	82.2	30,630,976 (66.7)	29.1	29.1	24.0	17.8	0.7	55.0
C4R_C	Control	1	4 / 1 / 2	84,948,846	75,326,016	98.5	35.6	51,282,141 (68.1)	86.9	20,350,359 (39.7)	24.8	29.9	30.7	14.5	0.9	57.2
C4WO_C	Control	1	4 / 1 / 2	74,208,598	65,366,566	98.1	35.6	45,628,957 (69.8)	84.8	19,809,792 (43.4)	26.9	27.0	28.6	17.5	1.2	57.4
C5R_AN	AOM/DSS proximal	2	3 / 2 / 4	80,441,596	72,530,412	99.2	36.3	63,310,848 (87.3)	86.6	27,491,943 (43.4)	26.9	35.9	22.4	14.8	0.7	58.4
C5R_T3	AOM/DSS tumor	2	4 / 1 / 3	87,995,464	78,171,196	99.0	35.9	65,586,025 (83.9)	86.6	21,051,434 (32.1)	15.8	38.3	30.8	15.1	0.8	59.2
C5WO_AN	AOM/DSS proximal	2	3 / 2 / 4	80,979,154	71,220,962	98.9	36.1	61,708,343 (86.6)	82.5	28,620,625 (46.4)	21.7	38.1	21.5	18.7	0.7	59.0
C5WO_T2	AOM/DSS tumor	2	4 / 1 / 2	83,888,018	73,344,368	98.4	35.8	58,312,854 (79.5)	86.8	19,779,633 (33.9)	18.5	30.4	34.5	16.6	1.1	59.5
C6R_AN	AOM/DSS proximal	2	3 / 2 / 4	80,973,088	71,884,672	99.0	36.2	60,124,386 (83.6)	88.4	23,682,952 (39.4)	25.0	34.0	27.5	13.4	0.9	58.9
C6R_T3	AOM/DSS tumor	2	3 / 3 / 3	87,257,966	77,228,084	99.1	36.3	65,180,419 (84.4)	87.9	18,388,272 (28.2)	14.4	36.1	35.3	14.2	0.9	59.5
C6WO_AN	AOM/DSS proximal	2	4 / 1 / 2	86,505,906	76,754,894	99.0	36.0	68,237,827 (88.9)	85.7	30,699,098 (45.0)	27.1	31.5	25.8	15.7	0.8	58.6
C6WO_T1	AOM/DSS tumor	2	4 / 1 / 3	80,415,248	71,415,430	99.0	35.9	59,548,876 (83.4)	86.6	19,137,978 (32.1)	14.5	39.4	29.8	16.3	0.9	59.9
C7WO_AN	AOM/DSS proximal	2	3 / 2 / 4	81,039,454	71,905,156	99.2	36.3	64,223,052 (89.3)	87.9	29,961,605 (46.7)	31.9	31.9	23.0	13.3	0.8	58.6
C7WO_T1	AOM/DSS tumor	2	3 / 1 / 3	91,750,070	87,054,362	99.7	37.1	76,206,352 (87.5)	86.7	26,840,167 (35.2)	19.2	33.6	31.4	15.7	1.0	59.4
C8WO_D	DSS	2	3 / 2 / 4	79,173,528	70,493,758	99.2	36.3	62,523,483 (88.7)	87.2	26,270,455 (42.0)	25.4	36.9	23.2	14.5	0.8	59.5

Sample	Treatment	Set	Run / lane / library	# raw reads	# filtered reads	Mean read length (filtered)	Mean base quality (filtered)	# mapped (%mapped)	% MQ=50	# duplicates (%duplicates)	% rRNA	% exonic (CDS/UTR)	% intronic	% intergenic	% s/as noise	% expressed genes
C9WO_D	DSS	2	3 / 3 / 3	72,008,288	63,564,392	99.1	36.3	55,218,366 (86.9)	88.5	21,047,716 (38.1)	24.6	30.1	31.8	13.5	1.0	58.8
C10R_C	Control	2	3 / 1 / 4	99,466,560	94,614,118	99.9	37.3	84,490,170 (89.3)	89.7	26,452,724 (31.3)	18.4	31.8	36.4	13.4	1.1	60.5
C10WO_C	Control	2	3 / 2 / 4	64,448,716	57,491,068	99.1	36.3	51,954,055 (90.4)	87.8	21,957,218 (42.3)	26.2	30.6	29.2	13.9	1.0	57.3
C11WO_AN	AOM/DSS proximal	3	3 / 3 / 4	80,127,746	70,951,730	99.4	36.4	61,702,091 (87.0)	87.3	31,356,431 (50.8)	35.2	31.5	20.0	13.4	0.8	58.3
C11WO_T2	AOM/DSS tumor	3	4 / 2 / 3	68,363,202	60,242,230	99.1	35.9	49,495,261 (82.2)	87.0	19,159,367 (38.7)	17.9	34.8	32.0	15.3	0.9	56.4
C12WO_AN	AOM/DSS proximal	3	4 / 2 / 2	86,206,406	75,927,222	98.7	35.8	64,853,364 (85.4)	86.4	30,193,322 (46.6)	30.4	30.2	24.9	14.4	0.8	57.8
C12WO_T1	AOM/DSS tumor	3	4 / 2 / 3	91,311,128	80,049,072	99.0	35.8	67,821,700 (84.7)	85.9	26,181,239 (38.6)	21.5	35.5	26.4	16.6	0.8	59.9
C15R_AN	AOM/DSS proximal	3	3 / 3 / 4	73,494,618	64,772,740	98.8	36.1	54,155,206 (83.6)	84.5	23,565,870 (43.5)	26.1	32.4	24.9	16.6	0.8	58.2
C15R_T3	AOM/DSS tumor	3	4 / 2 / 3	81,746,424	70,970,344	98.3	35.6	58,197,265 (82.0)	87.1	19,937,096 (34.6)	19.7	30.4	35.1	14.8	1.1	58.2
C18R_D	DSS	3	3 / 3 / 4	89,646,658	78,994,492	99.2	36.2	68,607,236 (86.9)	87.0	32,985,111 (48.1)	31.6	36.7	18.1	13.5	0.6	59.3
C19R_D	DSS	3	4 / 2 / 2	91,163,174	80,250,702	99.0	35.8	69,742,293 (86.9)	86.5	31,080,438 (44.6)	27.2	35.1	23.0	14.7	0.7	58.4
C20R_C	Control	3	3 / 3 / 4	67,686,500	59,671,836	99.2	36.3	52,096,107 (87.3)	87.4	22,232,701 (42.7)	27.1	35.3	23.4	14.2	0.8	58.2
C21WO_C	Control	3	4 / 2 / 2	99,747,638	87,626,650	98.8	35.7	73,815,741 (84.2)	87.2	29,848,800 (40.4)	25.4	31.8	29.3	13.5	0.8	58.9

Table S 34 Sequencing and mapping statistics of murine transcriptome samples
The first part of each sample name refers to the mouse ID, while the second part refers to the sample type (AN = formerly inflamed tissue from the proximal colon part from an AOM/DSS-treated mouse, Tx = Tumor tissue from an AOM/DSS-treated mouse, C = proximal colon tissue from an untreated control mouse, D = formerly inflamed tissue from the proximal colon part from a DSS-treated mouse.

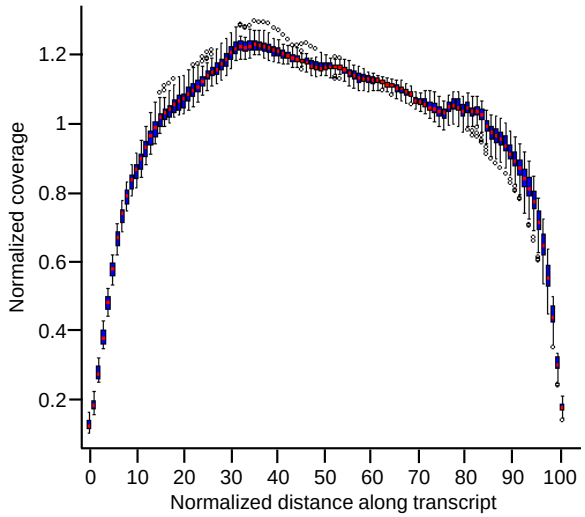


Figure S 44 Test for systematic 5'- or 3'-bias

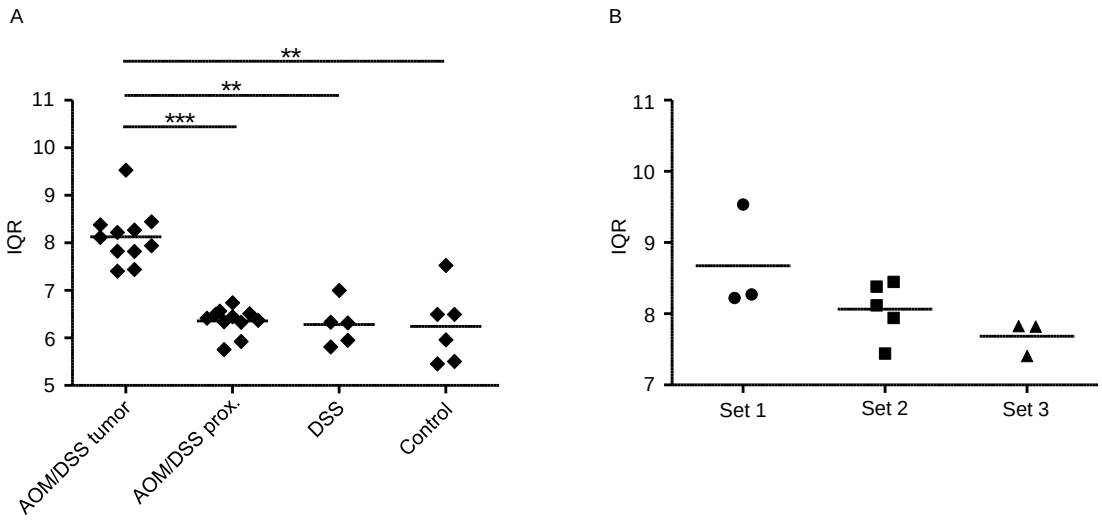


Figure S 45 Comparison of the IQRs of FPKM values

(A) Comparison of IQRs of expression values between the four treatment groups. (B) Comparison of the IQRs of expression values between the three tumor sample sets: The first set harbored samples from mice treated with high DSS dose, while the second and third set contained samples from mice treated with medium and low DSS dose, respectively.

GO term ID	GO term name	GO type	GO group	Adj. p-value
GO:0008152	Metabolic process	BP	18	8.56E-78
GO:0071840	Cellular component organization or biogenesis	BP	18	5.29E-69
GO:0051179	Localization	BP	18	5.21E-57
GO:0009987	Cellular process	BP	18	3.47E-37
GO:0032502	Developmental process	BP	18	1.50E-33
GO:0007275	Multicellular organism development	BP	18	1.80E-33
GO:0070887	Cellular response to chemical stimulus	BP	18	4.85E-29
GO:0051239	Regulation of multicellular organismal process	BP	18	3.10E-25
GO:1901700	Response to oxygen-containing compound	BP	18	4.42E-24
GO:0009719	Response to endogenous stimulus	BP	18	5.91E-22
GO:0035556	Intracellular signal transduction	BP	18	1.39E-21
GO:0042493	Response to drug	BP	18	3.26E-20
GO:0010033	Response to organic substance	BP	18	1.14E-19
GO:1901698	Response to nitrogen compound	BP	18	1.11E-15
GO:0009628	Response to abiotic stimulus	BP	18	1.79E-14
GO:0048583	Regulation of response to stimulus	BP	18	1.87E-14
GO:0006950	Response to stress	BP	18	3.57E-12
GO:0007267	Cell-cell signaling	BP	18	1.42E-11
GO:0010647	Positive regulation of cell communication	BP	18	7.47E-11
GO:0007166	Cell surface receptor signaling pathway	BP	18	7.05E-10
GO:0010035	Response to inorganic substance	BP	18	1.82E-09
GO:0008283	Cell proliferation	BP	18	2.65E-09
GO:0040007	Growth	BP	18	1.36E-08
GO:0040011	Locomotion	BP	18	6.75E-08
GO:0065007	Biological regulation	BP	18	1.89E-07
GO:0009968	Negative regulation of signal transduction	BP	18	2.71E-07
GO:0009636	Response to toxic substance	BP	18	5.57E-07
GO:0007584	Response to nutrient	BP	18	8.04E-07
GO:0022610	Biological adhesion	BP	18	1.75E-06
GO:0046677	Response to antibiotic	BP	18	2.12E-06
GO:0003013	Circulatory system process	BP	18	2.63E-06
GO:0003012	Muscle system process	BP	18	3.85E-06
GO:0009410	Response to xenobiotic stimulus	BP	18	6.47E-06
GO:0001503	Ossification	BP	18	7.98E-06
GO:0031668	Cellular response to extracellular stimulus	BP	18	1.82E-05
GO:1903578	Regulation of ATP metabolic process	BP	18	2.54E-05
GO:0050789	Regulation of biological process	BP	18	2.25E-04
GO:0044419	Interspecies interaction between organisms	BP	18	2.66E-04
GO:0001101	Response to acid chemical	BP	18	3.65E-04
GO:0009605	Response to external stimulus	BP	18	4.01E-04
GO:0035637	Multicellular organismal signaling	BP	18	4.98E-04
GO:0007610	Behavior	BP	30	3.10E-08
GO:0048536	Spleen development	BP	32	4.08E-04

Table S 35 GO terms harboring more upregulated genes than expected by chance
All upregulated genes with adjusted p-value smaller 0.001 and fold change larger four were considered for the gene set enrichment analysis. Exclusively GO terms of the highest GO hierarchical level and adjusted p-value smaller 0.001 are shown. The terms are ordered by GO group and subsequently by p-value.

GO term ID	GO term name	GO type	GO group	Adj. p-value
GO:0007610	Behavior	BP	1	2.89E-27
GO:0065008	Regulation of biological quality	BP	1	3.02E-25
GO:0007267	Cell-cell signaling	BP	1	4.94E-24
GO:0051179	Localization	BP	1	2.47E-16
GO:0032879	Regulation of localization	BP	1	1.10E-12
GO:0032502	Developmental process	BP	1	1.50E-12
GO:0042493	Response to drug	BP	1	8.24E-11
GO:0051239	Regulation of multicellular organismal process	BP	1	3.81E-09
GO:0050890	Cognition	BP	1	7.03E-09
GO:0032501	Multicellular organismal process	BP	1	4.79E-08
GO:0009719	Response to endogenous stimulus	BP	1	1.41E-07
GO:0050808	Synapse organization	BP	1	2.14E-07
GO:0044281	Small molecule metabolic process	BP	1	1.32E-06
GO:0006629	Lipid metabolic process	BP	1	2.00E-06
GO:1901700	Response to oxygen-containing compound	BP	1	2.70E-06
GO:0051050	Positive regulation of transport	BP	1	4.11E-06
GO:0019233	Sensory perception of pain	BP	1	5.05E-06
GO:0006082	Organic acid metabolic process	BP	1	6.18E-06
GO:0009306	Protein secretion	BP	1	9.69E-06
GO:0007269	Neurotransmitter secretion	BP	1	1.77E-05
GO:0042445	Hormone metabolic process	BP	1	2.91E-05
GO:0009410	Response to xenobiotic stimulus	BP	1	4.63E-05
GO:0009636	Response to toxic substance	BP	1	1.65E-04
GO:0031644	Regulation of neurological system process	BP	1	1.87E-04
GO:0009187	Cyclic nucleotide metabolic process	BP	1	3.15E-04
GO:0010876	Lipid localization	BP	1	3.61E-04
GO:0007200	Phospholipase C-activating G-protein coupled receptor signaling pathway	BP	1	4.76E-04
GO:0010646	Regulation of cell communication	BP	1	5.63E-04
GO:0050708	Regulation of protein secretion	BP	1	6.71E-04
GO:0070887	Cellular response to chemical stimulus	BP	1	8.24E-04
GO:0050905	Neuromuscular process	BP	2	1.08E-04
GO:0007215	Glutamate receptor signaling pathway	BP	3	3.85E-06
GO:0005215	Transporter activity	MF	5	1.18E-32
GO:0030594	Neurotransmitter receptor activity	MF	5	5.64E-07
GO:0008066	Glutamate receptor activity	MF	5	4.11E-06
GO:0015464	Acetylcholine receptor activity	MF	5	2.04E-04
GO:0003824	Catalytic activity	MF	20	7.57E-04
GO:0004089	Carbonate dehydratase activity	MF	22	1.22E-04
GO:0008528	G-protein coupled peptide receptor activity	MF	30	7.59E-05
GO:0001653	Peptide receptor activity	MF	30	1.06E-04
GO:0007187	G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	BP	44	2.79E-12
GO:0003310	Pancreatic A cell differentiation	BP	54	6.32E-04
GO:0046906	Tetrapyrrole binding	MF	61	3.43E-04
GO:0052689	Carboxylic ester hydrolase activity	MF	80	1.83E-04

GO term ID	GO term name	GO type	GO group	Adj. p-value
GO:0030165	PDZ domain binding	MF	82	7.87E-04
GO:0007218	Neuropeptide signaling pathway	BP	96	1.27E-05

Table S 36 GO terms harboring more downregulated genes than expected by chance
All downregulated genes with adjusted p-value smaller 0.001 and fold change larger four were considered for the gene set enrichment analysis. Exclusively GO terms of the highest GO hierarchical level and adjusted p-value smaller 0.001 are shown. The terms are ordered by GO group and subsequently by p-value.

TF class	TFBS matrix	P-value	Adj. p-value
V$MZF1_01	M00083	1.24E-10	3.49E-08
V$MAZR_01	M00491	3.80E-10	5.36E-08
V$PAX4_03	M00378	8.73E-08	8.21E-06
V$MZF1_02	M00084	2.40E-07	1.69E-05
V$TATA_01	M00252	3.26E-07	1.84E-05
V$AP1_C	M00199	1.14E-05	0.000533
V$NFKB_Q6	M00194	1.32E-05	0.000533
V$AP1_01	M00517	2.33E-05	0.000710
V$LMO2COM_01	M00277	2.44E-05	0.000710
V$NFKAPPAB_01	M00054	2.52E-05	0.000710
V$TATA_C	M00216	3.83E-05	0.000981
V$NFKB_C	M00208	5.78E-05	0.00136
V$HEN1_02	M00058	6.42E-05	0.00140
V$HEN1_01	M00068	9.62E-05	0.00194
V$SREBP1_02	M00221	0.000141	0.00266
V$STAT6_02	M00500	0.000165	0.00290
V$AREB6_03	M00414	0.000181	0.00300
V$AP1_Q6	M00174	0.000215	0.00335
V$NFKAPPAB50_01	M00051	0.000226	0.00335
V$SPZ1_01	M00446	0.000312	0.00440
V$BACH1_01	M00495	0.000343	0.00445
V$E47_01	M00002	0.000347	0.00445
V$E47_02	M00071	0.000378	0.00463
V$AP2REP_01	M00468	0.000403	0.00473
V$NF1_Q6	M00193	0.000421	0.00475
V$SP1_01	M00008	0.000497	0.00539
V$TAL1ALPHAE47_01	M00066	0.000624	0.00634
V$ARP1_01	M00155	0.000629	0.00634
V$AREB6_02	M00413	0.000713	0.00693
V$OLF1_01	M00261	0.00113	0.0106
V$AP2GAMMA_01	M00470	0.00131	0.0119
V$MYOD_Q6	M00184	0.00135	0.0119
V$BACH2_01	M00490	0.00156	0.0134
V$RP58_01	M00532	0.00180	0.0149
V$ZIC1_01	M00448	0.00253	0.0204
V$AP2ALPHA_01	M00469	0.00290	0.0227

TF class	TFBS matrix	P-value	Adj. p-value
V$LUN1_01	M00480	0.00333	0.0254
V$TAL1BETAITF2_01	M00070	0.00361	0.0268
V$P53_01	M00034	0.00462	0.0334
V$AP1_Q4	M00188	0.00491	0.0346
V$STAT_01	M00223	0.00617	0.0424

Table S 37 Transfac transcription factor binding site classes regulating more overexpressed genes than expected by chance
TF = transcription factor, TFBS = transcription factor binding site

TF class	TFBS matrix	P-value	Adj. p-value
ZNF263	MA0528.1	1.72E-08	9.98E-06
TBP	MA0108.2	2.83E-07	4.76E-05
INSM1	MA0155.1	3.02E-07	4.76E-05
Myod1	MA0499.1	4.99E-07	4.76E-05
NFKB2	MA0778.1	5.00E-07	4.76E-05
Tcf12	MA0521.1	5.43E-07	4.76E-05
BACH2	MA1101.1	6.80E-07	4.76E-05
Myog	MA0500.1	6.94E-07	4.76E-05
JUNB	MA0490.1	7.40E-07	4.76E-05
FOSL1	MA0477.1	1.59E-06	9.19E-05
ASCL1	MA1100.1	2.27E-06	0.000119
FOS	MA0476.1	2.49E-06	0.000120
FIGLA	MA0820.1	3.74E-06	0.000167
Ascl2	MA0816.1	4.18E-06	0.000173
TCF3	MA0522.2	4.76E-06	0.000183
Bach1::Mafk	MA0591.1	5.64E-06	0.000204
JUN(var.2)	MA0489.1	7.90E-06	0.000269
TCF4	MA0830.1	9.93E-06	0.000319
MZF1	MA0056.1	1.26E-05	0.000371
RELB	MA1117.1	1.30E-05	0.000371
JUND	MA0491.1	1.35E-05	0.000371
KLF4	MA0039.3	1.85E-05	0.000482
ZNF740	MA0753.1	1.92E-05	0.000482
MZF1(var.2)	MA0057.1	2.36E-05	0.000569
FOSL2	MA0478.1	2.82E-05	0.000650
KLF5	MA0599.1	2.92E-05	0.000650
FOSB::JUNB	MA1135.1	3.55E-05	0.000761
FOS::JUNB	MA1134.1	5.05E-05	0.00103
FOSL2::JUNB	MA1138.1	5.13E-05	0.00103
NFKB1	MA0105.4	5.49E-05	0.00106
PLAG1	MA0163.1	6.63E-05	0.00124
Nfe2l2	MA0150.2	7.53E-05	0.00136
FOS::JUND	MA1141.1	9.94E-05	0.00174
FOSL1::JUNB	MA1137.1	0.000107	0.00183

TF class	TFBS matrix	P-value	Adj. p-value
FOSL2::JUND	MA1144.1	0.000114	0.00189
ID4	MA0824.1	0.000142	0.00228
EWSR1-FLI1	MA0149.1	0.000153	0.00239
Klf1	MA0493.1	0.000163	0.00249
TCF7L2	MA0523.1	0.000182	0.00270
SP1	MA0079.3	0.000189	0.00274
FOSL2::JUN	MA1130.1	0.000196	0.00276
NEUROD1	MA1109.1	0.000210	0.00285
NFIX	MA0671.1	0.000212	0.00285
NHLH1	MA0048.2	0.000263	0.00346
TBX15	MA0803.1	0.000270	0.00347
Sox3	MA0514.1	0.000298	0.00375
THAP1	MA0597.1	0.000311	0.00383
TFAP2C	MA0524.2	0.000324	0.00391
FOSL1::JUN	MA1128.1	0.000362	0.00428
HIC2	MA0738.1	0.000383	0.00443
LEF1	MA0768.1	0.000406	0.00461
E2F6	MA0471.1	0.000443	0.00494

Table S 38 Jaspar transcription factor classes regulating more overexpressed genes than expected by chance
TF = transcription factor, TFBS = transcription factor binding site

TF class	TFBS matrix	P-value	Adj. p-value
V$MZF1_01	M00083	3.17E-09	8.95E-07
V$MZF1_02	M00084	9.49E-09	1.29E-06
V$MAZR_01	M00491	1.37E-08	1.29E-06
V$MYOD_Q6	M00184	4.82E-07	3.40E-05
V$PAX4_03	M00378	8.88E-07	5.01E-05
V$CP2_01	M00072	1.07E-06	5.03E-05
V$PPARG_03	M00528	3.39E-06	0.000126
V$ZID_01	M00085	3.57E-06	0.000126
V$NF1_Q6	M00193	6.24E-06	0.000188
V$HEN1_02	M00058	6.67E-06	0.000188
V$AP4_Q5	M00175	9.96E-06	0.000255
V$AP4_Q6	M00176	2.46E-05	0.000578
V$HNF4_01	M00134	3.31E-05	0.000718
V$ZIC2_01	M00449	5.98E-05	0.00120
V$OLF1_01	M00261	9.99E-05	0.00183
V$HNF4_01_B	M00411	0.000104	0.00183
V$HEN1_01	M00068	0.000112	0.00187
V$ZIC1_01	M00448	0.000131	0.00206
V$LMO2COM_01	M00277	0.000164	0.00232
V$SP1_01	M00008	0.000164	0.00232
V$NRSF_01	M00256	0.000175	0.00235
V$AREB6_03	M00414	0.000195	0.00250

TF class	TFBS matrix	P-value	Adj. p-value
V$RREB1_01	M00257	0.000237	0.00287
V$ZIC3_01	M00450	0.000244	0.00287
V$SP1_Q6	M00196	0.000327	0.00368
V$E47_01	M00002	0.000537	0.00582
V$MYOGNF1_01	M00056	0.000793	0.00814
V$AP2ALPHA_01	M00469	0.000809	0.00814
V$SPZ1_01	M00446	0.000915	0.00890
V$AP2GAMMA_01	M00470	0.000998	0.00938
V$NGFIC_01	M00244	0.00124	0.0112
V$PPARG_01	M00512	0.00171	0.0151
V$MYOD_01	M00001	0.00184	0.0157
V$AP2_Q6	M00189	0.00230	0.0191
V$SREBP1_02	M00221	0.00361	0.0291
V$EGR1_01	M00243	0.00415	0.0325
V$E47_02	M00071	0.00462	0.0352
V$EGR3_01	M00245	0.00484	0.0359
V$EGR2_01	M00246	0.00634	0.0450
V$AP4_01	M00005	0.00638	0.0450

Table S 39 Transfac transcription factor classes regulating more downregulated genes than expected by chance
TF = transcription factor, TFBS = transcription factor binding site

TF class	TFBS matrix	P-value	Adj. p-value
ZNF740	MA0753.1	3.31E-09	1.92E-06
Pparg::Rxra	MA0065.2	1.61E-08	4.65E-06
E2F6	MA0471.1	6.44E-08	1.24E-05
Spz1	MA0111.1	8.95E-08	1.30E-05
ASCL1	MA1100.1	1.65E-07	1.50E-05
ZNF263	MA0528.1	1.82E-07	1.50E-05
NR2C2	MA0504.1	1.95E-07	1.50E-05
Myod1	MA0499.1	2.07E-07	1.50E-05
KLF5	MA0599.1	3.53E-07	2.27E-05
MZF1	MA0056.1	4.05E-07	2.34E-05
NFIX	MA0671.1	5.50E-07	2.81E-05
SP1	MA0079.3	5.82E-07	2.81E-05
Myog	MA0500.1	1.94E-06	8.63E-05
ID4	MA0824.1	5.12E-06	0.000212
Tcf12	MA0521.1	6.55E-06	0.000243
0MZF1(var.2)	MA0057.1	6.71E-06	0.000243
RREB1	MA0073.1	1.52E-05	0.000519
TCF4	MA0830.1	2.82E-05	0.000907
SP2	MA0516.1	3.27E-05	0.000998
TFAP2A(var.3)	MA0872.1	6.93E-05	0.00195
INSM1	MA0155.1	7.13E-05	0.00195
TFAP2B	MA0811.1	7.41E-05	0.00195
TFAP2A(var.2)	MA0810.1	7.75E-05	0.00195
SNAI2	MA0745.1	0.000112	0.00267

TF class	TFBS matrix	P-value	Adj. p-value
TCF3	MA0522.2	0.000115	0.00267
TFAP2C(var.3)	MA0815.1	0.000121	0.00270
TFAP2B(var.3)	MA0813.1	0.000131	0.00280
EWSR1-FLI1	MA0149.1	0.000136	0.00282
Ascl2	MA0816.1	0.000159	0.00317
TFAP2C	MA0524.2	0.000205	0.00395
NFIC::TLX1	MA0119.1	0.000220	0.00411
HIC2	MA0738.1	0.000291	0.00509
0NHLH1	MA0048.2	0.000298	0.00509
HNF4G	MA0484.1	0.000299	0.00509
NFIA	MA0670.1	0.000363	0.00601
KLF9	MA1107.1	0.000376	0.00604
EBF1	MA0154.3	0.000417	0.00651
ZEB1	MA0103.3	0.000427	0.00651
ZIC3	MA0697.1	0.000529	0.00777
SP3	MA0746.1	0.000537	0.00777
PPARA::RXRA	MA1148.1	0.000589	0.00801
KLF16	MA0741.1	0.000593	0.00801
RARA::RXRG	MA1149.1	0.000612	0.00801
PLAG1	MA0163.1	0.000629	0.00801
EGR3	MA0732.1	0.000632	0.00801
EGR4	MA0733.1	0.000637	0.00801
E2F4	MA0470.1	0.000660	0.00813
NFIC	MA0161.2	0.000762	0.00920
CTCFL	MA1102.1	0.000906	0.0107
ZIC1	MA0696.1	0.00100	0.0116
SMAD2::SMAD3::SMAD4	MA0513.1	0.00114	0.0129
REST	MA0138.2	0.00180	0.0200
FIGLA	MA0820.1	0.00197	0.0215
ZIC4	MA0751.1	0.00205	0.0220
THAP1	MA0597.1	0.00211	0.0222
Hic1	MA0739.1	0.00214	0.0222
RXRB	MA0855.1	0.00245	0.0249
Klf1	MA0493.1	0.00275	0.0275
KLF4	MA0039.3	0.00312	0.0310
TGIF2	MA0797.1	0.00322	0.0311
Nr2f6	MA0677.1	0.00335	0.0318
Rxra	MA0512.2	0.00340	0.0318
PKNOX2	MA0783.1	0.00346	0.0318
SP8	MA0747.1	0.00398	0.0360
Hnf4a	MA0114.3	0.00464	0.0413
PKNOX1	MA0782.1	0.00513	0.0450
RXRG	MA0856.1	0.00567	0.0490

Table S 40 Jaspar transcription factor classes regulating more downregulated genes than expected by chance
TF = transcription factor, TFBS = transcription factor binding site

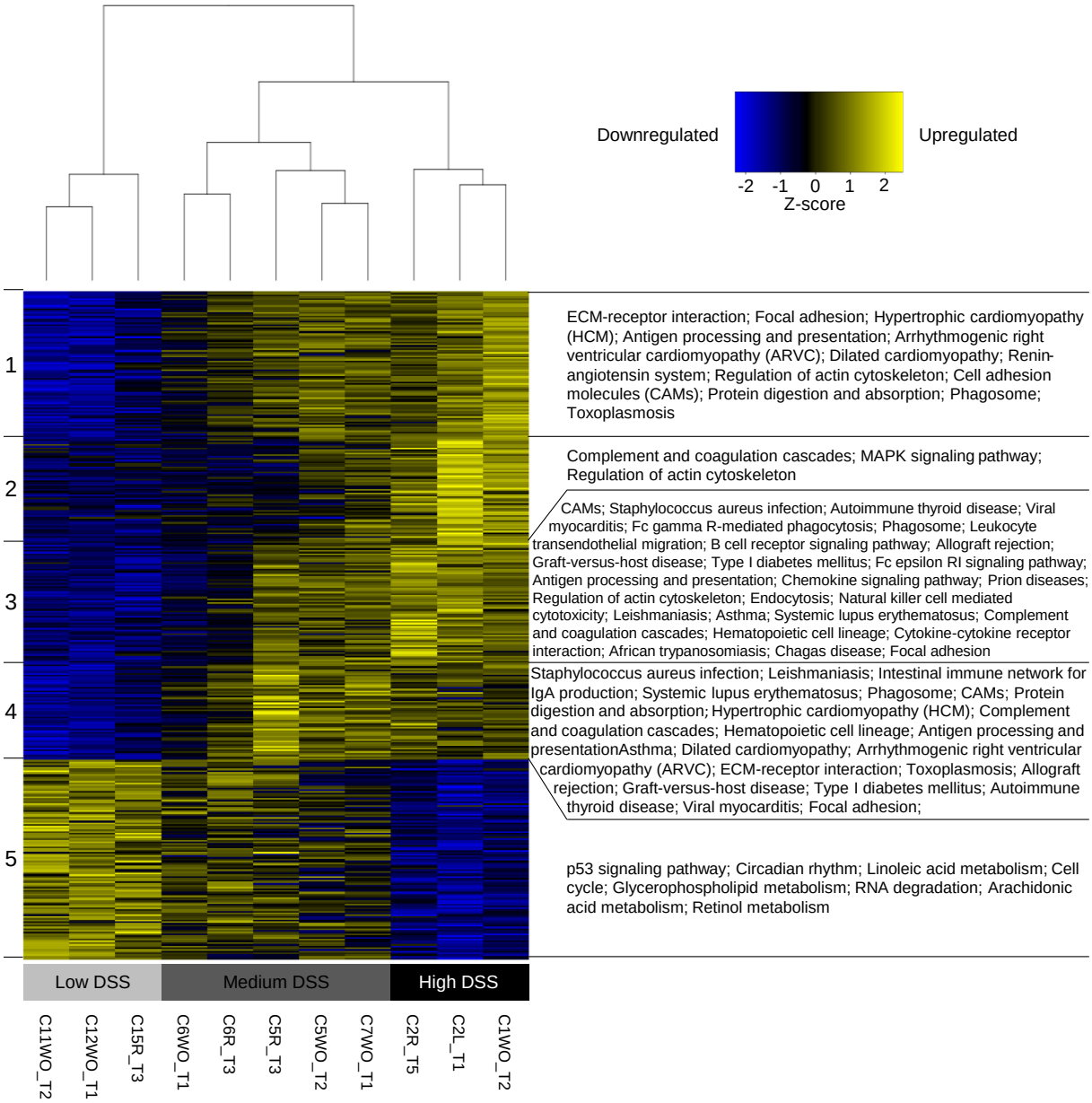


Figure S 46 Processes associated with DSS dose in tumor samples
A k-means clustering was performed on all genes, which were differentially expressed between tumors from mice treated with different DSS doses. KEGG pathways, which harbored significantly more genes of one cluster than expected by chance, are shown on the right site. The following abbreviation was used in the figure: CAMs = Cell adhesion molecules.

	GO term ID	GO term name	GO type	GO group	Adj. p-value
Cluster 1	GO:0022610	Biological adhesion	BP	28	3.11E-09
	GO:0032502	Developmental process	BP	2	1.32E-04
	GO:0040011	Locomotion	BP	2	1.89E-04
	GO:0051270	Regulation of cellular component movement	BP	2	7.53E-04
	GO:0034446	Substrate adhesion-dependent cell spreading	BP	28	0.0010
	GO:0048806	Genitalia development	BP	2	0.0017
	GO:0040012	Regulation of locomotion	BP	2	0.0020
	GO:0070206	Protein trimerization	BP	26	0.0021
	GO:0032501	Multicellular organismal process	BP	2	0.0028
	GO:0043171	Peptide catabolic process	BP	42	0.0038
	GO:0000003	Reproduction	BP	2	0.0092
	GO:0098634	Cell-matrix adhesion mediator activity	MF	40	0.0099
Cluster 3	GO:0002376	Immune system process	BP	14	1.25E-18
	GO:0001775	Cell activation	BP	14	6.76E-12
	GO:0050865	Regulation of cell activation	BP	14	1.92E-07
	GO:0050896	Response to stimulus	BP	14	2.74E-06
	GO:0040011	Locomotion	BP	14	2.98E-06
	GO:0030029	Actin filament-based process	BP	14	3.33E-06
	GO:0006928	Movement of cell or subcellular component	BP	14	9.98E-06
	GO:0008283	Cell proliferation	BP	14	1.00E-05
	GO:0022610	Biological adhesion	BP	14	1.17E-05
	GO:0042127	Regulation of cell proliferation	BP	14	5.48E-05
	GO:0032502	Developmental process	BP	14	6.68E-05
	GO:0048522	Positive regulation of cellular process	BP	14	3.67E-04
	GO:0051704	Multi-organism process	BP	14	7.05E-04
	GO:0032418	Lysosome localization	BP	14	7.05E-04
	GO:0051716	Cellular response to stimulus	BP	14	8.21E-04
	GO:0007165	Signal transduction	BP	14	0.0013
	GO:0023052	Signaling	BP	14	0.0018
	GO:0007154	Cell communication	BP	14	0.0021
	GO:0007010	Cytoskeleton organization	BP	14	0.0021
	GO:0006887	Exocytosis	BP	14	0.0024
	GO:0051179	Localization	BP	14	0.0032
	GO:0060627	Regulation of vesicle-mediated transport	BP	14	0.0035
	GO:0032036	Myosin heavy chain binding	MF	60	0.0045
	GO:0008219	Cell death	BP	15	0.0047
	GO:0042535	Positive regulation of tumor necrosis factor biosynthetic process	BP	14	0.0050
	GO:0033043	Regulation of organelle organization	BP	14	0.0056
	GO:0018212	Peptidyl-tyrosine modification	BP	14	0.0057
	GO:0065007	Biological regulation	BP	14	0.0063
	GO:0051128	Regulation of cellular component organization	BP	14	0.0065
	GO:0006468	Protein phosphorylation	BP	14	0.0065
	GO:0042509	Regulation of tyrosine phosphorylation of STAT protein	BP	14	0.0075

	GO term ID	GO term name	GO type	GO group	Adj. p-value
	GO:0050789	Regulation of biological process	BP	14	0.0078
	GO:0032501	Multicellular organismal process	BP	14	0.0081
	GO:0046427	Positive regulation of JAK-STAT cascade	BP	14	0.0089
	GO:0001573	Ganglioside metabolic process	BP	14	0.0097
	GO:0055094	Response to lipoprotein particle	BP	57	0.0097
Cluster 4	GO:0003366	Cell-matrix adhesion involved in ameboidal cell migration	BP	19	0.0036
	GO:0002682	Regulation of immune system process	BP	10	0.0042
	GO:0002376	Immune system process	BP	10	0.0078
Cluster 5	GO:0044237	Cellular metabolic process	BP	11	0.0013
	GO:0140110	Transcription regulator activity	MF	1	0.0013
	GO:0008152	Metabolic process	BP	11	0.0079
	GO:0003690	Double-stranded DNA binding	MF	7	0.0100

Table S 41 Functional groups associated with DSS dose in tumor samples
A k-means clustering was performed on all genes, which were differentially expressed between tumors from mice treated with different DSS doses. Afterwards, GO term enrichment analyses based on each gene cluster were executed. Processes with significant p-value are shown in the table. Exclusively GO terms of the highest GO hierarchical level and adjusted p-value smaller 0.001 are listed.

	GO term ID	GO term name	GO type	GO group	Adj. p-value
Cluster 1	GO:0018196	Peptidyl-asparagine modification	BP	6	0.0176
	GO:0006487	Protein N-linked glycosylation	BP	6	0.0292
	GO:0031490	Chromatin DNA binding	MF	9	0.0415
	GO:0006334	Nucleosome assembly	BP	8	0.0433
	GO:0016740	Transferase activity	MF	7	0.0440
	GO:0044237	Cellular metabolic process	BP	1	0.0446
	GO:0044238	Primary metabolic process	BP	3	0.0466
	GO:0031497	Chromatin assembly	BP	8	0.0482
	GO:0018279	Protein N-linked glycosylation via asparagine	BP	6	0.0176
	GO:0019538	Protein metabolic process	BP	3	0.0432
	GO:0044267	Cellular protein metabolic process	BP	3	0.0242
Cluster 2	GO:0051131	Chaperone-mediated protein complex assembly	BP	24	0.0006
	GO:0006415	Translational termination	BP	22	0.0006
	GO:1905323	Telomerase holoenzyme complex assembly	BP	13	0.0091
	GO:0045040	Protein import into mitochondrial outer membrane	BP	9	0.0091
	GO:0008079	Translation termination factor activity	MF	17	0.0091
	GO:0006636	Unsaturated fatty acid biosynthetic process	BP	2	0.0091
	GO:0003747	Translation release factor activity	MF	17	0.0091
	GO:0043933	Protein-containing complex subunit organization	BP	22	0.0309
	GO:0097581	Lamellipodium organization	BP	21	0.0309
	GO:0033559	Unsaturated fatty acid metabolic process	BP	2	0.0309
	GO:0030728	Ovulation	BP	12	0.0309
	GO:0090151	Establishment of protein localization to mitochondrial membrane	BP	9	0.0309

	GO term ID	GO term name	GO type	GO group	Adj. p-value
	GO:0070182	DNA polymerase binding	MF	11	0.0309
	GO:0046983	Protein dimerization activity	MF	10	0.0309
	GO:0003729	mRNA binding	MF	20	0.0309
	GO:0006986	Response to unfolded protein	BP	5	0.0309
	GO:0006260	DNA replication	BP	2	0.0309
	GO:0051054	Positive regulation of DNA metabolic process	BP	2	0.0309
	GO:0045723	Positive regulation of fatty acid biosynthetic process	BP	2	0.0309
	GO:0046457	Prostanoid biosynthetic process	BP	2	0.0309
	GO:0001516	Prostaglandin biosynthetic process	BP	2	0.0309
	GO:0035966	Response to topologically incorrect protein	BP	5	0.0338
	GO:0043624	Cellular protein complex disassembly	BP	22	0.0352
	GO:0006633	Fatty acid biosynthetic process	BP	2	0.0366
	GO:0042803	Protein homodimerization activity	MF	10	0.0366
	GO:0050790	Regulation of catalytic activity	BP	19	0.0421
	GO:0051973	Positive regulation of telomerase activity	BP	8	0.0421
	GO:0006457	Protein folding	BP	23	0.0421
	GO:0009725	Response to hormone	BP	16	0.0421
	GO:0000184	Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	BP	18	0.0421
	GO:1904407	Positive regulation of nitric oxide metabolic process	BP	2	0.0421
	GO:0044249	Cellular biosynthetic process	BP	2	0.0421
	GO:0006259	DNA metabolic process	BP	2	0.0421
	GO:0046456	Icosanoid biosynthetic process	BP	2	0.0421
	GO:0045923	Positive regulation of fatty acid metabolic process	BP	2	0.0421
	GO:0042304	Regulation of fatty acid biosynthetic process	BP	2	0.0421
	GO:0010605	Negative regulation of macromolecule metabolic process	BP	4	0.0421
	GO:0010243	Response to organonitrogen compound	BP	16	0.0421
	GO:0006692	Prostanoid metabolic process	BP	2	0.0421
	GO:1901576	Organic substance biosynthetic process	BP	2	0.0421
	GO:0071897	DNA biosynthetic process	BP	2	0.0421
	GO:0045429	Positive regulation of nitric oxide biosynthetic process	BP	2	0.0421
	GO:0034061	DNA polymerase activity	MF	1	0.0421
	GO:0006693	Prostaglandin metabolic process	BP	2	0.0421
	GO:0009058	Biosynthetic process	BP	2	0.0426
	GO:0048545	Response to steroid hormone	BP	16	0.0435
	GO:0043434	Response to peptide hormone	BP	16	0.0435
	GO:0006402	mRNA catabolic process	BP	6	0.0446
	GO:0140097	Catalytic activity, acting on DNA	MF	1	0.0462
	GO:2001233	Regulation of apoptotic signaling pathway	BP	3	0.0473
	GO:1901570	Fatty acid derivative biosynthetic process	BP	2	0.0473
	GO:0048519	Negative regulation of biological process	BP	4	0.0480
	GO:0009892	Negative regulation of metabolic process	BP	4	0.0480
	GO:0006401	RNA catabolic process	BP	6	0.0480
	GO:0051972	Regulation of telomerase activity	BP	8	0.0480
	GO:0009719	Response to endogenous stimulus	BP	16	0.0480

	GO term ID	GO term name	GO type	GO group	Adj. p-value
	GO:1901698	Response to nitrogen compound	BP	16	0.0480
	GO:0045933	Positive regulation of muscle contraction	BP	14	0.0480
	GO:0009891	Positive regulation of biosynthetic process	BP	2	0.0480
	GO:1903428	Positive regulation of reactive oxygen species biosynthetic process	BP	2	0.0480
	GO:0045740	Positive regulation of DNA replication	BP	2	0.0480
	GO:0097367	Carbohydrate derivative binding	MF	7	0.0480
	GO:0001882	Nucleoside binding	MF	7	0.0480
	GO:0032984	Protein-containing complex disassembly	BP	22	0.0480
	GO:1902743	Regulation of lamellipodium organization	BP	21	0.0480
	GO:0031328	Positive regulation of cellular biosynthetic process	BP	2	0.0480
	GO:0051052	Regulation of DNA metabolic process	BP	2	0.0480
	GO:0045428	Regulation of nitric oxide biosynthetic process	BP	2	0.0480
	GO:1901652	Response to peptide	BP	16	0.0480
	GO:0046889	Positive regulation of lipid biosynthetic process	BP	2	0.0489
	GO:0006809	Nitric oxide biosynthetic process	BP	2	0.0489
	GO:2001234	Negative regulation of apoptotic signaling pathway	BP	3	0.0495
Cluster 5	GO:0006366	Transcription by RNA polymerase II	BP	1	0.0331
	GO:0006357	Regulation of transcription by RNA polymerase II	BP	1	0.0331
	GO:0003677	DNA binding	MF	2	0.0376

Table S 42 Functional groups associated with malignant transformation
A k-means clustering was performed on all genes, which were differentially expressed between the proximal colon regions of mice treated with AOM/DSS and controls but not between the proximal colon regions of mice treated with DSS only and controls. The results of the subsequent GO term enrichment analyses are shown in the table. Exclusively GO terms of the highest GO hierarchical level and adjusted p-value smaller 0.001 are listed.

	GO term ID	GO term name	GO type	GO group	Adj. p-value
	GO:0032502	Developmental process	BP	26	9.80E-19
	GO:0022610	Biological adhesion	BP	26	2.94E-13
	GO:0040011	Locomotion	BP	26	1.87E-12
	GO:0032501	Multicellular organismal process	BP	26	1.65E-11
	GO:0071840	Cellular component organization or biogenesis	BP	26	1.73E-11
	GO:0051179	Localization	BP	26	4.17E-08
	GO:0050896	Response to stimulus	BP	26	6.59E-06
	GO:0050789	Regulation of biological process	BP	26	6.59E-06
	GO:0008283	Cell proliferation	BP	26	1.42E-05
	GO:0060346	Bone trabecula formation	BP	26	1.52E-05
	GO:0065007	Biological regulation	BP	26	1.86E-05
	GO:0006793	Phosphorus metabolic process	BP	26	2.27E-04
	GO:0009987	Cellular process	BP	26	3.79E-04
	GO:0044057	Regulation of system process	BP	26	9.07E-04
	GO:0005488	Binding	MF	81	5.11E-13
Cluster 2	GO:0051301	Cell division	BP	10	2.04E-04

	GO term ID	GO term name	GO type	GO group	Adj. p-value
Cluster 3	GO:0005509	Calcium ion binding	MF	5	1.82E-04
	GO:0051179	Localization	BP	23	5.65E-05
	GO:0006928	Movement of cell or subcellular component	BP	23	5.65E-05
	GO:0040011	Locomotion	BP	23	2.23E-04
	GO:0040012	Regulation of locomotion	BP	23	4.55E-04
	GO:0022610	Biological adhesion	BP	34	7.75E-05
	GO:0048519	Negative regulation of biological process	BP	34	2.23E-04
	GO:0048523	Negative regulation of cellular process	BP	34	7.08E-04
Cluster 4	GO:0065008	Regulation of biological quality	BP	19	8.16E-05
	GO:0032502	Developmental process	BP	26	2.65E-06
	GO:0010646	Regulation of cell communication	BP	26	2.15E-04
	GO:0023051	Regulation of signaling	BP	26	2.77E-04
	GO:0009966	Regulation of signal transduction	BP	26	2.89E-04
Cluster 5	GO:0140110	Transcription regulator activity	MF	2	1.91E-04
	GO:0003690	Double-stranded DNA binding	MF	7	1.76E-06
	GO:0001067	Regulatory region nucleic acid binding	MF	7	2.89E-06
	GO:0044212	Transcription regulatory region DNA binding	MF	7	2.89E-06
	GO:0043565	Sequence-specific DNA binding	MF	7	1.19E-04
	GO:0006805	Xenobiotic metabolic process	BP	15	7.01E-04
	GO:0032502	Developmental process	BP	25	1.20E-05
	GO:0006366	Transcription by RNA polymerase II	BP	25	1.44E-05
	GO:1902680	Positive regulation of RNA biosynthetic process	BP	25	1.19E-04
	GO:0048523	Negative regulation of cellular process	BP	25	3.57E-04
	GO:0001503	Ossification	BP	25	3.93E-04
	GO:2000288	Positive regulation of myoblast proliferation	BP	25	5.46E-04
	GO:0042127	Regulation of cell proliferation	BP	25	8.37E-04
	GO:1903507	Negative regulation of nucleic acid-templated transcription	BP	25	9.69E-04
	GO:0005515	Protein binding	MF	27	3.52E-04
Cluster 6	GO:0008152	Metabolic process	BP	1	4.64E-05
	GO:0007399	Nervous system development	BP	7	4.94E-04
	GO:0005488	Binding	MF	37	1.92E-05
Cluster 7	GO:0005488	Binding	MF	5	4.23E-06
	GO:0022610	Biological adhesion	BP	6	3.53E-18
	GO:0032502	Developmental process	BP	6	7.22E-13
	GO:0040011	Locomotion	BP	6	4.92E-12
	GO:0050896	Response to stimulus	BP	6	2.23E-09
	GO:0002376	Immune system process	BP	6	1.84E-08
	GO:0032501	Multicellular organismal process	BP	6	9.03E-07
	GO:0003013	Circulatory system process	BP	6	1.01E-06

	GO term ID	GO term name	GO type	GO group	Adj. p-value
	GO:0008283	Cell proliferation	BP	6	1.02E-06
	GO:0048518	Positive regulation of biological process	BP	6	1.02E-06
	GO:0023052	Signaling	BP	6	2.23E-06
	GO:0065007	Biological regulation	BP	6	2.60E-06
	GO:0048519	Negative regulation of biological process	BP	6	4.06E-06
	GO:0046068	cGMP metabolic process	BP	6	4.42E-06
	GO:0032963	Collagen metabolic process	BP	6	7.90E-06
	GO:0051179	Localization	BP	6	1.17E-04
	GO:0009187	Cyclic nucleotide metabolic process	BP	6	3.48E-04
	GO:0051174	Regulation of phosphorus metabolic process	BP	6	3.57E-04
	GO:0006793	Phosphorus metabolic process	BP	6	8.77E-04
	GO:0098657	Import into cell	BP	6	9.12E-04
	GO:0004713	Protein tyrosine kinase activity	MF	16	8.39E-04
	GO:0042056	Chemoattractant activity	MF	38	8.53E-04
	GO:0046946	Hydroxylysine metabolic process	BP	46	9.11E-05
	GO:0005201	Extracellular matrix structural constituent	MF	62	7.53E-05

Table S 43 Functional groups associated with tumor location
A k-means clustering was performed on all genes, which were differentially expressed between rectal, distal, and central located tumors in colorectal tissues from AOM/DSS treated mice. The results of the subsequent GO term enrichment analyses are shown in the table. Exclusively GO terms of the highest GO hierarchical level and adjusted p-value smaller 0.001 are listed. In each cluster, the results are first ordered by the GO group and subsequently by the p-value.

	Pathway	# DEGs	Pathway size	Adj. p-value
Cluster 1	ECM-receptor interaction	14	88	1.47E-07
	Focal adhesion	18	207	6.37E-06
	Protein digestion and absorption	11	89	3.92E-05
	Amoebiasis	11	119	0.000502
	Malaria	6	47	0.00513
	Leukocyte transendothelial migration	9	122	0.00994
	Cell adhesion molecules (CAMs)	10	154	0.0119
	Dilated cardiomyopathy	7	89	0.0198
	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	6	74	0.0314
	Axon guidance	8	130	0.0361
	Hypertrophic cardiomyopathy (HCM)	6	84	0.0478
Cluster 2	Ubiquitin mediated proteolysis	9	136	0.000506
	Cell cycle	7	123	0.00606
	Regulation of actin cytoskeleton	9	216	0.00606
	Progesterone-mediated oocyte maturation	5	89	0.0258
Cluster 3	Focal adhesion	12	207	0.000824
	ECM-receptor interaction	7	88	0.00411
Cluster 4	Metabolism of xenobiotics by cytochrome P450	10	63	1.07E-07

	Pathway	# DEGs	Pathway size	Adj. p-value
	Drug metabolism	8	65	1.80E-05
	Axon guidance	10	130	3.92E-05
	Retinol metabolism	7	87	0.000776
	Ascorbate and aldarate metabolism	4	26	0.00206
	Drug metabolism	5	50	0.00219
	Starch and sucrose metabolism	5	53	0.00247
	Pentose and glucuronate interconversions	4	35	0.00417
	Porphyrin and chlorophyll metabolism	4	39	0.00506
	Pantothenate and CoA biosynthesis	3	17	0.00500
	Wnt signaling pathway	7	141	0.00548
Cluster 6	TGF-beta signaling pathway	6	81	0.0277
Cluster 7	Focal adhesion	20	207	1.61E-07
	Staphylococcus aureus infection	11	51	1.61E-07
	Leishmaniasis	10	64	1.23E-05
	ECM-receptor interaction	11	88	2.28E-05
	Vascular smooth muscle contraction	13	128	2.28E-05
	Phagosome	14	165	6.39E-05
	Cell adhesion molecules (CAMs)	13	154	0.000129
	Hypertrophic cardiomyopathy (HCM)	9	84	0.000423
	Pathways in cancer	18	322	0.000549
	Dilated cardiomyopathy	9	89	0.000549
	Toxoplasmosis	10	113	0.000549
	Asthma	5	23	0.000549
	Systemic lupus erythematosus	11	139	0.000616
	Complement and coagulation cascades	8	78	0.000899
	Calcium signaling pathway	12	181	0.00139
	Allograft rejection	6	48	0.00186
	Protein digestion and absorption	8	89	0.00186
	Graft-versus-host disease	6	50	0.00208
	Viral myocarditis	7	73	0.00250
	Amoebiasis	9	119	0.00250
	Antigen processing and presentation	7	75	0.00281
	Type I diabetes mellitus	6	55	0.00286
	Intestinal immune network for IgA production	5	41	0.00462
	Autoimmune thyroid disease	6	62	0.00496
	Cytokine-cytokine receptor interaction	13	264	0.00761
	Melanoma	6	71	0.00922
	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	6	74	0.0109
	Primary immunodeficiency	4	35	0.0147
	Prion diseases	4	35	0.0147
	Gap junction	6	87	0.0218
	Glioma	5	65	0.0259
	Regulation of actin cytoskeleton	10	216	0.0289

	Pathway	# DEGs	Pathway size	Adj. p-value
	Endocytosis	10	225	0.0366
	Chagas disease (American trypanosomiasis)	6	104	0.0440
	Lysine degradation	4	51	0.0455

Table S 44 Pathways associated with tumor location
A k-means clustering was performed on all genes, which were differentially expressed between rectal, distal, and central located tumors in colorectal tissues from AOM/DSS treated mice. All significant results of the subsequent KEGG pathway enrichment analyses are shown in the table.

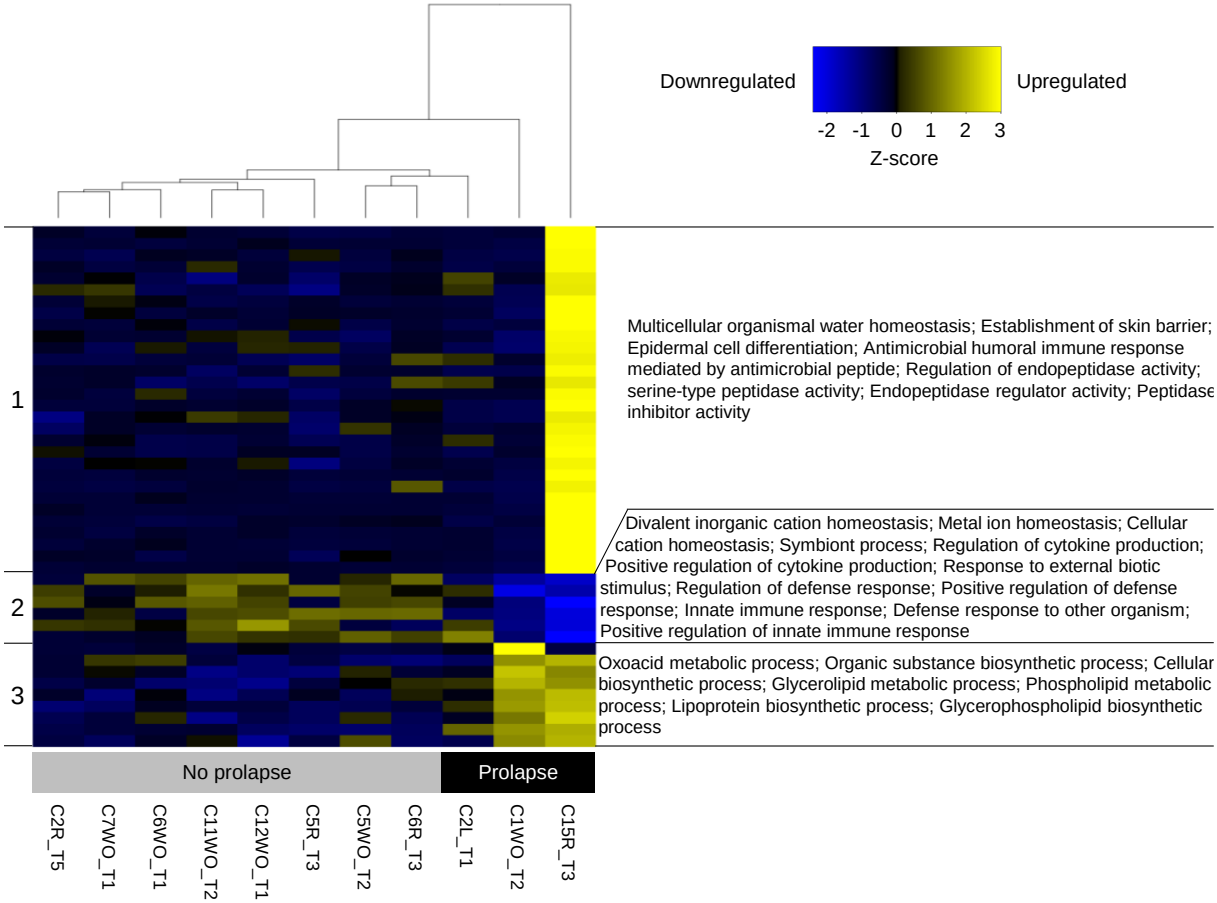


Figure S 47 Processes associated with the development of an intestinal prolapse
A k-means clustering was performed on all genes, which were differentially expressed between tumors from mice with and without intestinal prolapse. GO terms enriched for genes of one of the resulting clusters are shown on the right site of the plot. Exclusively GO terms of the highest GO hierarchical level and an adjusted p-value smaller 0.001 are listed. Only genes of the first cluster existed more often than expected in two KEGG pathways ['Pantothenate and CoA biosynthesis' ($p = 0.00031$) and 'Metabolism of xenobiotics of cytochrome P450' ($p = 0.0022$)].

ID	Term	Group	P-value
GO:0048016	Inositol phosphate-mediated signaling	1	0.0498
GO:0098772	Molecular function regulator	2	1.46E-13
GO:0035615	Clathrin adaptor activity	3	0.0451
GO:0098748	Endocytic adaptor activity	3	0.0451
GO:0044699	Single-organism process	5	5.08E-30
GO:0009987	Cellular process	5	1.99E-26
GO:0008152	Metabolic process	5	4.92E-25

ID	Term	Group	P-value
GO:0071840	Cellular component organization or biogenesis	5	4.38E-22
GO:0032502	Developmental process	5	4.59E-21
GO:0051179	Localization	5	1.00E-17
GO:0065007	Biological regulation	5	1.24E-16
GO:0050789	Regulation of biological process	5	6.50E-15
GO:0032501	Multicellular organismal process	5	8.68E-08
GO:0022610	Biological adhesion	5	7.33E-07
GO:0050896	Response to stimulus	5	3.28E-06
GO:0040007	Growth	5	3.72E-06
GO:0040011	Locomotion	5	9.23E-06
GO:0003012	Muscle system process	5	2.55E-05
GO:0002376	Immune system process	5	0.0003
GO:0032024	Positive regulation of insulin secretion	5	0.0004
GO:0042770	Signal transduction in response to DNA damage	5	0.0008
GO:0006376	mRNA splice site selection	5	0.0009
GO:0071867	Response to monoamine	5	0.0032
GO:0023052	Signaling	5	0.005
GO:0032412	Regulation of ion transmembrane transporter activity	5	0.006
GO:1900063	Regulation of peroxisome organization	5	0.0079
GO:0070831	Basement membrane assembly	5	0.0086
GO:0002793	Positive regulation of peptide secretion	5	0.0099
GO:0048525	Negative regulation of viral process	5	0.0105
GO:0010951	Negative regulation of endopeptidase activity	5	0.0127
GO:0009306	Protein secretion	5	0.0138
GO:0043632	Modification-dependent macromolecule catabolic process	5	0.0149
GO:0035773	Insulin secretion involved in cellular response to glucose stimulus	5	0.0169
GO:0008380	RNA splicing	5	0.0171
GO:2000736	Regulation of stem cell differentiation	5	0.0173
GO:0001837	Epithelial to mesenchymal transition	5	0.0184
GO:1903426	Regulation of reactive oxygen species biosynthetic process	5	0.0185
GO:0003254	Regulation of membrane depolarization	5	0.02
GO:0007420	Brain development	5	0.0228
GO:0060606	Tube closure	5	0.0228
GO:0033031	Positive regulation of neutrophil apoptotic process	5	0.0231
GO:0035022	Positive regulation of Rac protein signal transduction	5	0.0231
GO:1901844	Regulation of cell communication by electrical coupling involved in cardiac conduction	5	0.0231
GO:1902083	Negative regulation of peptidyl-cysteine S-nitrosylation	5	0.0231
GO:1903690	Negative regulation of wound healing, spreading of epidermal cells	5	0.0231
GO:0032647	Regulation of interferon-alpha production	5	0.0239
GO:0072175	Epithelial tube formation	5	0.0241
GO:0006397	mRNA processing	5	0.027
GO:0002428	Antigen processing and presentation of peptide antigen via MHC class Ib	5	0.0271
GO:1901201	Regulation of extracellular matrix assembly	5	0.0271
GO:1901292	Nucleoside phosphate catabolic process	5	0.0278
GO:1903427	Negative regulation of reactive oxygen species biosynthetic process	5	0.0291

ID	Term	Group	P-value
GO:0044283	Small molecule biosynthetic process	5	0.0299
GO:0046887	Positive regulation of hormone secretion	5	0.0302
GO:1901699	Cellular response to nitrogen compound	5	0.0317
GO:0090317	Negative regulation of intracellular protein transport	5	0.0321
GO:0042308	Negative regulation of protein import into nucleus	5	0.0326
GO:0090257	Regulation of muscle system process	5	0.0338
GO:1901385	Regulation of voltage-gated calcium channel activity	5	0.0349
GO:0051704	Multi-organism process	5	0.0356
GO:1901077	Regulation of relaxation of muscle	5	0.0363
GO:0097300	Programmed necrotic cell death	5	0.0389
GO:2000738	Positive regulation of stem cell differentiation	5	0.0393
GO:0044257	Cellular protein catabolic process	5	0.0408
GO:0009166	Nucleotide catabolic process	5	0.042
GO:1904062	Regulation of cation transmembrane transport	5	0.043
GO:0098902	Regulation of membrane depolarization during action potential	5	0.0451
GO:1902106	Negative regulation of leukocyte differentiation	5	0.0453
GO:0090002	Establishment of protein localization to plasma membrane	5	0.0454
GO:1900746	Regulation of vascular endothelial growth factor signaling pathway	5	0.0472
GO:0003013	Circulatory system process	5	0.049
GO:0001913	T cell mediated cytotoxicity	5	0.0498
GO:0035458	Cellular response to interferon-beta	6	0.0298
GO:0090484	Drug transporter activity	8	0.0138
GO:0032452	Histone demethylase activity	9	0.0291
GO:0038044	Transforming growth factor-beta secretion	10	0.0231
GO:0003993	Acid phosphatase activity	11	0.0195
GO:0004382	Guanosine-diphosphatase activity	12	0.0079
GO:0003824	Catalytic activity	13	0.0000
GO:0008234	Cysteine-type peptidase activity	13	0.0038
GO:0022804	Active transmembrane transporter activity	13	0.0015
GO:0035739	CD4-positive, alpha-beta T cell proliferation	14	0.0011
GO:2000561	Regulation of CD4-positive, alpha-beta T cell proliferation	14	0.0011
GO:2000562	Negative regulation of CD4-positive, alpha-beta T cell proliferation	14	0.0231
GO:2001279	Regulation of unsaturated fatty acid biosynthetic process	15	0.0086
GO:0060337	Type I interferon signaling pathway	16	0.0349
GO:0071357	Cellular response to type I interferon	16	0.0416
GO:0019676	Ammonia assimilation cycle	17	0.0231
GO:0005220	Inositol 1,4,5-trisphosphate-sensitive calcium-release channel activity	18	0.0231
GO:0005198	Structural molecule activity	19	0.0002
GO:0022027	Interkinetic nuclear migration	20	0.0195
GO:0051647	Nucleus localization	20	0.0022
GO:1903894	Regulation of IRE1-mediated unfolded protein response	21	0.0451
GO:0071614	Linoleic acid epoxygenase activity	22	0.0451
GO:0007169	Transmembrane receptor protein tyrosine kinase signaling pathway	23	0.0265
GO:0071378	Cellular response to growth hormone stimulus	23	0.0472
GO:0051560	Mitochondrial calcium ion homeostasis	24	0.0239
GO:0005488	Binding	25	1.68E-31

ID	Term	Group	P-value
GO:0033293	Monocarboxylic acid binding	25	0.0022
GO:0042605	Peptide antigen binding	25	0.0055
GO:1990837	Sequence-specific double-stranded DNA binding	25	0.0478
GO:1990715	mRNA CDS binding	26	0.0451
GO:1902476	Chloride transmembrane transport	27	0.0435
GO:0052768	Mannosyl-oligosaccharide 1,3-alpha-mannosidase activity	28	0.0136
GO:0004142	Diacylglycerol cholinephosphotransferase activity	29	0.0079
GO:0043403	Skeletal muscle tissue regeneration	32	0.0191
GO:0043906	Ala-tRNA(Pro) hydrolase activity	33	0.0136
GO:0001076	Transcription factor activity, RNA polymerase II transcription factor binding	34	0.0323
GO:0052767	Mannosyl-oligosaccharide 1,6-alpha-mannosidase activity	35	0.0136
GO:2001185	Regulation of CD8-positive, alpha-beta T cell activation	36	0.0195
GO:0002281	Macrophage activation involved in immune response	37	0.0271
GO:0098660	Inorganic ion transmembrane transport	38	0.0302
GO:0005388	Calcium-transporting ATPase activity	39	0.0363
GO:1903243	Negative regulation of cardiac muscle hypertrophy in response to stress	40	0.0079
GO:0010616	Negative regulation of cardiac muscle adaptation	40	0.0231
GO:0019626	Short-chain fatty acid catabolic process	42	0.0231
GO:0071495	Cellular response to endogenous stimulus	43	0.0283
GO:0002218	Activation of innate immune response	44	0.0460
GO:0043907	Cys-tRNA(Pro) hydrolase activity	45	0.0136
GO:1904152	Regulation of retrograde protein transport, ER to cytosol	46	0.0472
GO:0005159	Insulin-like growth factor receptor binding	47	0.0052
GO:0045134	Uridine-diphosphatase activity	48	0.0451
GO:0030282	Bone mineralization	49	0.0451
GO:0090557	Establishment of endothelial intestinal barrier	50	0.0133
GO:0014722	Regulation of skeletal muscle contraction by calcium ion signaling	51	0.0231
GO:0004062	Aryl sulfotransferase activity	52	0.0451
GO:0005547	Phosphatidylinositol-3,4,5-trisphosphate binding	53	0.0103
GO:0034454	Microtubule anchoring at centrosome	54	0.0451
GO:0031625	Ubiquitin protein ligase binding	55	0.0473
GO:0031852	Mu-type opioid receptor binding	56	0.0451
GO:0016055	Wnt signaling pathway	57	0.0226
GO:0032469	Endoplasmic reticulum calcium ion homeostasis	58	0.0349
GO:0050830	Defense response to Gram-positive bacterium	59	0.0149
GO:0070012	Oligopeptidase activity	60	0.0079
GO:0000414	Regulation of histone H3-K36 methylation	61	0.0451
GO:0006826	Iron ion transport	62	0.0032

Table S 45 GO terms enriched for genes affected by at least one alternative splice variant in tumor samples
Only processes of the highest GO hierarchy level are displayed. The GO terms were ordered by the GO group followed by the adjusted p-value.

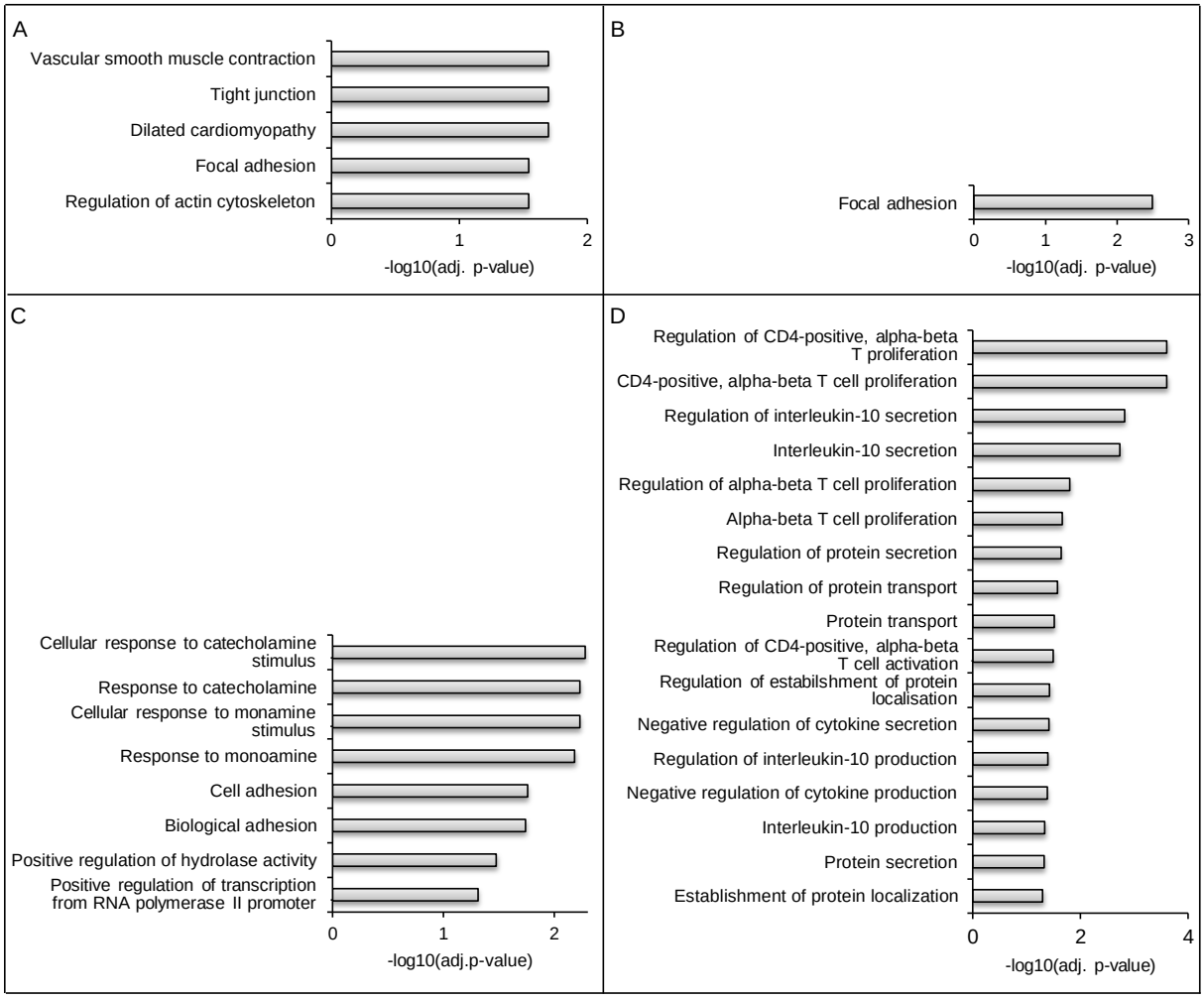


Figure S 48 Processes enriched for genes with different exon retention rate between tumor and all non-tumor samples
(A) KEGG pathways enriched for genes with at least one exon, which was more often retained in the tumor compared to all non-tumor samples. (B) KEGG pathways enriched for genes with at least one exon, which was more often spliced in the tumor compared to all non-tumor samples. (C) GO terms enriched for genes with at least one exon, which was more often retained in the tumor compared to all non-tumor samples. (D) GO terms enriched for genes with at least one exon, which was more often spliced in the tumor compared to all non-tumor samples.

GO ID	GO term	GO group	Adj. p-value
GO:0044699	Single-organism process	1	2.43E-14
GO:0009987	Cellular process	1	5.52E-11
GO:0051179	Localization	1	1.29E-10
GO:0032502	Developmental process	1	1.37E-10
GO:0022610	Biological adhesion	1	8.37E-10
GO:0065007	Biological regulation	1	6.16E-08
GO:0008152	Metabolic process	1	2.40E-07
GO:0071840	Cellular component organization or biogenesis	1	5.03E-07
GO:0050789	Regulation of biological process	1	1.13E-05
GO:0040011	Locomotion	1	2.54E-05
GO:0032501	Multicellular organismal process	1	1.09E-04
GO:0060560	Developmental growth involved in morphogenesis	1	5.77E-04

GO ID	GO term	GO group	Adj. p-value
GO:0048583	Regulation of response to stimulus	1	0.0012
GO:0032787	Monocarboxylic acid metabolic process	1	0.0052
GO:0051493	Regulation of cytoskeleton organization	1	0.0054
GO:0019219	Regulation of nucleobase-containing compound metabolic process	1	0.0066
GO:0010628	Positive regulation of gene expression	1	0.0070
GO:0009605	Response to external stimulus	1	0.0074
GO:0016049	Cell growth	1	0.0088
GO:2000112	Regulation of cellular macromolecule biosynthetic process	1	0.0095
GO:0045893	Positive regulation of transcription, DNA-templated	1	0.013
GO:0051173	Positive regulation of nitrogen compound metabolic process	1	0.013
GO:0006935	Chemotaxis	1	0.015
GO:0018212	Peptidyl-tyrosine modification	1	0.022
GO:0007420	Brain development	1	0.023
GO:0044723	Single-organism carbohydrate metabolic process	1	0.026
GO:0070887	Cellular response to chemical stimulus	1	0.026
GO:0002684	Positive regulation of immune system process	1	0.027
GO:0044702	Single organism reproductive process	1	0.036
GO:0008285	Negative regulation of cell proliferation	1	0.04
GO:0031324	Negative regulation of cellular metabolic process	1	0.044
GO:0097659	Nucleic acid-templated transcription	1	0.049
GO:0055080	Cation homeostasis	3	0.024
GO:0098771	Inorganic ion homeostasis	3	0.031
GO:0042060	Wound healing	5	0.0049
GO:0006950	Response to stress	7	0.0059
GO:0098772	Molecular function regulator	10	0.0047
GO:0005488	Binding	13	6.18E-19
GO:0022804	Active transmembrane transporter activity	19	0.0094
GO:0016757	Transferase activity, transferring glycosyl groups	26	0.050
GO:0003012	Muscle system process	27	0.0050
GO:0090257	Regulation of muscle system process	27	0.0087
GO:0003824	Catalytic activity	45	0.0040

Table S 46 GO terms enriched for genes with at least one exon more often retained in tumor samples
The gene set enrichment analysis was performed with all genes having at least one exon, which was more often retained in the tumor samples compared to the controls (gene selection based on unadjusted $p < 0.05$). Each process consisted of at least nine affected genes. Exclusively processes of the highest level are shown in the table. Therefore, processes such as 'Primary metabolic process', which is a 'Metabolic process', 'Cell cycle', which is a 'Cellular process' as well as a 'Single-organism process', and 'Cytoskeleton organization', which is a 'Cellular component organization or biogenesis' and a 'Cellular process' are not listed in the table. The GO terms were ordered by the GO group followed by the p-value.

GO ID	GO term	GO group	Adj. p-value
GO:0002028	Regulation of sodium ion transport	7	0.0061
GO:0003824	Catalytic activity	10	2.17E-19
GO:0019722	Calcium-mediated signaling	12	0.018
GO:0008234	Cysteine-type peptidase activity	13	0.0010
GO:0005488	Binding	15	1.09E-45
GO:0006066	Alcohol metabolic process	16	0.033
GO:0030522	Intracellular receptor signaling pathway	21	2.09E-04
GO:0098772	Molecular function regulator	26	6.23E-06
GO:0008152	Metabolic process	29	3.78E-28
GO:0009987	Cellular process	29	4.44E-28
GO:0071840	Cellular component organization or biogenesis	29	4.36E-23
GO:0048518	Positive regulation of biological process	29	2.09E-21
GO:0044699	Single-organism process	29	3.23E-20
GO:0051179	Localization	29	6.08E-16
GO:0065007	Biological regulation	29	1.17E-15
GO:0048519	Negative regulation of biological process	29	1.72E-09
GO:0032502	Developmental process	29	2.64E-09
GO:0035556	Intracellular signal transduction	29	4.18E-09
GO:0009966	Regulation of signal transduction	29	4.54E-09
GO:0023057	Negative regulation of signaling	29	0.00103
GO:0003012	Muscle system process	29	0.0018
GO:0034599	Cellular response to oxidative stress	29	0.0025
GO:1901652	Response to peptide	29	0.0038
GO:1901701	Cellular response to oxygen-containing compound	29	0.0040
GO:2000736	Regulation of stem cell differentiation	29	0.0054
GO:0040007	Growth	29	0.0067
GO:0050767	Regulation of neurogenesis	29	0.0097
GO:0044419	Interspecies interaction between organisms	29	0.012
GO:1901699	Cellular response to nitrogen compound	29	0.012
GO:0030155	Regulation of cell adhesion	29	0.014
GO:0090316	Positive regulation of intracellular protein transport	29	0.018
GO:0032434	Regulation of proteasomal ubiquitin-dependent protein catabolic process	29	0.019
GO:0045860	Positive regulation of protein kinase activity	29	0.021
GO:0042542	Response to hydrogen peroxide	29	0.022
GO:0006090	Pyruvate metabolic process	29	0.023
GO:0050770	Regulation of axonogenesis	29	0.026
GO:0006757	ATP generation from ADP	29	0.028
GO:0006397	mRNA processing	29	0.028
GO:0007030	Golgi organization	29	0.028
GO:0032271	Regulation of protein polymerization	29	0.031
GO:0045862	Positive regulation of proteolysis	29	0.032
GO:1901657	Glycosyl compound metabolic process	29	0.035
GO:0051047	Positive regulation of secretion	29	0.036

GO ID	GO term	GO group	Adj. p-value
GO:0050896	Response to stimulus	29	0.036
GO:0046128	purine ribonucleoside metabolic process	29	0.042
GO:0019362	Pyridine nucleotide metabolic process	29	0.043
GO:0034248	Regulation of cellular amide metabolic process	29	0.046
GO:0002218	Activation of innate immune response	31	0.043
GO:0055072	Iron ion homeostasis	32	0.025
GO:0001012	RNA polymerase II regulatory region DNA binding	35	0.037
GO:0043624	Cellular protein complex disassembly	37	0.049
GO:0060048	Cardiac muscle contraction	52	0.045
GO:0051427	Hormone receptor binding	54	0.035
GO:0005200	Structural constituent of cytoskeleton	60	0.0089

Table S 47 GO terms enriched for genes with at least one exon more often skipped in tumor samples
The gene set enrichment analysis was performed with all genes having at least one exon, which was more often skipped in the tumor samples compared to the controls (gene selection based on unadjusted $p < 0.05$). Only processes consisting of at least nine affected genes are displayed. Exclusively processes of the highest level are shown in the table. Therefore, processes like ‘P53 binding’, which is a ‘Binding’ process, ’I-KB kinase / NF-KB signaling’ and ‘TOR signaling’, which are both, amongst others, a ‘Response to a stimulus’ and a ‘Biological regulation’, and ‘Cell cycle’, which is in turn a ‘Cellular process’ as well as a ‘Single organism process’, are not listed in the table. The GO terms were ordered by the GO group followed by the adjusted p-value.

6.2.2.8 Combination of data layers

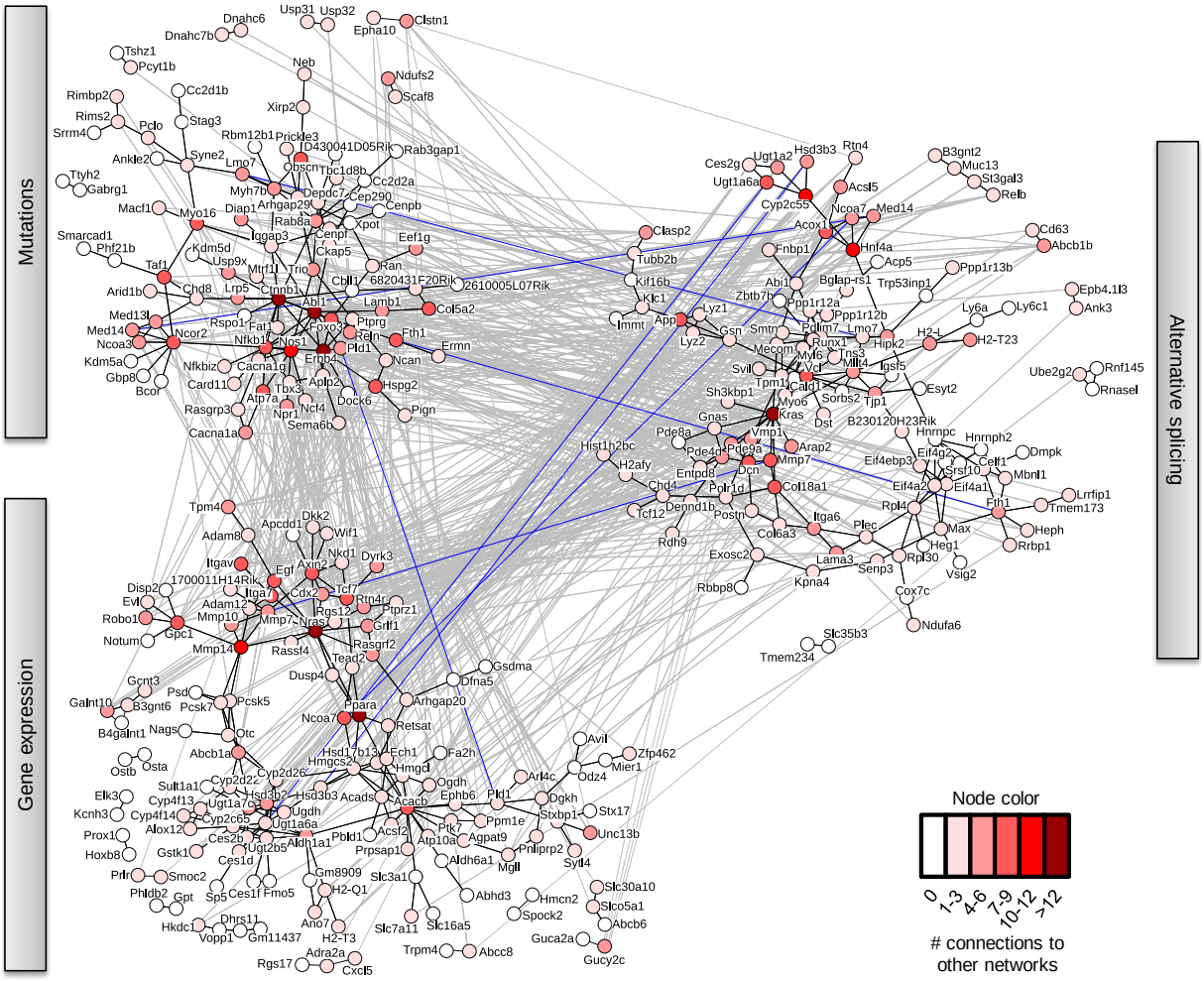


Figure S 49 Protein-protein interaction network connecting differentially mutated with differentially spliced and differentially expressed genes

The subnetwork based on genes, which were more often mutated in tumor than in control samples, is based on the WES results described in chapter 3.2.2.5. The gene expression subnetwork is based on the 250 differentially expressed genes with lowest p-value. The splicing subnetwork includes all genes affected by different exon usage or different intron retention rate as well as the top 100 genes with alternative splice site (lowest p-value) in the tumor samples. The grey lines indicate direct connections between two genes of different subnetworks. Blue lines connect genes existing in two subnetworks. The node color is based on the number of connections between the respective gene and the genes out of other subnetworks.

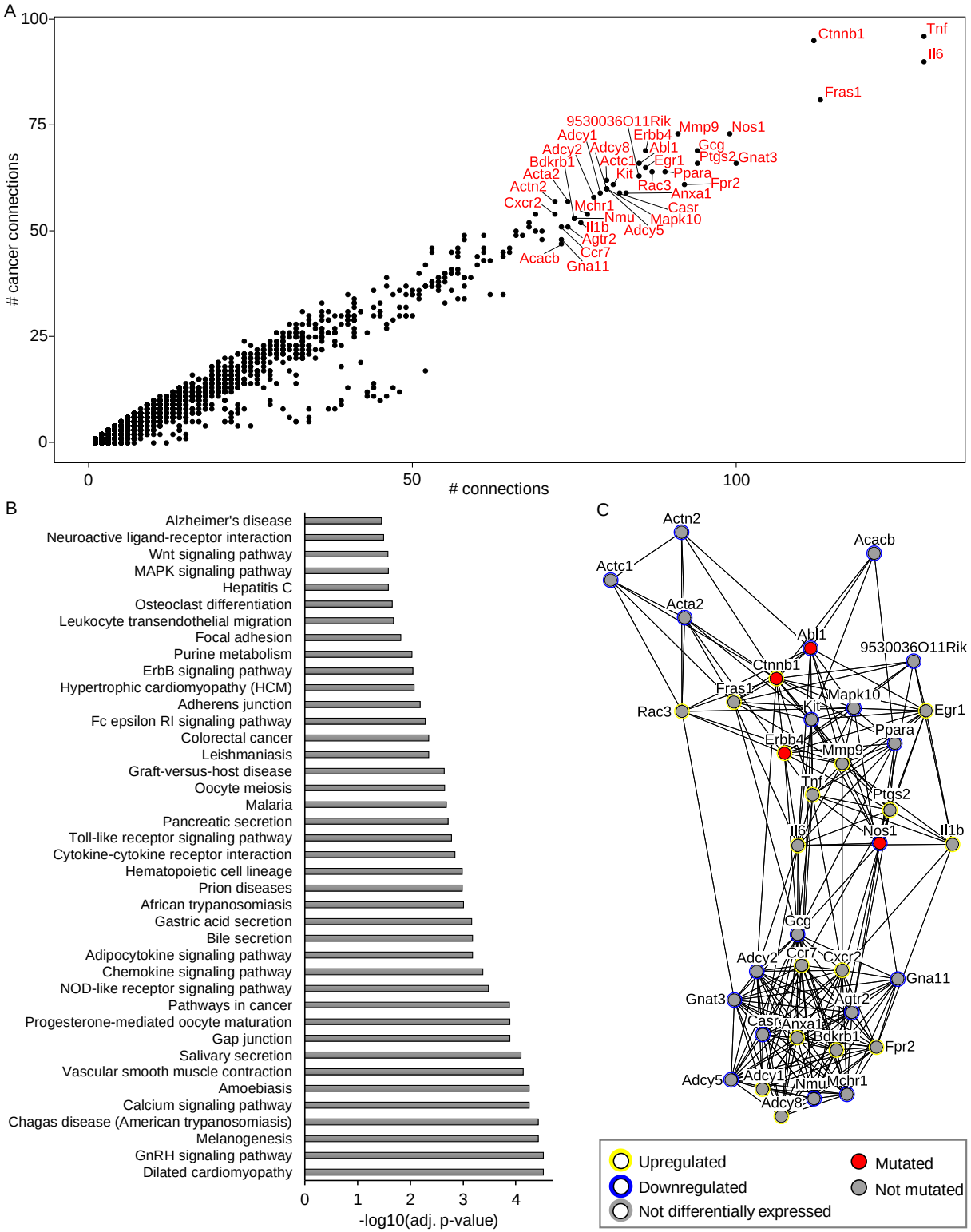


Figure S 50 Genes with highest connection rate within the protein-protein interaction network based on differentially mutated and differentially expressed genes

(A) The number of connections to other genes, which were more often mutated in tumor than in control samples or differentially expressed, was plotted against the number of connections to altered cancer-related genes annotated in the Cosmic database. Genes with red label harbored the highest number of connections and were investigated more detailed. (B) KEGG pathways enriched for highly connected genes shown in (A). Exclusively processes with at least three affected genes are listed. (C) Protein-protein interaction network based on highly connected genes shown in (A). The colors of the nodes indicate the observed alterations.

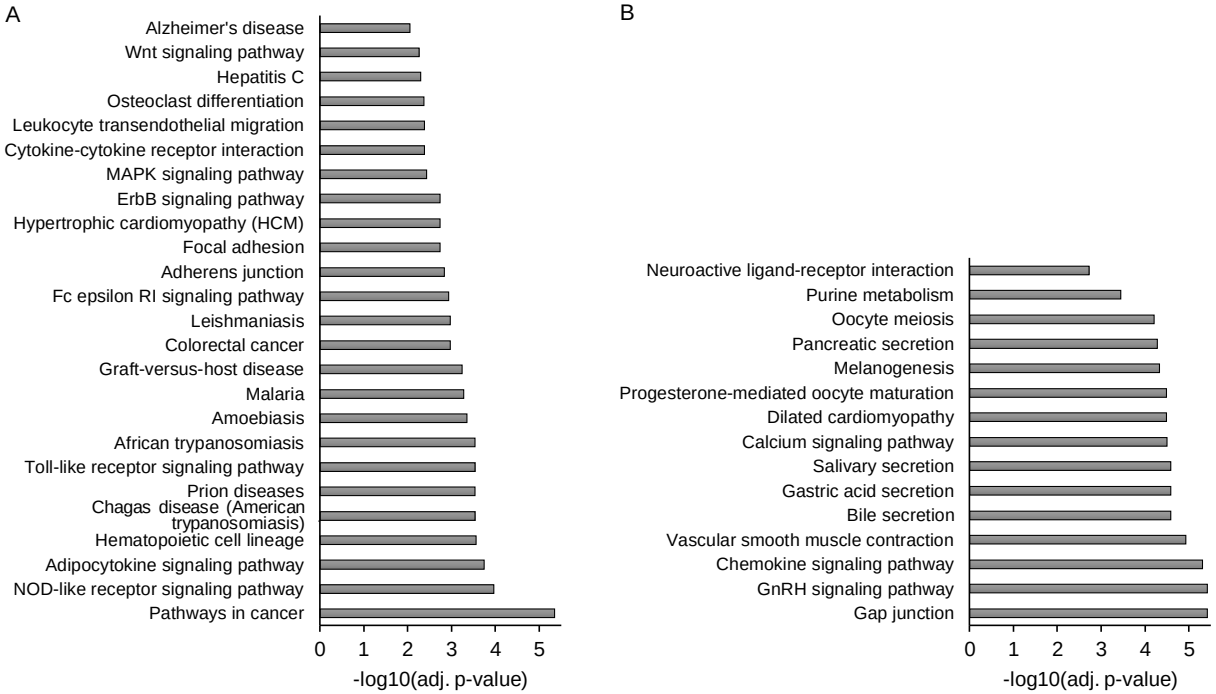


Figure S 51 KEGG pathway enrichment analyses based on clusters of highly connected genes in the protein-protein interaction network of differentially expressed and differentially mutated genes

A protein-protein interaction network based on all genes, which were more often mutated in tumor than in control samples or differentially expressed, was created. Genes with the highest connection rates formed a protein-protein interaction network with two subclusters (Figure S 50). The results of the KEGG enrichment analyses are shown for the first upper subcluster in figure (A) and for the lower subnetwork in (B).

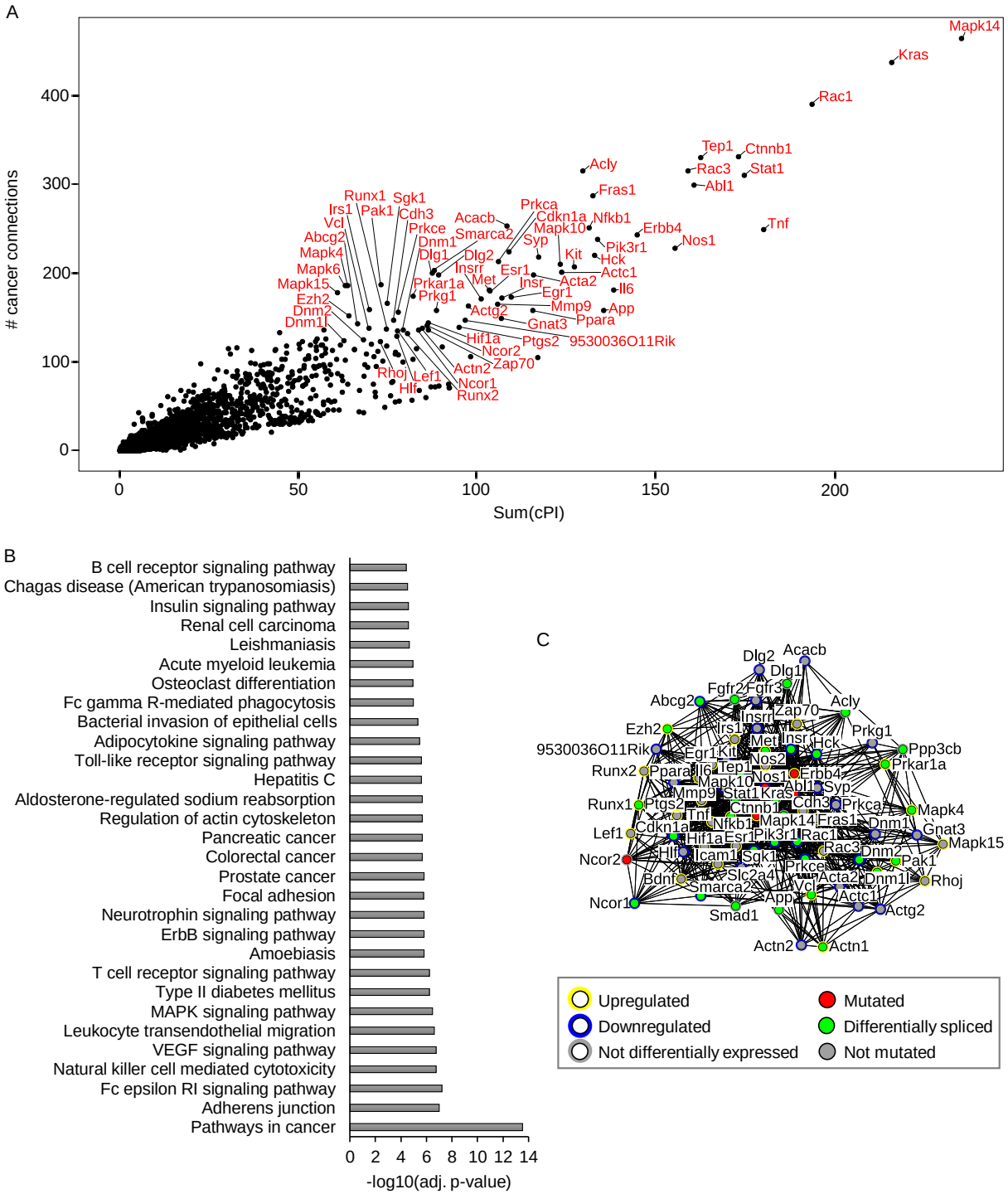


Figure S 52 Genes with highest cPI sum and highest number of connected cancer genes in a protein-protein interaction network based on differentially mutated, differentially expressed, and differentially spliced genes

A protein-protein interaction network based on all differentially mutated, differentially expressed, and differentially spliced genes was created. (A) Genes with the highest sum of absolute cPI values of all interaction partners and the highest number of connections to cancer-related genes were filtered and labeled in red in the scatter plot. For this analysis, all interaction partners including those, which were not altered and therefore not part of the network, were considered. (B) KEGG pathways enriched for highly connected genes with red label in (A). Exclusively processes with at least two affected genes are shown. (C) Protein-protein interaction network based on highly connected genes labeled in red in (A). The colors of the nodes indicate the observed alterations.

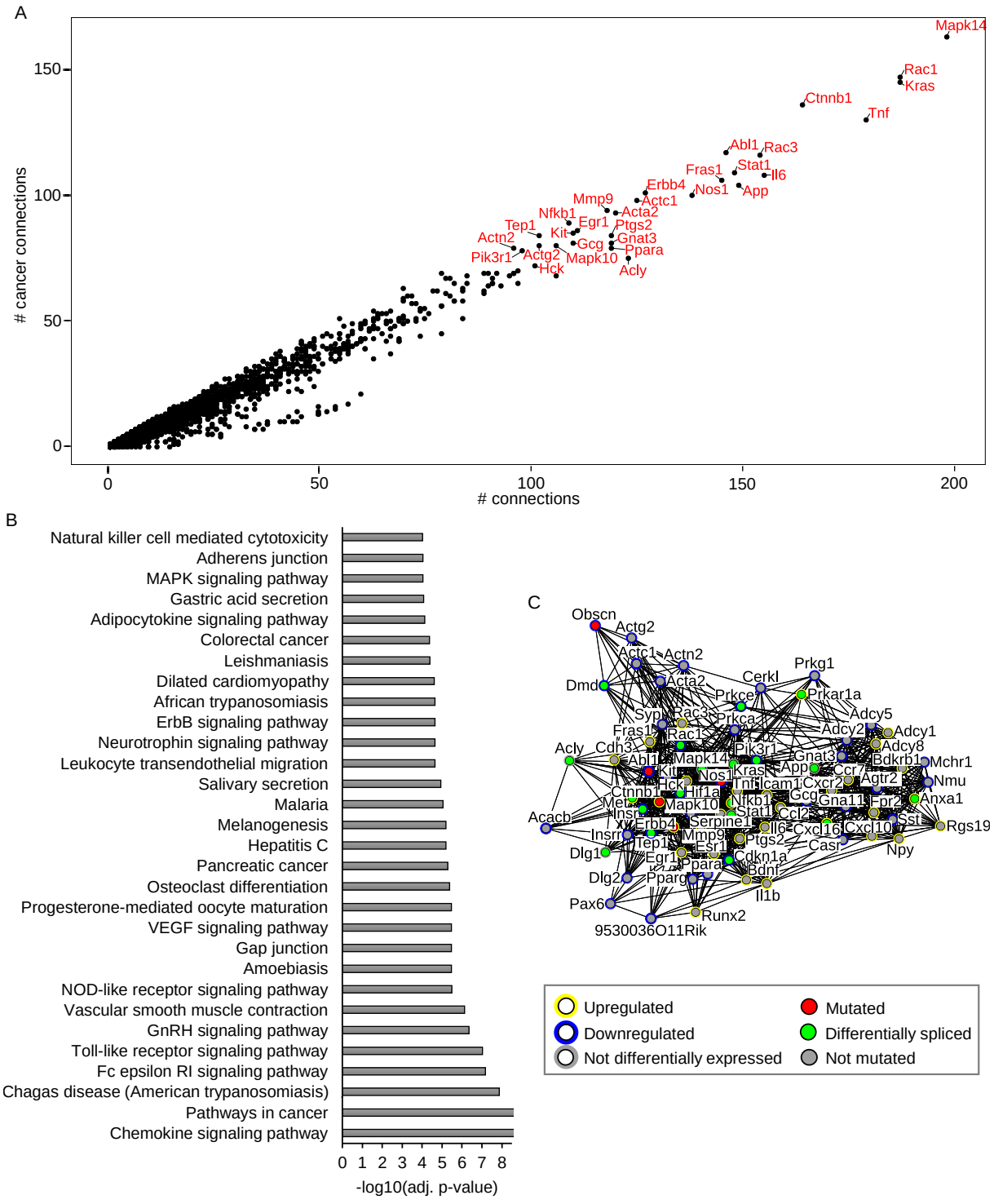


Figure S 53 Genes with highest connection rate within a protein-protein interaction network based on differentially mutated, differentially expressed, and differentially spliced genes
(A) The number of connection to other differentially mutated, differentially expressed or differentially spliced genes was plotted against the number of connection to altered cancer-related genes annotated in the Cosmic database. Genes with red label harbored the highest number of connections and were investigated more detailed. (B) KEGG pathways enriched for highly connected genes labeled in red in (A). Exclusively processes with at least three affected genes are listed (C) Protein-protein interaction network based on highly connected genes labeled in red in (A). The colors of the nodes indicate the observed alterations.

7 Summary (English)

Cancer causes the highest number of deaths among all diseases in economically developed countries. The course of the disease and the therapeutic success are strongly influenced by molecular factors. Therefore, a better understanding of genomic mechanisms underlying tumor development would create novel treatment possibilities and thus improve the patients' outcome. In the current study, next-generation sequencing (NGS)-based methods were employed to analyze the molecular background of two gastrointestinal cancer types with the following experimental parts: (i) Human primary specimens from two exemplary gastric cancer (GC) patients were examined at high resolution. (ii) The genomic landscape of colitis-associated colorectal cancer (CAC) was investigated in a mouse model.

The first part of my studies aimed to decipher novel molecular insights into GC, which causes ~10% of all cancer deaths worldwide and shows a 5-year survival rate of only 10-30%. Therefore, whole genomes and whole exomes of one microsatellite stable and one microsatellite instable gastric carcinoma as well as surrounding control tissues from female patients were investigated. To decipher relevant molecular features with this small sample number, a novel approach using two different NGS technologies and population-based resources was suggested. This revealed specific potentially GC-associated variant patterns, SNVs in *RERE*, *TP53*, and *PIK3CA*, small InDels in *BRAF* and *ARID1A*, structural variants in *MLL3*, *MCC*, and *VHLL* as well as altered processes, such as 'DNA cytosine deamination' and 'Negative regulation of transposition'. The mutational landscapes differed between the individual tumors, which might influence the therapy response. This illustrates the potential importance of NGS-based approaches to support diagnostics and treatment choice in GC patients.

In the second part of my studies, a mouse model for CAC was investigated. CAC is a serious complication of inflammatory bowel diseases and has a high mortality rate due to its aggressive biological behavior and a high tendency to metastasize. Although CAC is clinically well characterized, little is known about the underlying molecular processes. To receive novel insights into the genetic basis of inflammation-triggered colorectal cancer, the well-established AOM/DSS mouse model for the investigation of CAC was analyzed using whole exome and transcriptome sequencing. Early and late stage tumor samples from 15 AOM/DSS-treated mice were compared with formerly inflamed colonic tissue samples from mice exposed to either DSS (five mice) or AOM/DSS (15 mice) as well as with colonic tissue samples from six untreated mice. Bioinformatic analyses based on variant, gene, and process level as well as integrative networks combining exome and transcriptome data revealed molecular commonalities between the mouse model and human CAC (e.g. altered *Tnf* and a modified

Wnt signaling pathway). The results provided also hints for further genes, including *Nfκbiζ* and *Ppara*, which might play a role in the development of inflammation-triggered colorectal cancer. Moreover, expression patterns of disease characteristics, such as tumor stage, tumor localization, and intestinal prolapse, as well as specific SNV signatures were elucidated. However, the results demonstrated also differences between the human disease and the mouse model, including the lack of mutations in *Trp53* in the murine tumor samples. These differences might explain phenotypic discrepancies, such as missing metastasis in AOM/DSS-treated mice, and should be considered while using the AOM/DSS mouse model for the investigation of human CAC.

As part of these cancer studies, I developed novel bioinformatic tools and pipelines. These include (i) a population-based approach to identify disease-relevant genes and processes in human studies with a low sample number, (ii) a new algorithm to discover large insertions in the genome, and (iii) a program for the detection and characterization of novel transcriptionally active regions.

Taken together, these two studies further emphasize the power of investigating cancer pathophysiology via NGS in individual clinical settings (few samples, high sequencing depth) as well as in experimental animal models. Both studies revealed novel molecular insights into the analyzed gastrointestinal cancer types, which is an important step towards the development of novel treatment and diagnostic approaches.

8 Zusammenfassung (Deutsch)

Krebs verursacht in Industrieländern mehr Todesfälle als jede andere Krankheit. Krankheitsverlauf und Therapieerfolg werden dabei sehr stark durch molekulare Faktoren beeinflusst. Ein besseres Verständnis der genomischen Mechanismen, die der Tumorentwicklung zugrunde liegen, würde neue verbesserte Therapiemöglichkeiten erschließen. Im Rahmen dieser Arbeit wurden mit Methoden, die auf ‚Next-Generation' Sequenzierung (NGS) basieren, die molekularen Ursachen zweier gastrointestinaler Krebsarten mit folgenden experimentellen Ansätzen analysiert: (1) Exemplarisch wurden Biopsien von zwei Patientinnen mit Magenkarzinomen in hoher Auflösung untersucht. (2) Genomische Signaturen von Colitis-assoziiertem Dickdarmkrebs (CAC) wurden in einem Mausmodell erforscht.

Der erste Teil meiner Arbeit beschäftigt sich mit Magenkrebs, der weltweit etwa 10% aller Krebstodesfälle verursacht und eine 5-Jahres-Überlebensrate von nur 10-30% aufweist. In meinen Studien wurden Genom- und Exomsequenzierungen für ein Magenkarzinom mit stabilen und eines mit instabilen Mikrosatelliten sowie für das jeweils umgebende Kontrollgewebe durchgeführt. Um mit dieser kleinen Probenanzahl relevante molekulare Merkmale aufzudecken, wurde ein neues Analyseverfahren entwickelt. Dieses kombiniert zwei unterschiedliche NGS Technologien und integriert populationsbasierte Datenquellen. Dadurch konnten in den Tumorproben spezifische Mutationsmuster, Basenaustausche in den Genen *RERE*, *P53* und *PIK3CA* sowie InDels in *BRAF* und *ARID1A* und Strukturvarianten in *MLL3*, *MCC* und *VHLL* identifiziert werden. Außerdem wurden Veränderungen der Prozesse der DNA-Cytosin-Desaminierung und der negativen Regulation der Transposition in den Tumorproben festgestellt. Die Ergebnisse zeigten auch, dass sich die molekularen Signaturen der beiden Tumore deutlich unterscheiden, was den Therapieerfolg beeinflussen kann. Daher wäre der Einsatz von NGS-basierten Ansätzen für die Diagnose und Behandlungswahl bei Patienten mit Magenkarzinomen von enormer Bedeutung.

Im zweiten Teil meiner Studien wurde ein Mausmodell für den Dickdarmkrebs CAC untersucht. CAC ist eine schwerwiegende Komplikation bei chronisch-entzündlichen Darmerkrankungen und weist eine hohe Todesrate auf, die durch ein aggressives Tumorwachstum mit hoher Metastasen-Wahrscheinlichkeit verursacht wird. Obwohl CAC klinisch sehr gut charakterisiert ist, ist wenig über die zugrunde liegenden molekularen Prozesse bekannt. Um neue Einblicke in den genetischen Hintergrund von CAC zu erhalten, wurde das gut etablierte AOM/DSS Mausmodell für die Untersuchung von entzündungsassoziiertem Dickdarmkrebs mithilfe von Exom- und Transkriptomsequenzierung analysiert. Tumore aus unterschiedlichen Stadien wurden mit dem jeweils umgebenden, vormals entzündeten Darmgewebe aus 15 AOM/DSS

behandelten Mäusen sowie mit proximalem Darmgewebe aus fünf Mäusen, die DSS mit dem Trinkwasser erhalten haben und dadurch eine chronische Entzündung im Darm aufwiesen, und mit sechs Kontrollen verglichen. Bioinformatische Analysen basierend auf Varianten-, Gen- und Prozesslevel sowie integrative Netzwerke zur Kombination von Exom- und Transkriptomdaten offenbarten molekulare Veränderungen, die auch bei humanem CAC beobachtet wurden, wie z.B. Veränderungen in dem Gen *Tnf* oder dem Wnt-Signalweg. Darüber hinaus wurden Hinweise für eine funktionale Rolle weiterer Gene bei der Entwicklung von entzündungsassoziiertem Dickdarmkrebs gefunden (z.B. *Nfκbiζ* und *Ppara*). Außerdem konnten Expressionsmuster mit Eigenschaften wie Tumorstadium, Tumorposition und Darmvorfall in Verbindung gebracht sowie spezifische Mutationssignaturen charakterisiert werden. Es existierten allerdings auch Abweichungen zwischen humanem CAC und dem Mausmodell, wie beispielweise fehlende murine Mutationen in *Trp53*. Diese genetischen Unterschiede könnten phänotypische Gegensätze, wie z.B. fehlende Metastasenbildung im Mausmodell, erklären und sollten bei der Anwendung des Models für die Untersuchung von CAC berücksichtigt werden.

Im Rahmen der beiden Krebsstudien habe ich neuartige bioinformatische Programme und Analyseverfahren entwickelt. Dazu zählen unter anderem (1) ein populationsbasierter Ansatz, um relevante Gene und Prozesse in humanen Studien mit niedriger Probenanzahl zu identifizieren, (2) ein neuer Algorithmus, um große Insertionen im Genom zu finden, sowie (3) ein Programm für die Detektion und Charakterisierung von nicht annotierten, transkriptionell aktiven Regionen.

Zusammenfassend unterstreichen diese beiden Studien die enormen Möglichkeiten bei der Analyse der Pathophysiologie von Krebserkrankungen mit NGS unter Nutzung von individuellen klinischen Bedingungen oder mithilfe eines Mausmodells. Durch beide Studien konnten neue Einblicke in die molekularen Mechanismen der beiden untersuchten gastrointestinalen Krebsarten gewonnen werden. Das ist ein wichtiger Schritt, um neue Behandlungsmethoden und Therapieansätze entwickeln zu können.

9 Publications

Esser, D., Falk-Paulsen, M., Ito, G., Kuiper, J., Aden, K., Billmann-Born, S., Schreiber, S., Kaleta, C., Rosenstiel, P., Mutational and transcriptional landscape of murine inflammation-induced colorectal cancer. **Submitted.**

Fabian, A., Stegner, S., Miarka, L., Zimmermann, J., Lenk, L., Rahn, S., Buttlar, J., Viol F., Knaack, H., **Esser, D.**, Schäuble, S., Großmann, P., Marinos, G., Häsler, R., Mikulits, W., Saur, D., Kaleta, C., Schäfer, H., Sebens, S. (2019). Metastasis of pancreatic cancer: An uninflamed liver micromilieu controls cell growth and cancer stem cell properties by oxidative phosphorylation in pancreatic ductal epithelial cells. **Cancer Letters (accepted).**

Esser, D.[§], Lange, J.[§], Marinos, G.[§], Sieber, M., Best, L., Prasse, D., Bathia, J., Rühlemann M.C., Boersch K., Jaspers C., Sommer F., **§ Equal contribution**. (2018). Functions of the Microbiota for the Physiology of Animal Metaorganisms. **Journal of Innate Immunity.**

Aden, K., Bartsch, K., Dahl, J., Reijns, M.A.M., **Esser, D.**, Sheibani-Tezerji, R., Sinha, A., Wottawa, F., Ito, G., Mishra, N., Knittler, K., Burkholder, A., Welz, L, van Es, J., Tran, F., Lipinski, S., Kakavand, N., Boeger, C., Lucius, R., von Schoenfels, W., Schafmayer, C., Lenk, L., Chalaris, A., Clevers, H., Röcken, C., Kaleta, C., Rose-John, S., Schreiber, S., Kunkel, T., Rabe, B., and Rosenstiel, P. (2018). Epithelial RNase H2 Maintains Genome Integrity and Prevents Intestinal Tumorigenesis in Mice. **Gastroenterology.**

Aramillo Irizar, P.[§], Schäuble, S.[§], **Esser, D.[§],** Groth, M., Frahm C., Priebe S., Baumgart M., Hartmann N., Marthandan, S., Menzel, U., Müller, J., Schmidt, S., Ast, V., Caliebe, A., König, R., Krawczak, M., Ristow, M., Schuster, S., Cellerino, A., Diekmann, S., Englert, C., Hemmerich, P., Sühnel, J., Guthke, R., Witte, O.W., Platzer, M., Ruppin, E., Kaleta, C., **§ Equal contribution**. (2018). Transcriptomic alterations during ageing reflect the shift from cancer to degenerative diseases in the elderly. **Nature Communications.**

Esser, D., Holze, N., Haag, J., Schreiber, S., Krüger, S., Warneke, V., Rosenstiel, P., Röcken, C. (2017). Interpreting whole genome and exome sequencing data of individual gastric cancer samples. **BMC Genomics.**

Greenwood, J.M., Milutinović, B., Peuß, R., Behrens, S., **Esser, D.**, Rosenstiel, P., Schulenburg, H., and Kurtz, J. (2017). Oral immune priming with Bacillus thuringiensis induces a shift in the gene expression of Tribolium castaneum larvae. **BMC Genomics.**

Häsler, R., Sheibani-Tezerji, R., Sinha, A., Barann, M., Rehman, A., **Esser, D.**, Aden, K., Knecht, C., Brandt, B., Nikolaus, S., Schäuble, S., Kaleta, C., Franke, A., Fretter, C., Müller, W., Hütt, M.-T., Krawczak, M., Schreiber, S., and Rosenstiel, P. (2016). Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. **Gut.**

Yang, W., Dierking, K., **Esser, D.**, Tholey, A., Leippe, M., Rosenstiel, P., and Schulenburg, H. (2015). Overlapping and unique signatures in the proteomic and transcriptomic responses of the nematode Caenorhabditis elegans toward pathogenic Bacillus thuringiensis. **Dev Comp Immunol.**

Behrens, S., Peuß, R., Milutinović, B., Eggert, H., **Esser, D.**, Rosenstiel, P., Schulenburg, H., Bornberg-Bauer, E., and Kurtz, J. (2014). Infection routes matter in population-specific responses of the red flour beetle to the entomopathogen Bacillus thuringiensis. **BMC Genomics.**

Lappalainen, T., Sammeth, M., Friedländer, M.R., ’t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., **Esser, D.**, Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D.G., Lek, M., Lizano, E., Buermans, H.P.J., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S.B., Donnelly, P., McCarthy, M.I., Flicek, P., Strom, T.M., Geuvadis Consortium, Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S.E., Häsler, R., Syvänen, A.-C., van Ommen, G.-J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigó, R., Gut, I.G., Estivill, X., Dermitzakis, E.T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. **Nature.**

Barann, M., **Esser, D.**, Klostermeier, U.C., Lappalainen, T., Luzius, A., Kuiper, J.W.P., Ammerpohl, O., Vater, I., Siebert, R., Amstislavskiy, V., Sudbrak, R., Lehrach, H., Schreiber, S., and Rosenstiel, P. (2013). Janus--a comprehensive tool investigating the two faces of transcription. **Bioinformatics.**

Kreck, B., Richter, J., Ammerpohl, O., Barann, M., **Esser, D.**, Petersen, B.S., Vater, I., Murga Penas, E.M., Bormann Chung, C.A., Seisenberger, S., Lee Boyd, V., Smallwood, S., Drexler, H.G., Macleod, R. A. F., Hummel, M., Krueger, F., Häsler, R., Schreiber, S., Rosenstiel, P., Franke, A., and Siebert, R. (2013). Base-pair resolution DNA methylome of the EBV-positive Endemic Burkitt lymphoma cell line DAUDI determined by SOLiD bisulfite-sequencing. **Leukemia.**

Niederreiter, L., Fritz, T.M., Adolph, T.E., Krismer, A.-M., Offner, F.A., Tschurtschenthaler, M., Flak, M.B., Hosomi, S., Tomczak, M.F., Kaneider, N.C., Sarcevic, E., Kempster, S.L., Raine, T., **Esser, D.**, Rosenstiel, P., Kohno, K., Iwawaki, T., Tilg, H., Blumberg, R.S., and Kaser, A. (2013). ER stress transcription factor Xbp1 suppresses intestinal tumorigenesis and directs intestinal stem cells. **J Exp Med.**

Schramm, A., Köster, J., Marschall, T., Martin, M., Schwermer, M., Fielitz, K., Büchel, G., Barann, M., **Esser, D.**, Rosenstiel, P., Rahmann, S., Eggert, A., and Schulte, J.H. (2012). Next-generation RNA sequencing reveals differential expression of MYCN target genes and suggests the mTOR pathway as a promising therapy target in MYCN-amplified neuroblastoma. **Int J Cancer.**

Publications as part of the GEUVADIS Consortium

Greger, L., Su, J., Rung, J., Ferreira, P.G., **GEUVADIS Consortium**, Lappalainen, T., Dermitzakis, E.T., and Brazma, A. (2014). Tandem RNA chimeras contribute to transcriptome diversity in human population and are associated with intronic genetic variants. **PLoS One.**

’t Hoen, P.A., Friedländer, M.R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S.Y., Laros, J.F.J., Buermans, H.P.J., Karlberg, O., Brännvall, M., **GEUVADIS Consortium**, den Dunnen, J.T., van Ommen, G.-J.B., Gut, I.G., Guigó, R., Estivill, X., Syvänen, A.-C., Dermitzakis, E.T., and Lappalainen, T. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. **Nat Biotechnol.**

Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., Tsang, E.K., Karczewski, K.J., Maller, J.B., Kukurba, K.R., DeLuca, D.S., Fromer, M., Ferreira, P.G., Smith, K.S., Zhang, R., Zhao, F., Banks, E., Poplin, R., Ruderfer, D.M., Purcell, S.M., Tukiainen, T., Minikel, E.V., Stenson, P.D., Cooper, D.N., Huang, K.H., Sullivan, T.J., Nedzel, J., GTEx Consortium, **GEUVADIS Consortium**, Bustamante, C.D., Li, J.B., Daly, M.J., Guigo, R., Donnelly, P., Ardlie, K., Sammeth, M., Dermitzakis, E.T., McCarthy, M.I., Montgomery, S.B., Lappalainen, T.,

MacArthur, D.G. (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. **Science.**

Publications as part of the ICGC consortium

Richter, J., Schlesner, M., Hoffmann, S., Kreuz, M., Leich, E., Burkhardt, B., Rosolowski, M., Ammerpohl, O., Wagener, R., Bernhart, S.H., Lenze, D., Szczepanowski, M., Paulsen, M., Lipinski, S., Russell, R.B., Adam-Klages, S., Apic, G., Claviez, A., Hasenclever, D., Hovestadt, V., Hornig, N., Korbel, J.O., Kube, D., Langenberger, D., Lawerenz, C., Lisfeld, J., Meyer, K., Picelli, S., Pischimarov, J., Radlwimmer, B., Rausch, T., Rohde, M., Schilhabel, M., Scholtysik, R., Spang, R., Trautmann, H., Zenz, T., Borkhardt, A., Drexler, H.G., Möller, P., MacLeod, R.A.F., Pott, C., Schreiber, S., Trümper, L., Loeffler, M., Stadler, P.F., Lichter, P., Eils, R., Küppers, R., Hummel, M., Klapper, W., Rosenstiel, P., Rosenwald, A., Brors, B., Siebert, R., **ICGC MMML-Seq Project** (2012). Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. **Nat Genet.**

CONFERENCE CONTRIBUTIONS AS FIRST AUTHOR

Young Investigator Research Day of the CRC 1182 (06/2018, Kiel), Esser, D., Best, P., Kaleta, C., Elucidating the metabolic host-microbiome crosstalk based on transcriptome and metagenome sequencing (talk + poster)

SysBio2018 (03/2018, Innsbruck), Esser, D., Kaleta, C., Investigation of comorbidity and therapy response using transcriptional signatures across chronic inflammatory diseases (talk + poster)

KLS student conference (11/2017, Kiel), Comorbidity and therapy response prediction across chronic inflammatory diseases (talk)

e:Med meeting on Systems Medicine (11/2017, Göttingen), Esser D., Kaleta, C., Investigation of comorbidity and therapy response using transcriptional signatures across chronic inflammatory diseases (poster)

Plant Science Conference 2017 (09/2017, Kiel). Transcriptomic alterations during aging reflect the shift from malignant to degenerative diseases in the elderly (invited talk)

e:Med meeting on Systems Medicine (10/2016, Kiel), Esser, D., Kaleta, C., Cross-disease analyses reveal common transcriptomic signatures of inflammatory phenotypes (talk + poster)

Summer School of the cluster of excellence “Inflammation at Interfaces” (07/2016, Timmendorfer Strand), Cross-disease analyses reveal common transcriptomic signatures of inflammatory phenotypes (talk)

Summer School of the cluster of excellence “Inflammation at Interfaces” (10/2015, Weißenhäuser Strand), Exomic and transcriptomic characterization of colitis-associated carcinoma in mice (talk)

Illumina User Meeting (06/2015, Hamburg), Whole exome and transcriptome sequencing of murine colitis-associated colorectal cancer (invited talk)

Studierendentagung zu den Life Sciences in Kiel (12/2014, Kiel), Untersuchung molekularer Mechanismen von entzündungsassoziierten kolorektalen Tumoren (talk)

64. Annual meeting of the "American Society of Human Genetics" (10/2014, San Diego). Esser D., Falk-Paulsen M., Aden K., Rosenstiel P., Exomic and transcriptomic patterns of colitis-associated carcinoma (poster)

Summer School of the cluster of excellence “Inflammation at Interfaces” (10/2014, Schleswig), Molecular signatures of colitis-associated colorectal cancer (talk)

Studierendentagung zu den Life Sciences in Kiel (12/2013, Kiel). Genomische Signaturen von Magenkarzinomen (talk)

63. Annual meeting of the "American Society of Human Genetics" (10/2013, Boston). Esser D., Haag J., Holze N., Schreiber S., Rosenstiel P., Röcken C., Genomic landscape of two gastric cancer cases (poster)

Annual meeting of the "European Society of Human Genetics" (06/2013, Paris). Esser D., Haag J., Holze N., Schreiber S., Rosenstiel P., Röcken C., Whole genome and whole exome sequencing of gastric cancer samples (poster)

Workshop "Theoretical Biology" (03/2013, Plön). Deciphering gastric cancer genomes (invited talk)

5. NGFN annual meeting in the program of Medical Genome Research (12/2012, Heidelberg), Esser D., Haag J., Holze N., Schreiber S., Rosenstiel P., Röcken C., Whole genome and whole exome sequencing of gastric cancer samples (poster)

10 Danksagung

Mein herzlicher Dank geht an viele großartige Menschen, ohne deren Unterstützung diese Arbeit nicht möglich gewesen wäre. Mein besonderer Dank gilt …

- Herrn Professor Philip Rosenstiel, der die Bearbeitung dieses interessanten Themas überhaupt erst ermöglicht und hervorragende Arbeitsbedingungen geschaffen hat. Besonders danken möchte ich ihm für das stetige Vertrauen in meine Arbeit und die damit verbundenen sehr vielen Freiheiten, um meine eigenen Ideen umsetzen und eigenständig arbeiten zu können. Danken möchte ich ihm auch für die stets zielführenden wissenschaftlichen Diskussionen, die Unterstützung meiner Arbeit durch sein außergewöhnlich umfangreiches Wissen und die Förderung einer effektiven wissenschaftlichen Arbeitsweise.
- Herrn Professor Thomas Roeder für die Bereitschaft, das Zweitgutachten zu übernehmen, sowie die entspannte und angenehme Atmosphäre bei der Kooperation.
- Herrn Professor Hinrich Schulenburg für die Bereitschaft, das Zweitgutachten zu übernehmen, sowie die erfolgreichen und konstruktiven Kooperationen.
- Herrn Professor Christoph Kaleta für die gute Zusammenarbeit in den letzten Jahren und die Motivation meine Promotion trotz etlicher Rückschläge fertigzustellen. Weiterhin möchte ich mich bei ihm für die fachliche sowie persönliche Unterstützung und Förderung bedanken.
- Herrn Professor Hauke Busch für das hervorragende Mentoring und die vielen wertvollen Ratschläge
- Professor Andre Franke für seine konstruktive Unterstützung, Motivation und sein offenes Ohr bei allen Problemen
- Herrn Professor Röcken für die erfolgreiche Zusammenarbeit in dem gemeinsam publizierten Magenkarzinom- Projekt
- Dr. habil. Friederike Flachsbart für die fachlichen und persönlichen Ratschläge, die Kommentare zu einigen Abschnitten meiner Dissertation sowie für die vielen schönen Momente bei der Arbeit und privat z. B bei den gemeinsamen Surfsessions
- Dr. Ateequr Rehman für die wissenschaftlichen und persönlichen Gespräche, hilfreichen Ratschlägen sowie seiner ansteckenden Begeisterung für Wissenschaft und Forschung, die mich stets motiviert hat.

- Dr. Maren Falk-Paulsen für stetige Unterstützung bei allen Maus-Präparationen sowie für die vielen fachlichen Diskussionen
- Dr. Matthias Barann, Dr. Jan Kuiper und Dr. Robert Häsler für die vielen sehr fruchtbaren wissenschaftlichen Diskussionen
- Lena Best für die Freundschaft, die sich während meiner Promotionszeit entwickelt hat
- Katharina Göbel für die vielen hilfreichen Tipps im Labor und die Einweisung in die Geheimnisse der DNA- und RNA-Extraktion
- Dr. Jaydeep Bhat und Dr. Silvio Waschina für die vielen Kommentare zu einigen Abschnitten meiner Dissertation
- Maren Reffelmann für die histologischen Arbeiten
- Dr. Konrad Aden für die Durchführung der Koloskopien
- Dem NGS-Labor des Instituts für Klinische Molekularbiologie für die zuverlässige Sequenzierung aller Proben
- Allen Mitarbeitern des Instituts für Klinische Molekularbiologie und den Mitgliedern der Arbeitsgruppe „Medizinische Systembiologie“ am Institut für Experimentelle Medizin für die angenehme und freundschaftliche Arbeitsatmosphäre
- Allen, die meine Arbeit finanziell unterstützt haben, insbesondere dem Exzellenzcluster Entzündungsforschung sowie dem EU-Konsortium Geuvadis.
- Meinen Eltern, die mich unermüdlich gestärkt, liebevoll unterstützt und immer an mich geglaubt haben. Sie haben mich in allen guten und schwierigen Phasen meines Lebens begleitet, waren auch immer für mich da und haben in jeglicher Hinsicht den Grundstein für meinen Weg gelegt.
- Allen Freunden und Verwandten außerhalb des IKMB und IEM, für die ich oftmals nicht genug Zeit hatte. Ich möchte ihnen besonders für das Verständnis sowie für die Freundschaft und die moralische Unterstützung danken.

11 Eidesstattliche Erklärung

Hiermit erkläre ich, Daniela Esser, an Eides statt, dass die vorliegende Abhandlung - abgesehen von der Beratung durch meinen Betreuer, Prof. Dr. Philip Rosenstiel - nach Inhalt und Form meine eigene Arbeit ist. Die Arbeit wurde weder ganz noch zum Teil an anderer Stelle im Rahmen eines Prüfungsverfahrens vorgelegt. Auszüge dieser Arbeit wurden bereits in Fachzeitschriften oder auf Posterbeiträgen veröffentlicht oder zur Veröffentlichung eingereicht (siehe Publikationsliste). Die Arbeit ist unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden. Es wurde mir kein akademischer Grad entzogen.

Kiel, den _____ _____

Daniela Esser