

# **Identification of Functional Genetic Variants in Inflammatory Bowel Disease by Genome and Transcriptome Sequencing**

Dissertation  
zur Erlangung des Doktorgrades  
der Mathematisch-Naturwissenschaftlichen Fakultät  
der Christian-Albrechts-Universität zu Kiel

vorgelegt von  
**Matthias Barann**

August, 2012  
Kiel, Deutschland



Referent/in: Prof. Dr. Philip Rosenstiel

Korreferent/in: Prof. Dr. Dr. h.c. Thomas Bosch

Tag der mündlichen Prüfung: 31.10.2012

Zum Druck genehmigt: 31.10.2012

gez. Prof. Dr. Wolfgang J. Duschl, Dekan

**Parts of this dissertation are contained in the following two manuscripts.**

Rosenstiel R<sup>†</sup>, **Barann M**<sup>†</sup>, Klostermeier UC, Sheth V, Ellinghaus D, Rausch T, Korbel J, Nothnagel M, Krawczak M, Gilissen C, Veltman J, Forster M, Stade B, McLaughlin S, Lee CC, Fritscher-Ravens A, Franke A<sup>†</sup>, Schreiber S<sup>†</sup>. *Whole Genome Sequence of a Crohn disease trio*

**Barann M**, Klostermeier UC, Esser D, Ammerpohl O, Siebert R, Sudbrak R, Lehrach H, Schreiber S, Rosenstiel R. *Janus – Investigating the two faces of transcription*

# Contents

<b>1. Introduction.....</b>	<b>7</b>
1.1. Pathogenesis of inflammatory bowel diseases .....	8
1.2. Physical barrier function of the gut .....	11
1.3. Pathogen recognition.....	12
1.4. Complement System.....	13
1.5. Autophagocytosis .....	14
1.6. Unfolded protein response.....	15
1.7. Proteasome.....	16
1.8. Major histocompatibility complex and antigen presentation .....	17
1.9. Cytokines .....	18
1.10. Next generation sequencing as diagnostic tool for clinical application.....	20
1.11. RNA sequencing .....	21
<b>2. Aims of the study.....</b>	<b>23</b>
<b>3. Material and methods.....</b>	<b>24</b>
3.1. Extraction of nucleic acids from blood samples.....	24
3.2. Library generation.....	24
3.3. Sequencing.....	25
3.4. Sequence analysis .....	26
3.5. SNV annotation .....	27
3.6. Detection of <i>de novo</i> SNVs .....	28
3.7. Analysis of short structural variations (sSVs) .....	28
3.8. Analysis of long structural variations (LSVs) .....	29
3.9. Demographic origin of sequenced subjects.....	29
3.10. Single nucleotide variant distribution along chromosomes.....	29
3.11. STRING network analysis of missense single nucleotide variants .....	29
3.12. Selection of differentially expressed transcripts.....	30
3.13. Regions of homozygosity.....	30
3.14. Calculation of child's CD risk relative to its parents.....	31
3.15. SNV verification by Sanger sequencing.....	31
3.16. Identification of sense/antisense pairs in transcriptomic data.....	31
<b>4. Results .....</b>	<b>38</b>
4.1. General mapping and variant calling statistics .....	38
4.2. Verification of <i>de novo</i> SNVs by Sanger sequencing.....	39

4.3 Geographic origin determination based on genetic variants .....	43
4.4. Genetic variants in genes associated with monogenic phenocopies of Crohn's disease....	44
4.5. Genetic variants with known CD-association and other variants in the associated genes.	45
4.6. Genetic variants in genes associated with other inflammatory diseases .....	52
4.7. SNVs predicted to be damaging including all genes .....	53
4.8. Identification of mutational hotspots .....	55
4.9. Pathway analysis of genes affected by missense variants in the child .....	56
4.10. Expression changes in the child compared to the parents and genetic variants in the respective genes.....	58
4.11. Strand specific transcriptome analysis.....	58
<b>5. Discussion .....</b>	<b>65</b>
5.1. Sequencing and variant calling performance.....	65
5.2. Investigation of <i>de novo</i> variants .....	66
5.3. Genetic variants concerning monogenic phenocopies of Crohn's disease .....	66
5.4. Crohn's disease associated risk variants.....	66
5.5. Regions of homozygosity in relation to known Crohn's disease risk loci .....	68
5.6. Genetic variants associated with other inflammatory diseases .....	68
5.7. Everything else - the remaining genetic variability .....	69
5.8. Differential expression and genetic variants in differentially expressed genes .....	73
5.9. Strand specific transcriptome analysis .....	75
5.10. Conclusions .....	77
5.11. Perspective .....	79
<b>6. Summary (English) .....</b>	<b>79</b>
<b>7. Summary (German) .....</b>	<b>80</b>
<b>8. References .....</b>	<b>82</b>
<b>9. List of Figures .....</b>	<b>91</b>
<b>10. List of Tables.....</b>	<b>92</b>
<b>11. List of abbreviations .....</b>	<b>93</b>
<b>12. Statement of authorship .....</b>	<b>95</b>
<b>13. Curriculum vitae .....</b>	<b>96</b>
<b>14. Publications .....</b>	<b>97</b>
<b>16. Acknowledgements .....</b>	<b>99</b>
<b>17. Supplementary Material.....</b>	<b>100</b>

## 1. Introduction

Ever since the discovery that DNA is the carrier of genetic information in 1944 by Oswald Avery and colleagues<sup>1</sup> and the revelation of the DNA structure in 1953 by Watson, Crick<sup>2</sup> and Franklin, humanity searched for changes in DNA that explain phenotypical traits, especially those that cause diseases. This search was simplified around 1975 as the Sanger DNA-Sequencing method was established, that allowed to obtain the nucleotide sequence of DNA molecules, which was used to generate the sequence of the complete human genome for the first time. It took about ten years and millions of dollars before the first two drafts of the human genome were published in 2001. With the advent of next generation sequencing in 2005, the ability to sequence whole genomes changed dramatically. Today, sequencing of whole genomes only takes days and costs some ten thousands of dollars. Additionally, the already published human (reference) genomes allow for an easy identification of genetic variation, which has been used to identify disease causing variants. So far, next generation sequencing has most notably helped to find the genetic cause of multiple Mendelian diseases and chances are that the genetic cause of most Mendelian diseases will be discovered within the next decade.<sup>3</sup>

In my thesis, I focus on two related aspects that aim to identify functional genetic variants underlying inflammatory bowel disease (IBD) as a complex trait by next-generation sequencing. In the first part of the study, next generation sequencing was used to analyze the whole genome and transcriptome of a family trio with a severe early onset case of Crohn's disease. Crohn's disease is a subentity of IBD in which more than 70 risk loci have been identified using genome-wide association studies. Elaborating on this knowledge of immune and non-immune related pathways that are linked to etiology, I developed an algorithm that allows interrogating the different layers of information comprised in the genome sequence data from SNVs to larger structural variants. The introduction will focus on the pre-existing genetic knowledge on IBD and describe molecular mechanisms (e.g. autophagy, innate immunity) that are disturbed in the diseases.

The second part of the thesis focuses on the transcriptome as a key element linking genome and function. The establishment of comprehensive RNA expression profiles, including information on expression levels, alternative splicing, allele specific expression patterns and mRNA editing is essential to our understanding of the molecular basis of a disease. In the Introduction, concepts for next generation sequencing of transcriptomes and the link between transcriptional landscapes and potential disease mechanisms will be presented.

## **1.1. Pathogenesis of inflammatory bowel diseases**

Chronic and progressive inflammatory diseases affecting the gastrointestinal tract, such as Crohn's disease (CD) and ulcerative colitis (UC), are linked to a breach of the intestinal mucosal barrier and defects in the innate and adaptive immune system. Among these defects are an impaired macrophage activation,<sup>4</sup> multiple genetic variants in receptors of the innate immune system such as NOD-like receptors (NLRs, i.e. *NOD2* in CD), and defects in Toll-like receptors (TLRs, i.e. *TLR4* in UC)<sup>5</sup> (Figure 1). Crohn's disease and ulcerative colitis are clinically distinct diseases; however both diseases have common symptoms, i.e. both can lead to anemia due to gastrointestinal bleeding, fever, abdominal pain, malabsorption and weight loss, joint pain and diarrhea (often bloody in UC). CD is commonly found in the ileum but can affect any part of the digestive tract and may manifest in inflammatory patches. The colon wall may be thickened and uneven, and the ulcers are deep, sometimes extending into all layers of the wall. UC is usually restricted to the large intestine and manifests continuously across the affected areas, the colon wall is thinner and the mucus may have ulcers that do not extend beyond the inner lining.

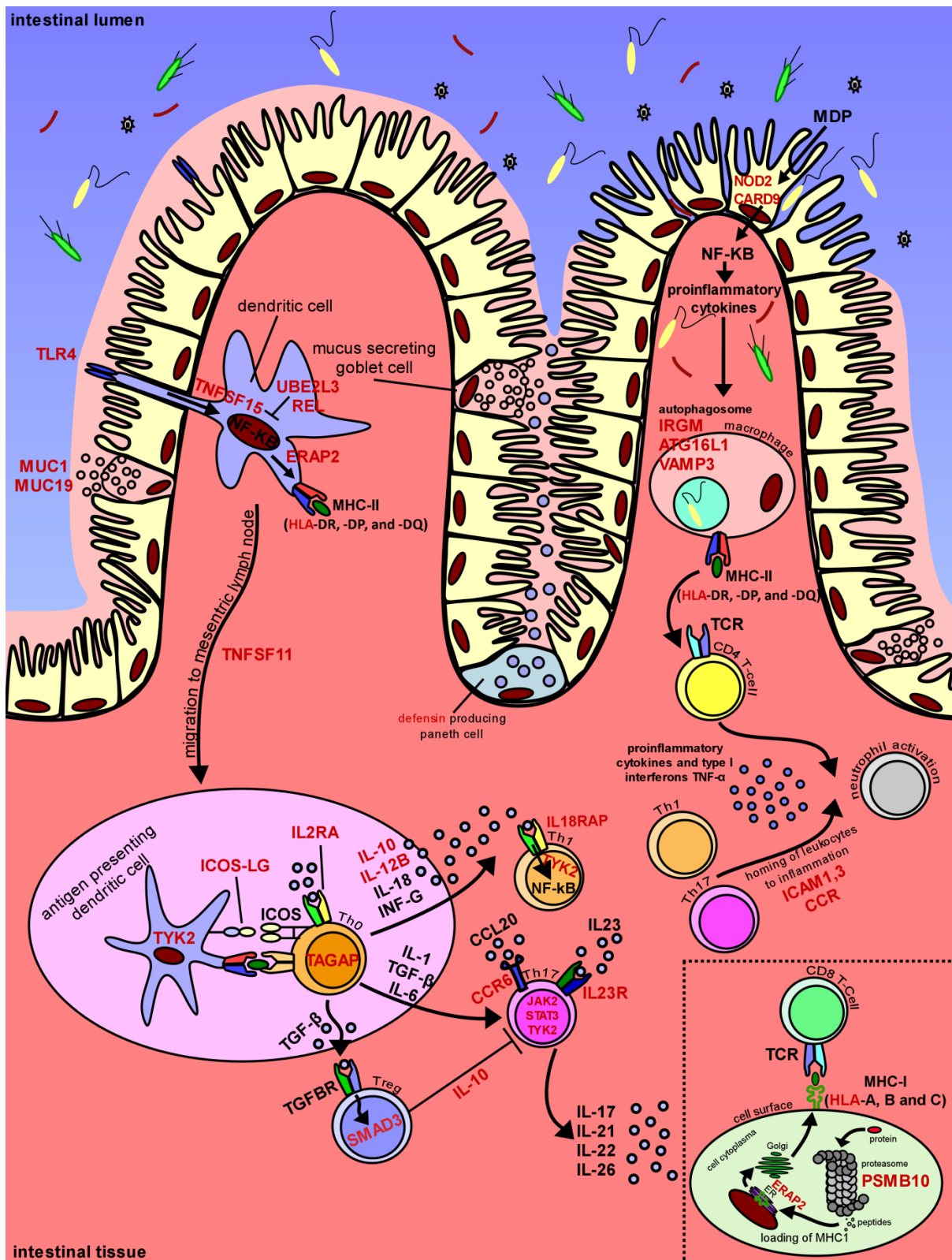
Studies have shown an increase of CD cases in distant, unrelated populations in the last few years, suggesting this to be a global trend.<sup>6</sup> One study showed a slight decrease of pediatric IBD incidence in a southwestern population of Ontario while in the same region the incidence of CD nearly doubled over the last decade.<sup>7</sup> A second study showed an increase of CD incidence in Korea between 2003 and 2008, where the mean annual incidence increased from 1.8 per 100,000 persons in 2003 to 2.7 per 100,000 persons in 2008.<sup>8</sup> Very recently a rapid increase of CD and UC cases was shown in the Irish population, where the number of cases increased 2-3 fold between 2001 and 2010.<sup>9</sup>

The genetic background of both diseases is different, although some risk genes are shared. As this work is focused on Crohn's disease, only the genetic background of CD will be mentioned in detail. Genetic factors are estimated to contribute to about 50-60% (sometimes up to 80%) of the risk to develop CD, but only about 20% of the genetic heritability has been identified yet.<sup>10,11,12</sup> More than 70 genes and loci have been associated with Crohn's disease. A large proportion of these loci have been identified by genome wide association studies (GWAS). Supplementary tables 1 and 2 show the genes and loci for CD included in the GWAS catalog ordered by highest odds ratio (<http://www.genome.gov/gwastudies/>). Most of these variants have been identified in individuals of western origin and studies have shown that risk variants detected in certain populations can not necessarily be transferred to other populations. *NOD2* is a prominent example of a susceptibility gene that differs between ethnicities. Genetic variants in *NOD2* have shown that *NOD2* plays a major role in CD pathogenesis among individuals of European ancestry, but this link could not be made for other populations. Recently, Adeyanju *et al.* (2012) studied the contribution of four known disease causing *NOD2* single nucleotide variants (SNVs) to CD in African Americans.<sup>13</sup> They



identified an association of one SNV with CD in African Americans, who are of about 80% West African and 20% European ancestry, but could not detect any known *NOD2* risk alleles in the West African reference genomes from the 1000 Genome project. They suggest that the association is a result of recent European admixture, i.e. by common children of West Africans and Europeans carrying the risk allele.

As mentioned above, the risk to develop CD can only be attributed by 50-60% to genetic variation. The remaining risk is explained by environmental factors, such as smoking, nutrition, composition of the bacterial flora and maybe other unknown factors. The probably most important environmental factor in CD is the composition of the intestinal bacterial flora. The intestinal epithelial cells are, separated by a mucus layer, permanently in close contact to potential pathogenic substances from the intestinal lumen. Under normal circumstances, a healthy host and its intestinal microflora have a symbiotic relation. Alterations in the microfloral composition however can turn the symbiosis into malignancy, causing a disease phenotype. Strong evidence for the involvement of intestinal bacteria in CD is provided by the observation, that mice of numerous IBD models do not develop disease after germ-free rederivation.<sup>11,14</sup> The presence of crosstalk between the host and microbiome is evidenced by numerous experiments. In 2008, Vaishnava *et al.* showed that Paneth cells sense enteric bacteria through toll-like receptor activation regulating the expression of antimicrobial molecules.<sup>15</sup> Further, Umesaki *et al.* demonstrated that the inoculation of germ-free mice by oral administration of a fecal suspension induces or increases the expression of T-cell receptors on T-cells.<sup>16</sup> Thus the presence or absence of microbes triggers responses of the host, that in turn can change the microflora. It is not conclusively determined if changes of the microflora cause IBD or if an individual's genetic background causes IBD and promotes alterations of the microflora. Most likely both play a role in IBD. It is therefore not surprising that many mechanisms impaired in CD affect microbial defense mechanisms. Genes associated with CD are generally involved in physical barrier function (i.e. mucins), pathogen recognition (TLRs, NLRs), antigen processing (autophagosome, proteasome, and MHC antigen presentation), B- and T-cell function and cytokine production.



**Figure 1** Modified schematic for Crohn's disease associated genes (red) and gene interactions according to (Fransen *et al.* 2011).<sup>17</sup>

Defects in the physical barrier, the mucosa (*MUC1*, *MUC19*) and epithelial cells (*ITLN*) lead to an increased number of intruding pathogens. Pattern recognition receptors, like the NOD-like receptors (NLRs, *NOD2*) and toll-like receptors (TLRs, *TLR4*) are activated and lead to a downstream activation of NF-κB. *CARD9* cross links the NLR/TLR signaling cascade with the signaling cascade of the ITAM (immunoreceptor tyrosine-based activation motif) tyrosine kinase pathway (i.e. induced by activation of members of the lectin family, such as the dectin-1 or CD16 receptors (not shown here)). Mediated by *CARD9*, both pathways can activate the effectors of the other, which in case of the TLR/NLR pathway is NF-κB activation and in the case of the ITAM-

TK pathway the induction of MAPK activation.<sup>18</sup> NF- $\kappa$ B signaling induced by pathogen recognition leads to production of pro-inflammatory cytokines and migration of antigen presenting cells into the intestinal mesenteric lymph node, where T-cell proliferation and differentiation is stimulated involving the IL-2 receptor (*IL2RA*) and the T-cell activation RhoGTPase activating protein (*TAGAP*). Interaction of ICOS-LG with ICOS enhances T-cell proliferation and IL-2 dependent cytokine production.<sup>19</sup> T-helper (Th<sub>0</sub>) cells differentiate in a cytokine dependent manner into various T-cell subtypes. Activation of the IL-23 receptor promotes differentiation into Th<sub>17</sub> cells, which are involved in several immune-related diseases. The JAK-STAT-TYK pathway is induced and production of pro-inflammatory cytokines enhanced. Activation of *TYK2* in dendritic cells induces IL-12B production, which is a subunit of the IL-23 receptor, already mentioned above, and IL-12, which is a strong promoter for Th<sub>1</sub> differentiation. It has been suggested that IL-18 is also involved in the development of Th<sub>1</sub> cells and associations between CD and IL18 receptor proteins have been made.<sup>12,20</sup> The transforming growth factor beta (TGF- $\beta$ ) promotes differentiation of naïve T-cells into regulatory T-cells (T<sub>reg</sub>). However, in the presence of IL-6, TGF- $\beta$  also promotes differentiation into TH<sub>17</sub> cells.<sup>21</sup> *SMAD3* was identified as a CD risk gene.<sup>12</sup> It acts as a mediator of transcriptional activation by the TGF- $\beta$  receptor and can induce strong ligand independent TGF- $\beta$  like responses in synergy with *SMAD4*.<sup>22,23</sup> Both helper T-cell subtypes (Th<sub>1</sub> and Th<sub>17</sub>) are pro-inflammatory, while differentiation into T-regulatory cells has an anti-inflammatory effect by IL-10 secretion. The autophagosome is another important pathway involved in CD pathogenesis (*ATG16L1*, *VAMP3*, and *IRGM*). In the autophagosome, microbial intruders are degraded for antigen presentation to CD4<sup>+</sup> T-cells, which triggers the production of inflammatory cytokines and maintains inflammation. Both, the activation of CD4<sup>+</sup> T-cells and the differentiation of pro-inflammatory Th<sub>1</sub> and Th<sub>17</sub> cells can lead to the recruitment of neutrophils to the site of inflammation (*CCR*, *ICAM1*, *ICAM3*), which ultimately increases inflammation, and thus ulceration and additional microbial penetrance. One of the CD risk genes with higher odds ratio (1.44) is *PSMB10*. *PSMB10* is part of the immunoproteasome in cells which degrades proteins for presentation via MHC class I molecules to CD8<sup>+</sup> cytotoxic T-cells. Linked to this, *ERAP2* is required for trimming of the antigen precursor molecules. Summarized this demonstrates the high complexity of pathways involved in CD and the large number of genes which can be involved in CD pathogenesis.

A detailed overview of mechanisms underlying CD pathogenesis and genetic variants affecting these mechanisms is presented in the following chapters.

## **1.2. Physical barrier function of the gut**

The intestinal physical barrier is the first line of protection of the body against luminal bacteria. A breach of the physical barrier allows bacteria to intrude the underlying tissue and cause various diseases. Epithelial cells are closely connected by tight-junctions that prevent trafficking of molecules from the apical site (here the intestinal lumen) into the intercellular space and *vice versa*. The cell-cell connections are further stabilized by additional junctions, the desmosomes and adherence junctions that consist of transmembrane proteins of the cadherine family. Disruption of the physical barrier, i.e. due to loss of cell-cell connections is a hallmark of Crohn's disease. In 2009, a study by Muise *et al.* raised evidence that polymorphisms in the *CDH1* (E-cadherin) gene are associated with CD and cause a cytoplasmatic accumulation of truncated E-cadherin.<sup>24</sup>

The intestinal epithelial cells are covered by a continuous layer of mucus that protects the cells from potentially aggressive environmental stimuli. The mucus layer consists mainly of mucin, O-glycoproteins with high density that influence the viscosity of the mucus,<sup>25</sup> and that are produced and secreted by goblet cells within the intestinal epithelial layer.<sup>26</sup> About 20 mucin genes (*MUC1-MUC22*) have been identified so far that can be divided into two groups of soluble and membrane bound proteins. In CD patients, deep ulceration of the mucus layer was observed, which might be

caused by variants affecting the expression or function of mucin genes, or variants that change the mucus viscosity, for example by altered secretion. Indeed, genetic variants conferring a higher CD risk were identified in CD patients by GWAS in proximity of the mucin genes *MUC1* and *MUC19*.<sup>12</sup>

The epithelial cells below the mucus layer not only serve as a physical barrier, separating the host from the intestinal contents, but also secrete antimicrobial defensins into the mucus layer to keep microbes at bay. The defensins can be divided into two classes,  $\alpha$ - and  $\beta$ -defensins, with the latter being more broadly distributed along the epithelial, and the former mainly expressed by Paneth cells, typically located at the bottom of the intestinal crypts, and neutrophils.<sup>27</sup> Disturbances in the defensin mediated antimicrobial defense, often associated with altered Paneth cell differentiation or a diminished number of beta-defensin gene clusters, seem to be critical in IBD pathogenesis.<sup>28</sup>

In this context, GWAS identified a CD associated risk variant in intelectin (*ITLN*) that is strongly expressed in goblet and Paneth cells. *ITLN* is thought to exhibit antimicrobial properties, and thus, changes affecting *ITLN* expression or function could be involved in alteration of the mucus characteristics, as has been shown by experiments in sheep respiratory tract epithelium and intestinal nematode infections of resistant mice strains.<sup>29</sup>

Should microbes overcome the physical barrier, other defense mechanisms, i.e. microbial sensing, become important.

### **1.3. Pathogen recognition**

Pattern recognition receptors (PRRs) are part of the innate immune system and can be divided into Toll-like receptors (TLRs), NOD-like receptors (NLRs), RIG-like helicases (RLRs) and C-type lectins (CLRs). PRRs recognize two distinct pattern types. The first type includes the so called damage-associated molecular patterns or DAMPs. For example uric acid and  $\beta$ -amyloid belong into this category. The second types of patterns are of microbial origin and are thus named pathogen associated molecular patterns (PAMPs). Under normal conditions, PRRs are able to discriminate between harmless bacteria and potential pathogenic microbes, which could be due to recognition of specific PAMP combinations. It is assumed that perturbed recognition of these patterns contributes to etiopathogenesis of CD. As an archetype of this disturbed recognition, NOD2, a member of the NLR family, has been identified as one of the strongest CD susceptibility genes in European individuals.

In the intestine, NOD2 has been reported to influence the composition of bacteria in mice, possibly by reduced bacterial killing ability of the host,<sup>30</sup> which is in agreement with strong NOD2 expression in Paneth cells and the suggested link between NOD2 and antimicrobial  $\alpha$ -defensin secretion.<sup>31</sup> After stimulation with the muramyl dipeptide (MDP), a component of bacterial cell wall peptidoglycan, NOD2 is thought to undergo conformational changes that lead to the activation of downstream signal cascades involving RIP2, TRAF6 and NEMO that induce the production of pro-

inflammatory cytokines by activation of NF- $\kappa$ B and MAPKs. Multiple variants associated with an increased CD risk have been identified in NOD2. They are often affecting the leucine-rich-repeat (LRR) domain and interfere with NOD2's ability to recognize PAMPs leading to diminished signal transduction ("loss-of-function variants"). NLRs and TLRs are likely activated simultaneously during pathogen infection, which amplifies cytokine production. For example, upon activation of TLR2 to -5, TLR7, and TLR9 by their corresponding ligands secretion of TNF- $\alpha$ , IL-1, IL-8, and IL-10 is increased in the presence of MDP.<sup>11,32</sup> The role of NOD2 in CD has not been precisely defined, speculation is that NOD2 might be involved in the negative regulation of TLR signaling, promotion of IL-10 expression, regulation of inflammatory cytokine production by CD4<sup>+</sup> T-cells in a NF- $\kappa$ B dependent way and Paneth cell function by direct regulation of  $\alpha$ -defensin expression.<sup>11</sup>

#### **1.4. Complement System**

The complement system is part of the innate immune system and involved in pathogen clearance. In some cases, deficiencies in the complement system have been reported to be associated with Crohn's disease. The complement system includes a large number of plasma proteins that react with each other to opsonize pathogens and induce inflammatory responses. Several components of the complement system are proteases that are present as precursor enzymes (zymogen) which require proteolytic cleavage to become active. The inactive zymogens are widely distributed throughout body fluids and become activated at sites of infection. The cleavage of the precursor enzymes triggers a cascade where one complement enzyme cleaves a second complement enzyme which then cleaves another. This allows a large amplification of the complement response. Three pathways can activate the complement system. The first pathway, also called classical pathway, is dependent on the complement component C1q. C1q can either bind directly to the pathogen surface or to antigen:antibody complexes, which is a key link for effector cascades between the adaptive and innate immune system. Activation of the C1s subunit of C1q leads to cleavage of other complement components (C4 and C2). Based on one case, it has been suggested that C2 deficiency might predispose for inflammatory bowel disease.<sup>33</sup> However, Marks *et al.* argued that this 'link' probably happened by chance, as the occurrence of C2 deficiency is the most common hereditary complement deficiency (1 in 10,000 people).<sup>34</sup> Marks *et al.* also reported a number of studies which showed CD-like small bowel enteritis in the context of C1 esterase deficiency. The second pathway depends on mannan-binding (MB) lectins, a serum protein that binds mannose-containing carbohydrates that can be found on bacteria. Very recently, in 2011 Bak-Romaniszyn *et al.* showed that the number of individuals carrying a *MBL2* (*mannose-binding lectin*) variant causing *MBL* deficiency was significantly higher in CD patients than in normal controls or children with UC.<sup>35</sup> The third pathway, called alternative pathway, is triggered by binding of spontaneously activated complement components to pathogen surfaces. Activation of

all three pathways converges by generation of the protease C3 convertase. This protease is covalently bound to the pathogen surface where it cleaves C3 to generate large amounts of C3a and C3b. C3b is the main effector molecule of the complement system and opsonizes pathogens for destruction by phagocytes. C3b also binds to the C3 convertase to form a C5 convertase that produces C5a and C5b. C5a is the most important small peptide mediator of inflammation and is involved in recruitment and activation of phagocytes. C5b activates the 'late' events of complement activation, which leads to the formation of a membrane-attack complex by multiple terminal complement components. This complex integrates into bacterial membranes and generates pores which leads to osmotic destruction of the pathogens.<sup>36</sup>

According to the literature available, a decisive involvement of the complement system in CD pathogenesis cannot be shown. However, some existing indications suggest a role of the complement system in some cases of CD or CD-like phenotypes.

A much stronger association with CD has been made for another mechanism involved in pathogen clearance, namely autophagy, involving *ATG16L1*, *NOD2*, *LRRK2* and *VAMP3*.

### **1.5. Autophagocytosis**

Autophagy evolved as a mechanism for removal of damaged organelles, invading microbes and recycling of cellular compounds during cellular starvation. The unwanted contents is enclosed by a double-layered membrane and fused with lysosomes, which then leads to degradation of the contents. Proper autophagosome function is linked to removal and regulation of the inflammasome. The inflammasome is induced by inflammatory cytokines (i.e. TNF- $\alpha$ ) and signaling resulting from recognition of DAMPs and PAMPs. Activation of the inflammasome activates caspase-1 which mediates the cleavage of immature IL-1 $\beta$  and IL-18 into their mature forms and extracellular secretion.<sup>11</sup> Polymorphisms in *NLRP3* (NOD-like pyrin containing protein 3), a critical component for caspase-1 activation in the inflammasome, have been strongly associated with CD.<sup>37</sup> However, the very strong association could not be replicated in a follow-up study.<sup>38</sup> Another gene possibly involved in inflammasome function that has been associated with increased CD risk is *ATG16L1*. Mice deficient for *ATG16L1* show unregulated inflammasome function in response to combined TNF and TLR stimulation.<sup>11</sup>

The pattern-recognition receptor NOD2 has also been identified as a key-regulator of autophagy induction. Upon activation, NOD2 induces the formation of autophagic vesicles in epithelial and dendritic cells, whereas *NOD2* deficiency results in reduced autophagic uptake and impaired intracellular bacterial killing.<sup>11</sup> In a study using *Shigella flexneri*, the authors could show that NOD2 localizes to the site of invasion and mediates translocation of ATG16L1 to these spots, involving direct physical interaction of both molecules. Whereas NOD2 molecules carrying CD risk variants were unable to recruit ATG16L1 and showed impaired autophagy induction.<sup>39</sup> ATG16L1 molecules

carrying the CD risk variant T300A also impair autophagy induction upon MDP stimulation, raising further evidence for the involvement of ATG16L1 and NOD2 in the same pathway.<sup>11</sup> Other than for autophagy induction in epithelial cells, RIP2 has been shown to be mandatory for NOD2 dependent induction of autophagy in dendritic cells. An impairment of autophagy in dendritic cells might affect the adaptive immune function.<sup>39</sup>

Besides the above mentioned genes, GWAS identified multiple loci attributing to a higher CD risk in proximity of other autophagy related genes. These genes include *IRGM*, that has been shown to induce autophagy and establish large autolysosomal organelles for elimination of intracellular bacteria;<sup>40</sup> *VAMP3* that has been linked to membrane traffic and mediates the delivery of TNF- $\alpha$  from the recycling endosome to sites of phagocytic cup formation at the cell surface;<sup>41</sup> and *LRRK2*.<sup>12</sup> Experiments with siRNA induced knockdown of *LRRK2* lead to an increased autophagy activity and reduced apoptosis in starving conditions. Further evidence for the involvement of *LRRK2* in CD are increased levels of *LRRK2* in inflamed tissue of CD patients compared to noninflamed tissue.<sup>42,43</sup>

Both, the sensing of pathogens by PRRs and autophagy are closely connected to endoplasmic reticulum (ER) stress. Several studies have shown that the *unfolded protein response* (UPR), a consequence of ER stress, induces autophagy and that decreased autophagy can increase ER stress.<sup>11</sup>

### **1.6. Unfolded protein response**

The unfolded protein response (UPR) is a result of endoplasmic reticulum (ER) stress and has been shown to be involved in CD. The *X box binding protein-1* (XBP1) has been identified as critical mediator of ER stress, and presents the first UPR component that has been associated with CD risk.<sup>44</sup> Various genetic variants in *XBP1* have been associated with CD and deletions of *XBP1* in intestinal epithelial cells results in spontaneous inflammation of the small intestine and increased susceptibility to DSS induced colitis, involving functional and phenotypic impairment of Paneth cells and hypersensitivity toward inflammatory cytokines by increased JNK and NF- $\kappa$ B signaling.<sup>11,44</sup> Later, a second gene linked to UPR, namely *ORMDL3*, has been associated with CD by GWAS. *ORMDL3* is a membrane bound protein that regulates the Ca<sup>++</sup> uptake from the cytosol to the ER. Variations changing *ORMDL3* function might thus lead to protein misfolding.<sup>45,11</sup> A third gene, *AGR2*, that regulates protein folding in mice has been linked to CD. Decreased levels of *AGR2* mRNA are associated with an increased risk for CD and UC. Also, *AGR2* deficient mice develop spontaneous enteritis and colitis, and these mice show increased ER stress in the intestinal epithelial indicated by higher levels of GRP78 and XBP1 with no mucus in goblet cells and expansion of Paneth cells prior inflammation.<sup>11</sup>

These findings demonstrate the high relevance of ER stress in CD. ER stress is generated by accumulation of misfolded proteins, caused by genetic changes (change of amino acid sequence)

or environmental factors (i.e. salt concentration, temperature, inflammation, excess of iron). Cells whose main function is secretion of proteins are highly dependent on a normal UPR. Among these cells are immune cells (plasma cells, plasmacytoid dendritic cells, macrophages, and CD8<sup>+</sup> cytotoxic T-cells), parenchymal cells and in context of CD very importantly, intestinal epithelial cells, especially Paneth and goblet cells.<sup>11</sup> Invocation of UPR by ER stress involves three pathways, which sense changes in the ER and lead to the expression of transcription factors that influence a variety of factors for stress adaptation, such as translation, the quantity of ER membranes by lipid metabolism, expression of chaperones, other protein quality controls through ER-associated degradative pathways and possibly apoptosis. Three gene products known to regulate UPR, namely PERK, IRE1 and ATF6 are in an inactive state while they are bound to the ER-associated chaperone GRP78. Upon engagement of GRP78 with unfolded proteins, the associated proteins are released and initiate UPR specific pathways. Release of PERK causes phosphorylation of eIF2 and thereby attenuation of translation, except for the transcription factor ATF4 that induces expression of proapoptotic *CHOP*. ATF6 migrates to the Golgi, where it undergoes proteolytic cleavage and releases its cytoplasmic domain that enters the nucleus and regulates transcription. The kinase activity of released IRE1 leads to activation of JNK and NF- $\kappa$ B pathways. Also, IRE1 exhibits an endoribonuclease activity directed at the mRNA of *XBP1*. Only the spliced *XBP1* mRNA leads to a functional protein, as proteins from the unspliced form are lacking the DNA binding domain.<sup>11</sup>

### **1.7. Proteasome**

Disregulation of the immune cells' ability to present antigens can negatively influence the success of pathogen clearance. Vital components of antigen presentation are autophagy and the allocation of small peptides by the proteasome. Proteasomes are responsible for degradation of poly-ubiquitylated proteins in eukaryotic cells. A 20S core catalyzes the degradation while at least one 19S subunit is attached to one end that exhibits a regulatory role by recognition and unfolding of proteins before degradation. Many essential cell processes are regulated and maintained by the proteasome. Among these are cell growth, cell differentiation, cell signaling, DNA repair, gene transcription, apoptosis, and generation of antigenic peptides presented by major histocompatibility (MHC) class I cell surface molecules to CD8<sup>+</sup> T-cells.<sup>46</sup>

A second form of proteasomes was identified, called the immunoproteasomes (IP) that shows affinity for low molecular weight polypeptides (LMP). These LMPs were identified by co-precipitation with MHC class I molecules and have small sizes of about 20 to 30 kDa. Structurally both types of proteasomes are very similar but some proteins have been identified that are unique to the IP. Among these proteins are PSMB9 (LMP2) and PSMB8 (LMP7), which replace the  $\beta$ 1 and  $\beta$ 5 subunits of the standard proteasome (SP) 20S core respectively. Other than genes of other proteasome subunits that are spread across the genome, *PSMB8* and *PSMB9* are concentrated



within the MHC class II region, upstream of the TAP-transporter (*transporter associated with antigen processing*) associated genes *TAP1* and *TAP2*. *PSMB8* and *PSMB9* are, like other molecules in the MHC regions, polymorphic and IFN- $\gamma$  induces their expression. Incorporation of both proteins into the SP has been reported to influence the quantity and quality of peptides generated by degradation of a 25-mer polypeptide containing the IE pp89 L<sup>d</sup>-restricted YPHFMPTNL naturally processed dominant antigenic peptide.<sup>46,47</sup> It has been shown that *PSMB8* or *PSMB9* deficient mice exhibit impaired presentation of bacterial- or viral-derived epitopes to CD8<sup>+</sup> cytotoxic T-cells and show a reduced number of naïve CD8<sup>+</sup> T-cells.<sup>46,48,49</sup>

GWAS identified a variant in a related proteasome component (*PSMB10*) that replaces the  $\beta$ 2 subunit of the SP in the IP, and is associated to confer a higher CD risk.<sup>46,50</sup> Additionally, *PSMB9* and *PSMB10* were found to be up-regulated during Crohn's disease pathogenesis,<sup>46</sup> and variants in *UBE2L3*, a gene involved in ubiquitination of the NF- $\kappa$ B precursor protein (p105) with possibly cytotoxic NK function, has been suggested to confer a higher CD risk.<sup>12,51</sup> Impaired proteasome function might therefore be involved in CD pathogenesis.

The immunoproteasomes are mainly found in hematopoietic cells like macrophages, T-cells and activated and resting B-cells. There is some evidence of IPs being expressed in dendritic cells, but current data on their role in DCs is not decisive. Ebstein *et al.* showed that in dendritic cells *PSMB8* to *10* expressions were only inducible by activation of TLRs, i.e. by LPS stimulation, which leads to an IFN- $\beta$  autocrine loop.<sup>46</sup>

*PSMB8* expression was also detected in other cells, i.e. in small intestine epithelial cells and colon. It has been suggested that IPs are more efficient in I $\kappa$ B degradation than SPs, and thus lead to an increased translocation of NF- $\kappa$ B into the nucleus. Disruption of the *PSMB8* subunit inhibits the release of pro-inflammatory cytokines like IL-23, TNF- $\alpha$  and IL-6 from peripheral blood mononuclear cells (PBMCs),<sup>46</sup> highlighting the potential involvement of the immunoproteasome in inflammatory processes.

The antigens provided by the proteasome are finally presented to immune cells by the components of the major histology complex. An impairment of the ability to present antigens could also contribute to a disease phenotype and is therefore discussed below.

### **1.8. Major histocompatibility complex and antigen presentation**

Genetic variants occurring in genes located in the *major histocompatibility complex* (MHC) regions seem to be of high importance in Crohn's disease pathogenesis. The MHC is located on chromosome 6 and spans 4-7 million base pairs containing about 200 genes.<sup>36</sup> Among these genes are the *Human Leukocyte Antigen* (HLA) genes (MHC I and MHC II), genes involved in proteasome function (MHC II) and genes of the complement system and cytokines (MHC III), like TNF- $\alpha$  and - $\beta$  which are strong mediators of apoptosis. The link between cytokines and CD will be explained in

the next chapter. Here, I will focus on the role of the HLA genes in antigen presentation and CD. Multiple studies tried to identify variants in the HLA genes that contribute to CD risk. In previous studies, four class II alleles (*HLA-DRB1\*07*, *HLA-DRB1\*0103*, *HLA-DRB1\*04*, and *HLA-DRB3\*0301*) show reproducible associations with CD.<sup>52</sup> Additionally, a study by Zhang *et al.* in 2011 reported that the *HLA-Cw\*12* allele might confer a higher CD risk, based on a study of 73 CD patients and 106 healthy controls.<sup>53</sup> Therefore, the HLA genes might play an important role in CD pathogenesis. The HLA genes can generally be categorized into three class I  $\alpha$ -chain genes (*HLA-A*, *-B* and *-C*) and three pairs of MHC class II  $\alpha$ - and  $\beta$ -chain genes (*HLA-DR*, *-DP*, and *-DQ*). Other than the mentioned  $\beta$ -chain genes, the  $\beta_2$ -microglobulin (*B2M*) gene is located on chromosome 15 and the invariant chain (*CD74*) on chromosome 5. MHC class I molecules consist of one  $\alpha$ -chain and one  $\beta_2$ -microglobulin chain. The  $\alpha$ -chain is divided into three domains ( $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$ ). The  $\alpha_3$  part contains a transmembrane domain and works as the anchor of the protein. The antigen is presented in a binding groove formed by the  $\alpha_1$  and  $\alpha_2$  domains that is stabilized by the  $\beta_2$ -microglobulin chain. MHC class I molecules present antigens from pathogens within a cell to CD8<sup>+</sup> cytotoxic T-Cells (i.e. during viral infection), which subsequently kill the infected cells. MHC class I molecules are highly expressed on all nucleated immune cells like T- and B-Cells, macrophages and other antigen presenting cells (i.e. dendritic cells) and neutrophils. Other nucleated cell types express MHC class I molecules on a low to moderate level and red blood cells do not show MHC class I molecules. MHC class II molecules consist of one  $\alpha$ -chain and  $\beta$ -chain. Both chains contain two domains, with the  $\alpha_2$  and  $\beta_2$  parts having a transmembrane domain each and  $\alpha_1$  and  $\beta_1$  forming the binding groove. MHC class II molecules interact exclusively with CD4<sup>+</sup> (helper) T-Cells and are highly expressed on B-cells, macrophages, other antigen presenting cells and in the epithelial cells of the thymus. Other cell types do not express MHC class II molecules, except for T-cells when they are in an activated state. B-cells presenting antigens via MHC class II are stimulated to produce antibodies when encountered by CD4<sup>+</sup> T-cells. Similar CD4<sup>+</sup> T-cells stimulate macrophages presenting antigens via MHC class II to destroy the pathogens in their vesicles (see chapter about autophagy). Besides the functional differences of the two MHC class molecules, they also differ in the size of the peptide they can bind. The binding groove of MHC class I molecules is not as open for antigen binding than the binding groove of class II molecules. As a result, MHC class I molecules binds smaller peptides of 8 to 10 amino acids by both ends, while class II molecules bind peptides which contain more than 13 amino acids.

### **1.9. Cytokines**

A broad range of genes related to cytokines and cytokine targets (i.e. *IL2RA*, *IL10*, *IL12B*, *IL18RAP*, *IL19*, *IL23R*, *IL27*, *CCR6*, *CCL2*, *CCL17*, *CPAMPD8*, *TNFSF11*, and *TNFSF15*) have been identified as potential Crohn's disease risk genes. One of the strongest CD risk genes is the gene coding for

IL23R that will be discussed in more detail later. Cytokines can be separated into three major structural families: members of the TNF family, members of the hematopoietin family and the chemokines. Members of the TNF family are involved in adaptive and innate immune function, most of them being soluble molecules. The hematopoietin family consists mainly of growth hormones and interleukins, which are involved in both parts of the immune system. The chemokines again can be divided into two major classes differentiated by different motifs at the amino terminus. The first class contains the CC chemokines, with the Cs representing adjacent cysteines near the amino terminus. CC-chemokines promote the migration of monocytes and other cell types from the blood stream into the surrounding tissue and the infiltration of leukocytes (T-effector cells) into tissues. The other class contains the CXC chemokines, in which an additional amino acid is introduced between the two cysteines at the amino terminus. These chemokines include two subgroups, members of the first group contain a Glu-Leu-Arg (ELR) motif before the first cysteine that promote migration of neutrophils and the members of the second group that are lacking this motif are usually involved in homing of lymphocytes, like effector T-cells.

Explaining the functions of cytokines involved in CD pathogenesis would be beyond the scope of this work, thus only IL2RA and IL23R will be exemplified.

GWAS identified the receptor of IL-2 (IL2RA) and *TAGAP* (that is co-regulated with *IL2* during T-cell activation) as potential CD risk genes. Interleukin-2 (IL2) is produced by lectin- or antigen-activated T-cells and acts as a growth hormone for both, B- and T-cells and might be involved in programming the development of CD8<sup>+</sup> memory T-cells.<sup>12,54,55,56</sup>

One of the strongest risk genes for CD is the gene coding for the IL-23 receptor (*IL23R*). IL-23 does not bind to IL-23R directly, but to IL12-RB1, which pairs with IL23-R in cells co-expressing both molecules. IL-23 is a composite molecule of the IL-12 p40 (IL12B) subunit and IL-23A (p19). Variants detected in *IL12B* have also been associated with an increased CD risk. IL-23, together with IL-6 and IL-1 $\beta$ , mediates generation of Th<sub>17</sub> cells by initiation of IL-17 production in naïve precursor cells independently of TGF- $\beta$ . Additionally, IL-23 enhances IFN- $\gamma$  secretion by memory T-cells in an IL-2 dependent manner.<sup>12,57,58,59</sup>

In this context, variants in *BCL3* might also be associated with a higher CD risk. Impaired *BCL3* expression leads to an increase of IL-23 production in dendritic cells stimulated with bacterial lipopolysaccharides of *IL10* deficient mice, which might lead to an increased adaptive immune response in CD patients that show lower expression of *BCL3*.<sup>51</sup>

As shown in the above chapters, Crohn's disease is a complex disease involving numerous genes and mechanisms. Although each CD associated gene and CD risk loci identified by GWAS could be investigated by conventional methods, such as PCR, chances are high to miss disease relevant variants in yet unknown CD risk genes, as only about 20% of the genetic heritability is explainable

by the known risk loci. Therefore a more complete assessment of the genetic variation in the individuals of the presented case trio is desirable. To accomplish this, next generation sequencing was used to sequence the whole genome of all three individuals that allows, in theory, the identification of all genetic variants.

### **1.10. Next generation sequencing as diagnostic tool for clinical application**

Next generation sequencing (NGS) is still a young technology. The first commercial machine was launched in 2005 by 454 Life Science (GS20, Roche) and since then several companies started selling their own sequencers. Currently, the most popular sequencers are from Illumina (HiSeq2000, miSeq) and Applied Biosystems (SOLiD, Ion Torrent/Proton), generating a large amount of sequencing data (200 Gb per run HiSeq2000; 90 Gb per run SOLiD 5500) for moderate costs, basically allowing to sequence one whole human genome in about 3-4 days in a single run with a genome-wide coverage depth of 30x (See the material and methods section for the workflow of SOLiD sequencing). In early 2011, Kahn illustrated how rapid the development of NGS proceeds. Even Moore's Law, a rule of thumb for the inexpensive doubling of transistors on an integrated circuit in a two year period falls far behind. The first GS20 generated roughly 200 Mb per day in 2005, current NGS machines generate 10-20 Gb per day.<sup>\*,†,60</sup>

In the past few years the cost of whole genome sequencing (WGS) decreased from several millions to less than 50.000\$ for one genome in 2010, about half as much today (2012) and is still dropping.<sup>61</sup> This drop in costs makes WGS more feasible in disease diagnostics that can be already noticed by the emerging number of published disease related genomes and exomes.<sup>62-72</sup>

Several studies were successful in the discovery of causative genes in Mendelian diseases, here exemplified by two studies. In an exome sequencing approach, Ng *et al.* were able to identify the causative gene responsible for Miller syndrome.<sup>66</sup> In a different study, whole genome-sequencing of a patient with metachondromatosis, revealed *PTPN11* as the likely disease gene.<sup>71</sup> Recently, Green *et al.* predicted, probably also encouraged by the high success rate in the identification of Mendelian disease genes by next generation sequencing that most Mendelian disease genes will be identified over the next decade.<sup>3</sup>

However, using NGS to solve the genetic basis of complex diseases proved, unsurprisingly, to be much more difficult. Baranzini *et al.* studied multiple sclerosis in discordant monozygotic twin pairs but could not detect any differences in the genome, epigenome or transcriptome that could explain the observed discordant differences.<sup>63</sup>

---

\* 5500 SOLiD™ System Information Sheet

† HighSeq™ 2000 Sequencing System, Specification Sheet

Very recently, an exome sequencing study was published, which aimed at the identification of rare and novel variants in IBD patients. The study focused on the detection of variants in known CD risk genes and was able to detect some novel variants, besides several already known variants.<sup>73</sup>

In this thesis, I aimed for the extraction of the full genetic variants in the genomes of a family trio of two healthy parents and one child with early onset of a severe case of CD. Other than the above mentioned study, this thesis is not restricted to known CD risk genes but rather aims to identify the causative variant(s) in the child's genome, assuming a strong genetic background and mendelian model of inheritance. Additionally, next generation sequencing was used to investigate the transcriptomes of the family that allows identification of expression changes and functional studies of genetic variants that occur in regulatory regions. Extending transcriptome work done in the trio, a method for assessing sense/anti-sense transcription events was developed that allows integrating of transcriptomic data with genetic and epigenetic information. The tool may ultimately be used to assess context-dependent regulatory events.

### **1.11. RNA sequencing**

Transcriptome sequencing (or RNA-Seq[ueencing]) can be a very powerful tool for investigation of functional consequences of genetic variations. This is especially true if the underlying genetic changes are known, as SNV-calling from RNA-Seq data is difficult for various reasons (i.e. allelic imbalance, RNA editing). Genetic variants affecting splice-sites for example can lead to decreased expression of an isoform (i.e. heterozygous SNVs) or even the complete loss of expression (homozygous SNVs). Also, nonsense SNVs or SNVs preventing correct protein folding may cause visible changes in gene expression (i.e. the absence of a protein might lead to an up-regulation of the corresponding gene as a compensatory mechanism or transcription aborts early). The effect of SNVs outside of exonic regions is difficult to assess with genomic data alone, but RNA-Seq data can add information towards functional relevance. Expression changes of genes involved in inflammation can be detected between inflamed and non-inflamed samples, which can be helpful to identify genes involved in inflammatory diseases. The following two examples further underline the use of RNA-Seq in disease interrogation. Using RNA-Seq, Hawkins and Kearney identified several novel candidate genes for epilepsy in mice.<sup>74</sup> Wu *et al.* compared the transcriptomes of 9 patients with schizophrenia and 9 controls and identified more than 2000 genes with schizophrenia-associated alternative promoter usage and more than 1000 differentially spliced genes.<sup>75</sup> Transcription by itself is a very complex network involving a broad range of regulatory mechanisms, such as allelic imbalance, RNA editing, RNA-RNA interactions, alternative splicing and antisense transcription. Most of these mechanisms are part of research by many groups, also in connection with NGS. However, the genome-wide assessment of antisense transcription by NGS is still hardly touched. Antisense transcripts might have a regulatory effect on the expression of the

sense transcript, which has been reported for several sense/antisense (S/AS) pairs before. One very well known S/AS pair (*XIST* and *TSIX*) is involved in X-chromosomal inactivation.<sup>76</sup> Another example is given by Morris *et al.*, who reported that expression of the p21 antisense transcript mediates methylation of the p21 sense promoter by recruitment of epigenetic regulatory complexes and showed that this effect was reversible by siRNA induced knockdown of the antisense transcript.<sup>77</sup> Published data about antisense transcription is sometimes contradictory, underlining the high complexity of the matter. In 2008, He *et al.* investigated S/AS patterns in different human cell lines using the Illumina GA and detected a high abundance of antisense tags within exons.<sup>78</sup> In contrast, in a paper published by Klevebring *et al.* in 2010, who used the SOLiD system, only few antisense tags were detected in the coding regions of genes.<sup>79</sup> However, both papers agree on an abundance of antisense transcription occurring in promoter and terminator regions.

To fill the gap left by no publicly available tool for the identification of sense/antisense transcript pairs on the genome-wide level, I developed a program, namely *Janus*, during this thesis. *Janus* observes transcription of the two stranded DNA, which can occur in sense and antisense direction.

## 2. Aims of the study

The setup includes a family with a severe case of Crohn's disease and unusual early onset (1.5 years). Whole genome and transcriptome sequencing potentially allows the identification of known and novel functional genetic variants in all genes and other genetic regions without an *a priori* biased approach. In this exemplary setup I have used next generation sequencing to address the following questions:

- Do genetic variations occur in genes known to cause monogenetic diseases phenocopying Crohn's disease?
- Do variants following the recessive model of inheritance or genes showing compound heterozygous variants occur in the child?
- Are there any *de novo* variants in the child, which effect known or novel potential risk genes?
- How much of the known risk loci does the child carry and how does this contribute to CD risk compared to the parents and other unrelated individuals?
- Are there any novel variants in known Crohn's disease risk genes?
- Genetic variants might occur in genes associated with other inflammatory diseases, which could hint towards a common link to inflammation. Does the child carry genetic variants in these genes?
- Are there any hitherto unassociated genes showing genetic variants which might contribute to CD risk?
- Regions of homozygosity (ROH) can hint towards risk loci. How do the ROHs differ between the three individuals and does the child have more known CD risk genes located in ROHs?
- Genetic variants might have an effect on gene expression. How different is the gene expression in the child compared to the parents? Which are the genes differentially expressed? Are the most differentially expressed genes affected by genetic variants?
- How prevalent are regulative mechanisms like antisense transcription? Are sense/antisense pairs differentially- or co-expressed? How do (epi)genetic variants influence sense/antisense transcription and is there a potential link to disease related transcript signatures?

## 3. Material and methods

### **3.1. Extraction of nucleic acids from blood samples**

DNA was extracted from frozen blood samples using the “Blood Giga kit” from Invitex™ (Berlin, Germany) following the manufacturer’s protocol. In the first step erythrocytes are lysed and removed. Afterwards, DNA from leukocytes is extracted in a second lysis step, without interfering hemoglobin. Protein components are digested using Proteinase K and DNA is precipitated with a 96% ethanol containing solution, washed with 70% ethanol, dried and finally resuspended in low salt buffer for downstream applications.

Preparation of total RNA was performed with “mirVana miRNA isolation kit” from Ambion™ (Paisley, UK) according to the manufacturer’s instructions. Similar to the DNA extraction method, at first a cell lysing step is performed. Afterwards, most DNA is removed by Acid-Phenol:Chloroform extraction of the samples. Ethanol is added and the solution is passed through a glass-fiber filter that binds the RNA. The filter is washed multiple times before total RNA is eluted with a low ionic-strength solution.

### **3.2. Library generation**

Independent of the type of application (miRNA, mRNA, exome enrichment, genomic DNA (gDNA)) the first step of library generation is the fragmentation of the input material. The following description is for fragment library preparation. The DNA (gDNA or cDNA in the case of RNA libraries) is sheered by sonication to fragments with a mean length of 160 bp (150-180 bp). Afterwards enzymes for end-repair are used to generate blunt-ended fragments (End-polishing enzyme 2) with a phosphorylated 5'-end (End-polishing enzyme 1 + ATP). Libraries are purified with PureLink™ columns before ligation of adapter sequences. A final size selection for fragments of 200 to 230 bp of the library is performed, which corresponds to the length of the fragmented DNA plus adapter sequences. To close potential gaps in the sequences, the eluate of the size selection undergoes nick-translation before the library is amplified by PCR. Before sequencing the library is purified with the “SOLiD™ Library Column Purification Kit” and measured by quantitative PCR.

The generation of long mate-pair (LMP) libraries requires a slightly different workflow. Library generation aims for an insert size distribution of fragments with lengths from ca. 1500-3500 bp. LMP CAP adapters are ligated to the ends of the sheered DNA fragments, which are then circularized by biotinylated internal adapters. The LMP CAP adapters do not have 5'-phosphate in one of their nucleotides, which results to a circular DNA molecule with nicks in both strands. The nick is then ‘moved’ in 5'- to 3'-direction by DNA polymerase I in a time and temperature



dependent manner. Subsequently T7 exonuclease and S1 nuclease cut the circle at positions opposite to the nicks, which releases a piece of DNA that contains sequences of both ends of the original fragment. Afterwards adapter sequences are ligated and library preparation proceeds as described above.

One library used the “Agilent Sure Select kit Human All Exon Kit V1” for targeted sequencing of exons. This introduces an additional step after the purified library has been generated. The prepared library is hybridized to the SureSelect capture library, consisting of biotinylated RNA library ‘baits’ for exon sequences. The hybridized libraries are incubated with streptavidin coated magnetic beads, which binds to the biotin. Beads and supernatant are magnetically separated. The beads are washed to separate the beads and the DNA and RNA libraries. After RNA digestion of the capture library, an exon specific DNA library (exome library) remains.

For each individual, four libraries were generated from genomic DNA and one library from RNA. The four genomic libraries consist of three paired-end libraries, two for the whole genome and one restricted to exonic regions, and one long mate-pair library. The RNA library was generated as paired-end library (see Table 1 for library and run chemistry used).

**Table 1 Overview of applied library and run chemistry**

<b>library type</b>	<b>library kit</b>	<b>run chemistry</b>
<b>long mate-pair</b>	SOLiD™ Long Mate-Paired Library Construction Kit V3+	SOLiD™ Top Mate Paired Sequencing Kit MM 50/50 v4
<b>paired-end (1<sup>st</sup>)</b>	SOLiD™ Paired-End Library Construction Kit V3	SOLiD™ Top Paired End Seq DNA Frag Kit MM 50/25 v3
<b>exome and paired-end (2<sup>nd</sup>)</b>	SOLiD™ Paired-End Library Construction Kit V4	SOLiD™ Top Paired End Seq DNA Frag Kit MM 50/35 v4
<b>transcriptome</b>	50+25 bp, SOLiD™ Total RNA-Seq Kit V4	Opti Fragment Library Sequencing Kit MM 50 v3+

### **3.3. Sequencing**

Before sequencing, an emulsion PCR (ePCR) is performed for all prepared libraries. The ePCR is required to generate clonal bead populations for sequencing. An aqueous phase containing the PCR ingredients (template, primers, polymerase and SOLiD™ P1 DNA Beads) and an oil phase containing emulsifiers are prepared and emulsion is created using the ULTRA-TURAXX® Tube Drive from IKA®. This results in micro-reactors (small droplets) containing a small fraction of the aqueous components. The quantity of aqueous components is chosen to favor droplets containing only one bead and template. The P1 adapter sequence of the template binds to the P1 sequence of the DNA bead, where the template is then amplified by PCR. Depending on the number of different templates and beads per droplet multiple results are possible, i.e. micro reactors containing no template or template fragments with two P2 adapters attached will not bind to the bead and will not be amplified, micro reactors containing multiple templates will result in non-clonal beads and

micro reactors containing multiple beads may lead to clonal beads, but lead to a small amplification bias of a fragment. The emulsion is abrogated by addition of 2-butanol and the beads are washed to remove any remaining oil and emulsifiers. Polystyrene beads coated with a P2 adapter are used for enrichment of amplifying beads. Non-amplifying beads and beads containing only fragments with P1 adapter will not bind to the polystyrene beads. In the subsequent centrifugation step polystyrene beads (with and without attached beads) result in a layer at the top and the non-amplifying beads form a layer at the bottom. The polystyrene beads are extracted and denatured to release the template beads. A 3'-terminal dUTP is added to template sequences before deposition of the beads on a glass slide for sequencing.

Sequencing was done on the SOLiD 4 platform, which uses 2-base encoding in a sequencing-by-ligation system. Generated sequences are written in SOLiD specific "color space". Sequential rounds of ligation with different labeled probes are carried out. The probes are eight nucleotide long molecules, two known bases which are interrogated, followed by three degenerated bases and three universal bases. Only four different dyes are used, thus each color corresponds to four dinucleotide combinations (including the 3 degenerated bases, there are  $4^5/4 = 256$  combinations per dye). For each cycle another sequencing primer is used which binds to multiple positions in the adapter sequence, shifted by one base-pair. As a consequence, each base is interrogated twice, by two different probes (i.e. in a subsequence of ATC two probes, ATNNN and TCNNN will bind and the T is interrogated twice). Each produced sequence starts with a T, followed by number corresponding to the colors (0-3). This allows for bijective translation into base space. This allows for a high accuracy, as valid nucleotide exchanges require two adjacent color space mismatches in relation to the reference sequence, which permits differentiation from sequencing errors (single color space errors).

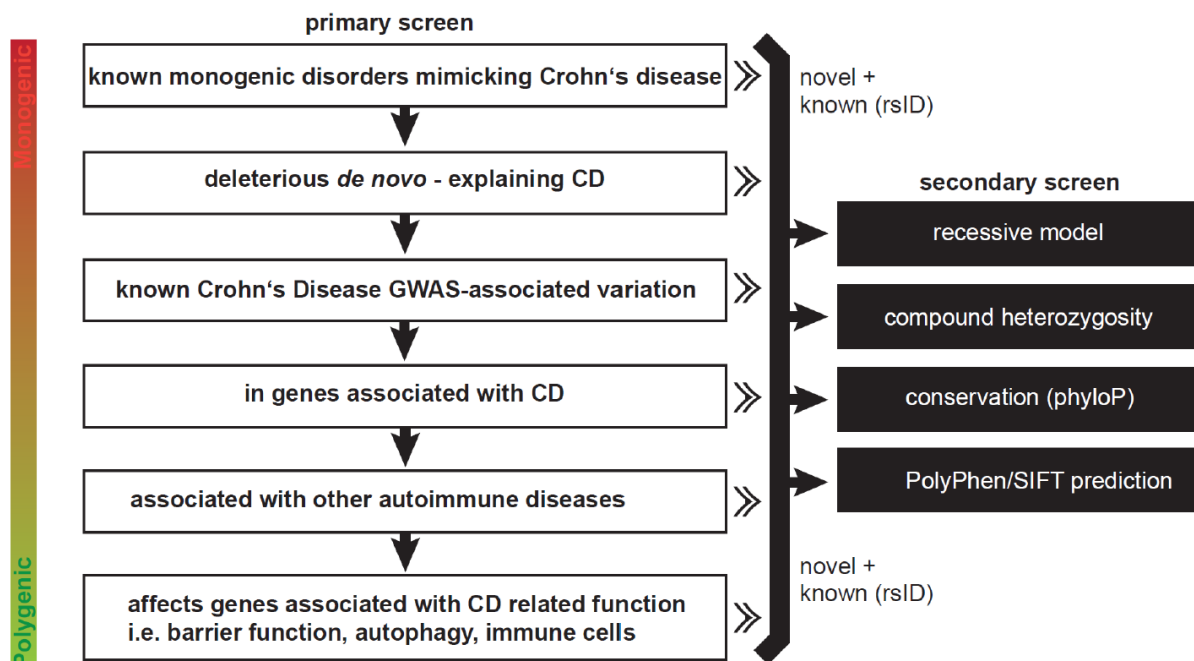
### **3.4. Sequence analysis**

Primary analysis (image analysis/base calling) and secondary analysis (mapping, calling of single nucleotide variants (SNVs) and detection of insertions and deletions) was performed with Bioscope™ (Applied Biosystems™). All sequence data was mapped against the non-masked human reference (hg18) using Bioscope's standard parameters. BAM files (the binary format of SAM (Sequence Alignment/Map format),<sup>80</sup> <http://samtools.sourceforge.net>) were generated with Bioscope's "mark duplicates" option and coverage has been calculated using Samtools (v.0.1.8) mpileup command. Ambiguous bases (Ns) in the reference and sequences marked as duplicates were ignored for the calculation. SNV calling was restricted to perfect pairs (same strand, correct orientation and without deletion and insertion) with a minimum quality value of 6 and a mapped bases to read length ratio of  $\geq 0.85$ .

For tertiary analysis I developed a multi-threaded Windows application in C/C++.

### 3.5. SNV annotation

The SNVs included in dbSNP build 130 were used to identify known and novel SNVs. SNVs present in dbSNP130 were assigned the respective rsIDs. Transcript information (refGene Oct. 2012, available at the UCSC Genome Browser<sup>81</sup> Website <http://genome.ucsc.edu/>) for transcripts within 2 kb range of the SNV was added to the SNVs. This includes the refSeq ID, gene symbol, the relative position of the SNV to the transcripts ([2 kb] neighborhood, 3'- or 5'- UTR, exonic, intronic), the type of nucleotide exchange (synonymous, nonsense, altered M[ethionine], new M[ethionine], read-through and splice-junction), the reference triplet and amino acid and the triplets including allele A and allele B of the SNV, as well as the respective amino acids. Information concerning the conservation of the SNV position was added using PhyloP<sup>82</sup> scores for mammals (available at the UCSC Genome Browser Website). PhyloP generates negative scores for predicted acceleration and positive scores for predicted conservation. For all SNVs contained in the result tables, allele frequency information for the CEU population based on the 60 sample pilot of the 1000 Genome project was added, when available. Additionally, SNV lists for each individual were complemented for SNVs only present in the other individuals by inclusion of consensus calls. Modified SNV data has been correlated to the Illumina SNV-Chip data that was converted from Illumina's TOP/BOT annotation to forward strand designation first. Additionally genes were screened for compound heterozygous variants (genes in the child that carry at least one but exclusive heterozygous SNV from each parent) and SNVs that are heterozygous in the parents but homozygous in the child. Figure 2 illustrates the workflow of SNV characterization.



**Figure 2** Filtering step for SV characterization.

On the top level genetic variants are compared to a list of monogenic disorders phenocopying CD. Next *de novo* variants are searched that could be functionally linked to CD pathogenesis. During the following two levels the detected variants are compared with known associated variants identified by GWAS and novel variants in the associated genes are searched. On the

next lower level, the search is extended to other autoimmune diseases. The bottom level regards all genes that could present potential novel CD risk targets. To narrow down the more likely disease causing variants on every level a secondary filter was applied that screens for variants following the recessive model of inheritance, accumulation of variants unique for either parent (compound heterozygosity), conservation and prediction of protein changes.

In a top-down approach, variants were screened for in genes involved in diseases phenocopying CD<sup>83</sup> (with the addition of two monogenic syndromes involving *IL10* and *XIAP*), genes and loci involved in CD pathogenesis based on data available in the GWAS catalog, meta-analysis<sup>12</sup> and OMIM database, genes known to be involved in other inflammatory diseases (based on the GWAS catalog) and lastly all other genes, restricted to SNVs that follow the recessive model, are compound heterozygous, located in a site of high conservation or were predicted to have a damaging effect on protein function by SIFT<sup>84</sup> and/or PolyPhen<sup>85</sup>. Detection of *de novo* SNVs was performed using two different approaches (see below) and exonic SNVs detected by both approaches were sequenced by Sanger.

### **3.6. Detection of *de novo* SNVs**

Two different approaches were used to detect *de novo* SNVs in the child. In the first, simple approach, Mendel errors (*de novo* SNVs and back mutations) were extracted from the SNV-Calling results of the child. The number of false positive Mendel errors was reduced by investigating the raw allele counts. Alleles with  $\geq 5\%$  of the total coverage were considered as present. SNVs with more than 2 different alleles were excluded, as these might pose difficult regions for SNV-Calling.

In the second approach, variants that were detected in the parents were excluded from further analysis. The remaining variants (~230,000) were annotated using a custom annotation pipeline, including overlap with dbSNP, overlap with our in-house variant database, genomic location according to refSeq and predicted amino acid consequences. I prioritized for unknown non-synonymous variants that were detected in five reads or more that made up at least 30% of the total amount of reads. This resulted in 37 potential *de novo* variants that were subsequently checked by retrieving the raw sequence calls of these positions in the parents' data. Subsequently, all exonic SNVs that were detected by both approaches were sequenced using Sanger sequencing.

### **3.7. Analysis of short structural variations (sSVs)**

Similar to SNVs short insertions and deletions (InDels) were categorized according to current gene annotations. InDels following the recessive model of inheritance were identified and described by using the same databases as for the SNVs.

### **3.8. Analysis of long structural variations (ISVs)**

Two approaches were used to detect ISVs. For the first approach structural rearrangements were detected using paired-end mapping.<sup>86</sup> Such methods rely on the expected mapping distribution and orientation of the sequenced library. The sensitivity is dependent on the spanning coverage (synonymously called physical coverage) and hence, large insert libraries can increase the power to detect structural rearrangements. All discordantly mapped paired-ends either show an abnormal distance or abnormal orientation and any such pair can indicate a rearrangement. To confidently call structural variants, discordant paired-ends were clustered into groups that support the same, putative structural rearrangement. For the second approach the Bioscope “Find Large Indels” tool was used applying standard parameters. All large deletions  $\geq 3.000$  bp that were detected by both approaches were manually checked, independent of the individual in whom they were detected.

### **3.9. Demographic origin of sequenced subjects**

The geographic position of all three individuals was assessed in a SNP-based Principal component analysis (PCA) with 1,296 individuals from 11 HapMap populations<sup>87</sup> using 41,163 SNPs shared between subjects. Principal component analysis was performed with default settings using the Eigensoft program<sup>88</sup> and genotype data available from the HapMap ftp site:[http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2009-01\\_phaseIII/plink\\_format/](http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2009-01_phaseIII/plink_format/).

### **3.10. Single nucleotide variant distribution along chromosomes**

To assess changes in the SNV distribution across the genomes of the family trio compared to the genomes of the ‘normal’ CEU population, R was used to plot the chromosomes reflecting regions of high SNV density in red and regions of low SNV density in blue. Each pixel represents 137360 bp (= 1800 pixel for chromosome 1). The color of the pixel is determined by the number of SNVs in the region, where 500 or more SNVs equals an RGB value of 1,0,0 (=red) and zero SNVs equals 0,0,1 (=blue).

### **3.11. STRING network analysis of missense single nucleotide variants**

Enrichment of Gene Ontology terms was performed for all missense SNVs in the child using Fisher’s exact test, followed by Benjamini-Hochberg correction for multiple testing. To investigate SNV burden of networks, STRING (<http://string-db.org/>) interaction analysis was performed for all genes contained within the GO term “immune system process”. Ten white nodes were allowed and connections based on text mining, gene fusion and neighborhood were excluded.

### **3.12. Selection of differentially expressed transcripts**

Between 127 and 151 million reads (5.0-5.7 Gb) of transcriptomic data were generated per individual. Of the transcriptomic data 60% of forward and 17% of reverse reads could be mapped to the hg18 reference. Expression was calculated with Cufflinks v0.9.3 (*cufflinks.cccb.umd.edu*)<sup>89</sup> using the UCSC *knownGene* table (March 2011) as reference that contains 66.8k transcripts. Between 35.5k and 38.0k transcripts were detected by at least 0.01 *Fragments Per Kilobase of exon per Million fragments mapped* (FPKM) per individual (Supplementary table 3). The number of higher abundant transcripts with FPKMs above 1 varied between 11k and 13k. About 29k to 31k transcripts could not be detected (< 0.01 FPKM) per individual. The RNA-Seq expression profiles reflect the child's inflammatory condition. Several IBD and general inflammation associated genes (e.g. *MAPK14*, *MMP9*, *TLR5* and members of the S100/calgranulins family) are differentially expressed, which means at least 5-fold higher or lower, in the child compared to its parents. For higher expressed transcripts, only those with a FPKM > 0 in the parents and FPKM > 1 in the child were regarded, while lower expressed transcripts required the parents' FPKMs to be > 1 and child FPKM > 0.

### **3.13. Regions of homozygosity**

A reasonable assumption for at least some proportion of variants, such as loss-of-function mutations, is a recessive risk model. Therefore it was investigated if known risk genes/loci for Crohn's disease are overrepresented in autozygous regions. Autozygosity presents as extended homozygous regions, or runs of homozygosity (ROH), in the human genome. ROHs are genetic regions that contain a large number of homozygous SNVs with a very limited number of interrupting heterozygous variants. ROHs were identified in all three individuals using an adapted version of the previously suggested GWAS-based definition in PLINK [REF Purcell et al. 2007 AJHG]. Due to the close relation to GWAS, PLINK uses only few SNVs compared to the much denser set of NGS-inferred variants that increases the chance that a heterozygous variant might interrupt the homozygous stretch. To strike a balance between very few ROHs, due to a long sliding window, and too many ROHs, due to a small sliding window, a sliding window of 250 kb (instead of 1000 kb) was used with 1) at least 50 SNVs, 2) at most one heterozygous SNV, 3) at most five missing SNVs and 4) a proportion of overlapping homozygous windows of 0.05. A ROH was defined by 1) a minimum size of 50 kb, 2) containing at least 100 SNVs, 3) a density of 1 SNV/10 kb and 4) a maximum distance between SNVs of 50 kb.

Finally, the overlap of ROHs with the 71 published CD risk loci<sup>1</sup> was calculated, combining the length of ROHs overlapping any risk loci and dividing by the combined length of all ROHs detected.

### **3.14. Calculation of child's CD risk relative to its parents**

Assuming small and independent contributions of known risk mutations to the probability of becoming affected with CD, it is possible to calculate the relative mutational load, or burden, of the child given the genotypes of the parents. Consider  $1, \dots, n$  autosomal variants and let  $X_{j,i}$  denote the odds ratio conferred by the  $j$ -th pair of alleles at variant  $i$  present in the child. Assuming non-preferential transmission of alleles, the expected logOR conferred by variant  $i$  in the child conditional on the parental alleles is the average of the four possible parental allele combinations

$$EX_i = \sum_{j=1}^4 \frac{1}{4} \log OR_{j,i}$$

with expected variance

$$\text{Var}X_i = \sum_{j=1}^4 \frac{1}{4} \log^2 OR_{j,i} - (EX_i)^2$$

In accordance with the central-limit theorem, the sum of equally weighted standardized variant risk contributions in the child

$$Z = \sum_{i=1}^n \frac{\log X_i - EX_i}{\sqrt{n \text{Var}X_i}}$$

follows a standard normal distribution:  $Z \sim N(0,1)$ . Evaluation of  $Z$  allows assessing the mutational load of the child relative to the parents.

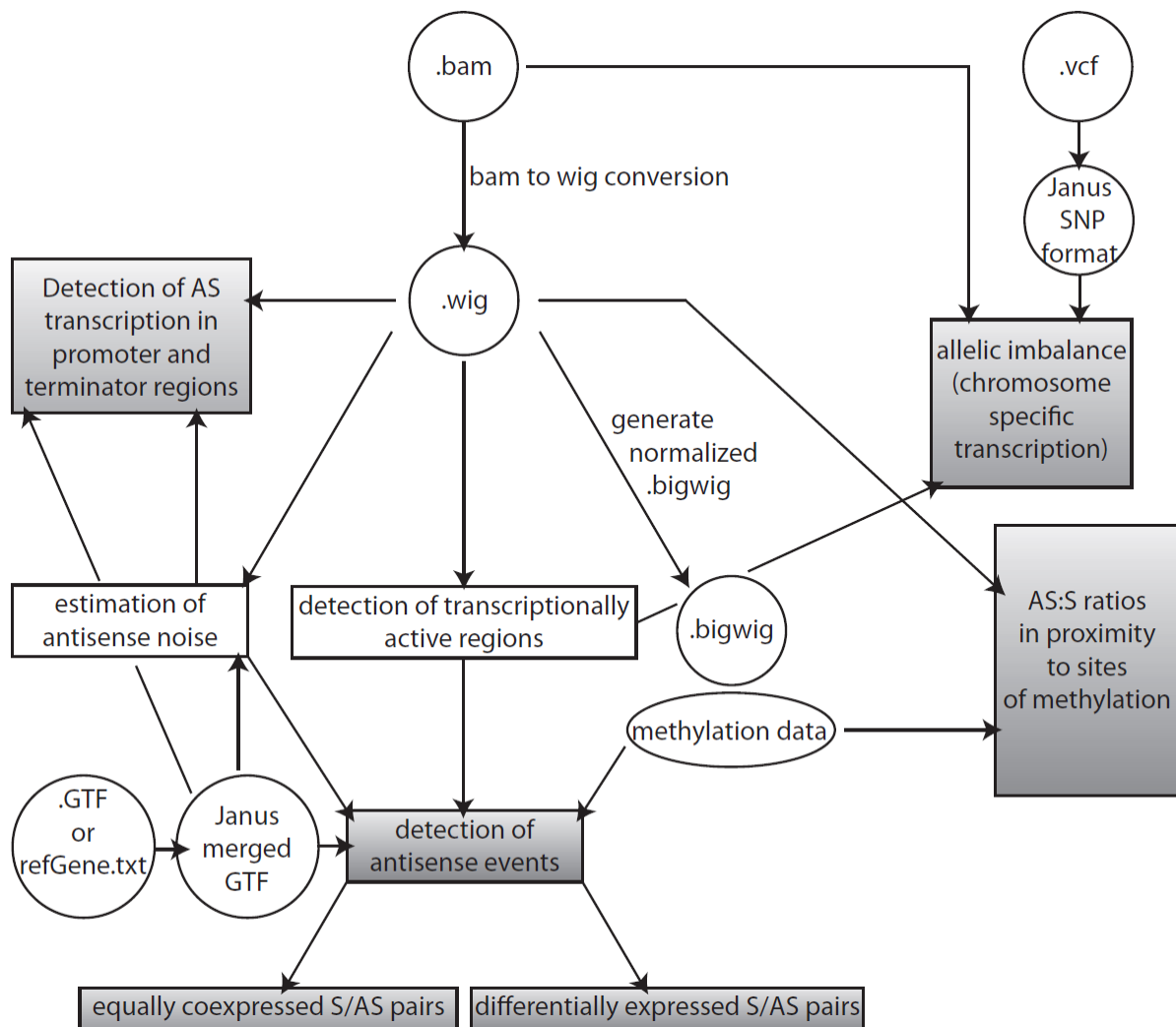
### **3.15. SNV verification by Sanger sequencing**

Polymerase-Chain-Reaction (PCR) was performed to amplify regions containing SNVs using the primers given in Supplementary table 4 and genomic DNA as input material. A second PCR was performed with the product of the first with separate reactions for the forward and reverse primer. The product of the second PCR has been sequenced using Sanger sequencing.

### **3.16. Identification of sense/antisense pairs in transcriptomic data**

As currently no tool is publicly available to deal with strand-specific transcriptomic data sets from next-generation sequencing platforms in an automated way, I developed a tool (*Janus*) that allows the identification of sense/antisense pairs in this type of transcriptomic data sets. *Janus* is written in C/C++ and can be run under both Windows and Linux but requires a 64 bit environment. The workflow of *Janus* is shown in Figure 3. *Janus* expects a BAM file as input, and uses the reads to generate one wig file per chromosome and strand. A wig file contains the information on how many reads cover each base of the genome that can be used to assess the genomic coverage, e.g. for graphical representation. For further analysis, a transcript annotation file in gene transfer format

(GTF) is required. *Janus* is designed to convert an existing GTF file or the refGene.txt, which can be obtained by the table browser of the UCSC Genome Browser Website (<http://genome.ucsc.edu/cgi-bin/hgTables>), into a merged GTF file that contains one entry per gene symbol, including the merged meta-exons of all transcript isoforms of the same gene. Although *Janus* can make use of commonly available gene annotations in GTF format, it is recommended running the built-in converter of *Janus* to improve the results. If the built-in converter is omitted, GTF-files might generate multiple, similar results for different isoforms of the same gene.

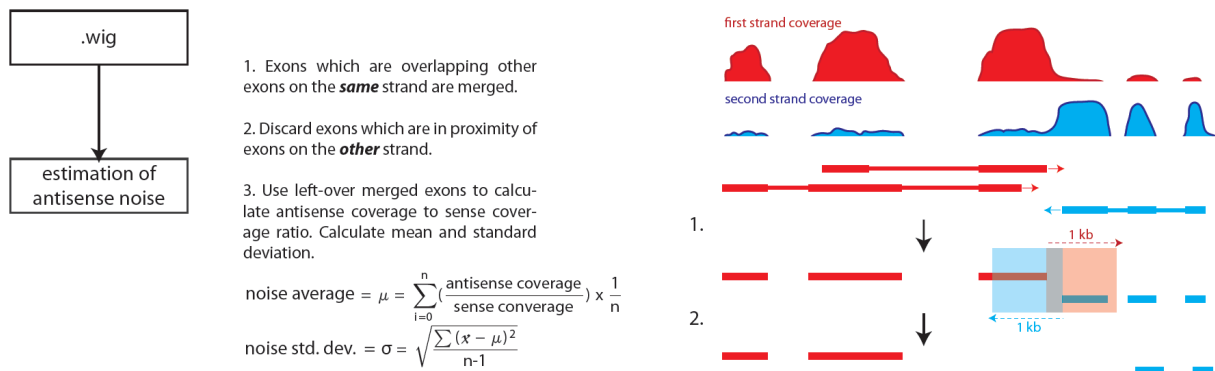


**Figure 3** *Janus* workflow.

Input files are marked by circles and final analysis output in boxes shaded in grey. *Janus* generates wig files from a BAM file, which is the primary input. The wig files can be converted into bigwigs and are used to estimate the level of antisense noise and to detect transcriptionally active regions. To determine the location of an antisense event relative to other transcripts, *Janus* depends on a gene annotation file, which can be either a refGene.txt file or a GTF file. Finally *Janus* detects antisense events above the estimated background noise and annotates them using the gene annotation file. Optionally *Janus* can incorporate methylation data. Multiple lists of sense-antisense (S/AS) pairs can be compared with *Janus* to detect differentially expressed and equally co-expressed S/AS pairs. Additionally *Janus* includes the functionality to determine AS:S ratios in proximity of methylation sites, quantify the amount of antisense transcription in promoter and terminator regions and the ability to detect strand specific allelic imbalance.



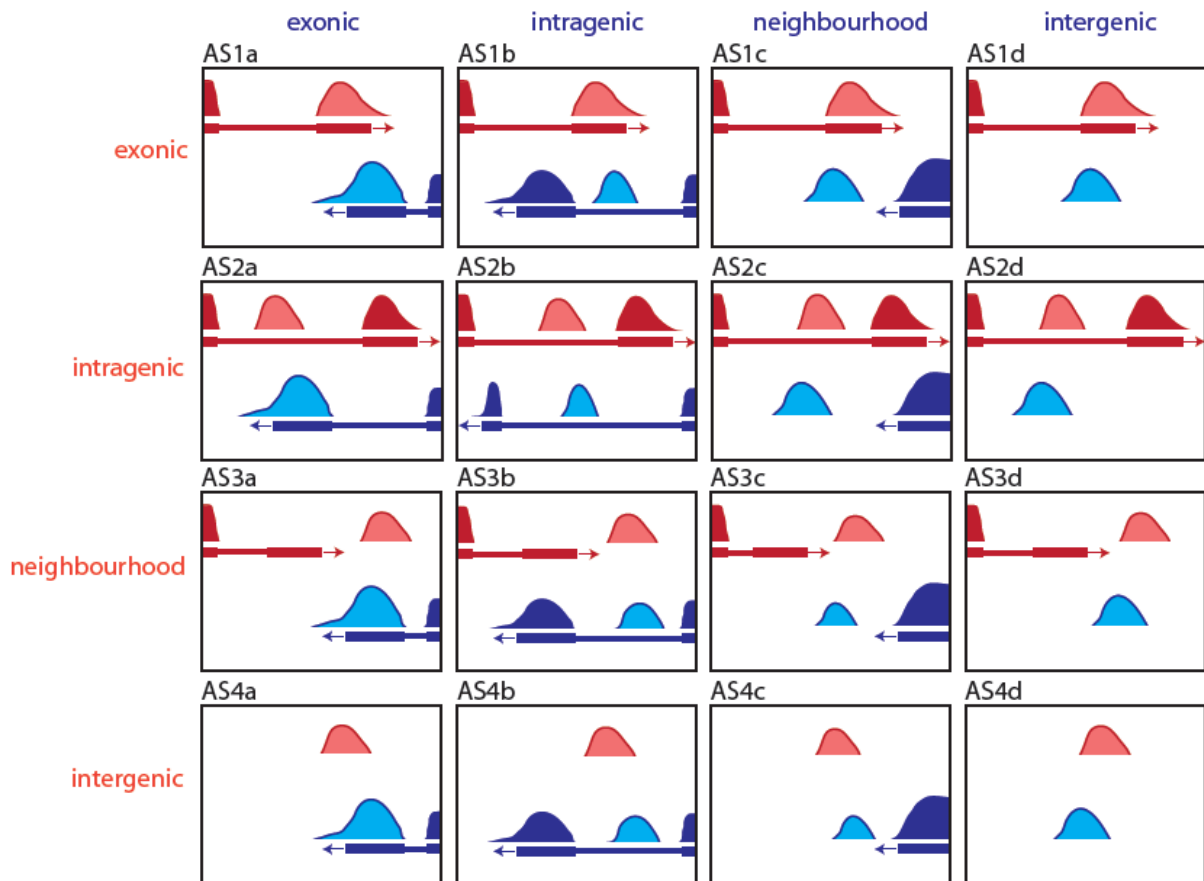
Using the reference annotation, the level of 'antisense-noise' is calculated (Figure 4). I defined this noise as transcription on the opposite strand of known exons, which is a result of imperfect library generation and mismapping of reads. Often, transcription extends beyond the 3'-end of annotated transcripts. As this would bias the noise estimation in case of two adjacent transcripts in tail-to-tail orientation, exons within 1 kb of other exons on the opposite strand were excluded. The mean and standard deviation of the noise are calculated and reported in a detailed tab-delimited file for every investigated meta-exon. Additionally the average mean and standard deviation are reported in a stats file that also contains the total coverage of the sample.



**Figure 4 Estimation of antisense noise level.**

At first all exons which are located in the same strand are merged. Next all exons which within 1 kb range of another exon, which is located on the opposite strand, are discarded. For all remaining merged exons the antisense to sense ratio is calculated using the coverage contained in the wig files. Finally the mean and standard deviation of the 'noise' level is calculated.

The wig files are used to identify transcriptional active regions (TARs) (Supplementary figure 1). Three criteria must be met for a TAR to be regarded as an antisense event. First, the TAR must have a minimum length (default = 51 bp) and minimum average coverage per base (default = 5). The coverage criterion uses unnormalized values, i.e. the raw mapped coverage for the identification of TARs. Second, on the strand opposite to the identified TAR, transcription must occur within a given distance (default = 10 kb). Third, each TAR requires an antisense- to sense-transcription ratio above the estimated noise level. Events that do not differ by more than 1.96 standard deviations ( $p\text{-value} \leq 0.05$ ) from the mean noise level are discarded. An antisense event is regarded as known if it is overlapping an exon on the same strand and novel otherwise. Each antisense event is classified depending on its location in relation to the sense event (Figure 5).

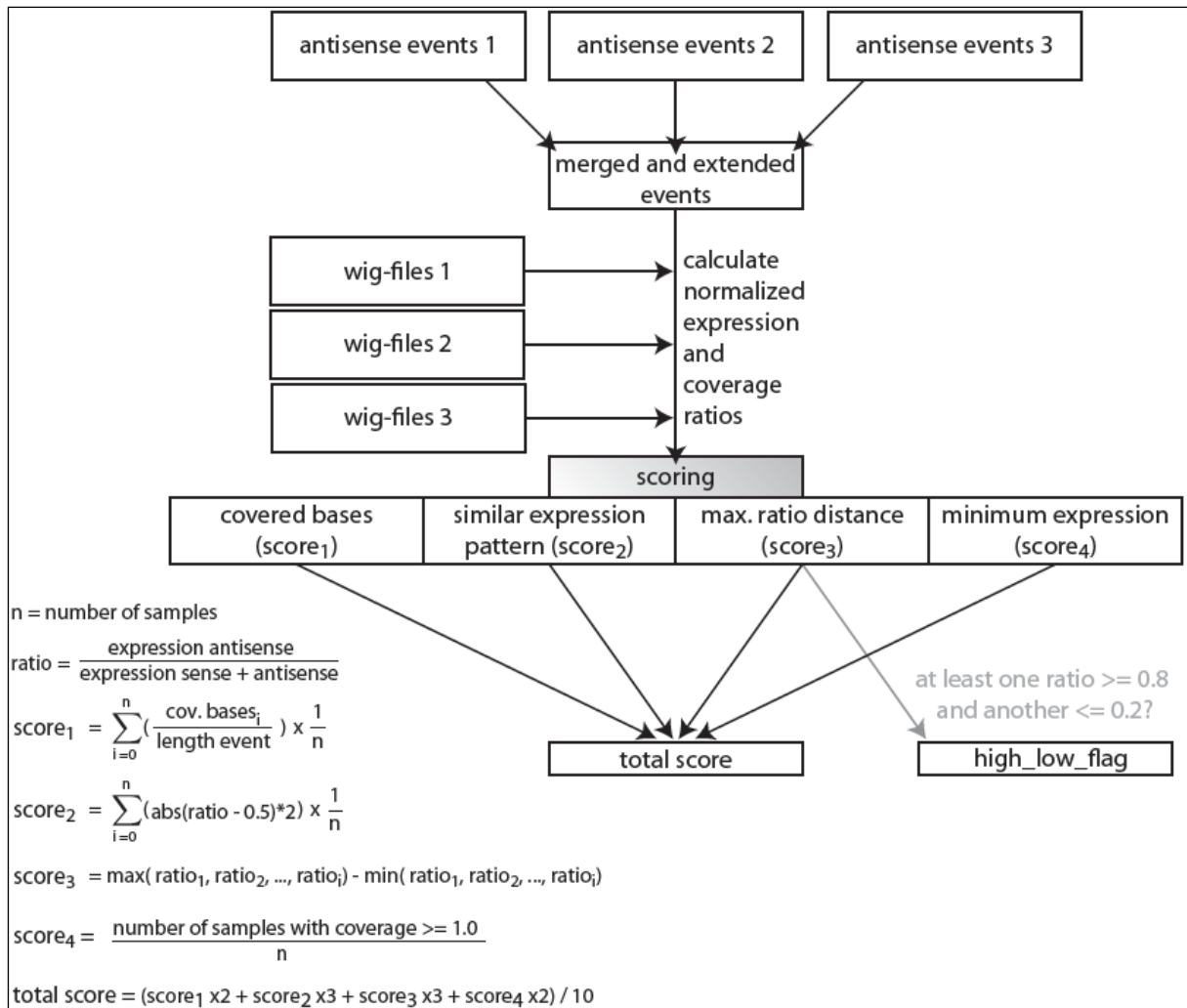


**Figure 5 Antisense classes.**

Coverage is shown for the first strand (red) and second strand (blue). Regions showing coverage on both strands (S/AS pairs) are shown in lighter colors.

*Janus* includes the functionality to compare antisense events of multiple samples to identify differences in gene expression that might affect only one strand or both strands equally. Supplementary figure 2A-C illustrates possible types of AS events, which show equally strong expression as the sense transcript and Supplementary figure 2D an S/AS pair with mutually exclusive expression of either the sense or antisense transcript. The workflow to identify potential self-regulatory S/AS pairs is shown in Figure 6. First, all overlapping antisense events are merged for all samples and expression values are calculated for each strand and sample. The expression values are normalized by event length and total genomic coverage before scoring. Two quality scores are incorporated in the scoring algorithm: the mean value of covered length to full event length ( $score_1$ ) and the number of samples which have a minimum coverage of 1 for the whole event ( $score_4$ ). For identification of differential expressed S/AS pairs two scores based on the AS:S ratio are used: the mean of the difference of the AS:S ratio from an equal distribution (1:1) and the distance between the minimum and maximum AS:S ratio. For the final score, the four sub-scores are weighted slightly in favor of the ratio-depending scores (2 and 3) as they are more important for detection of expression differences than the quality scores (1 and 4). Additionally a flag is

generated, which marks the more extreme cases, in which one sample shows an AS:S ratio below 0.2 and another above 0.8 representing empirical borders determined from our training data set.

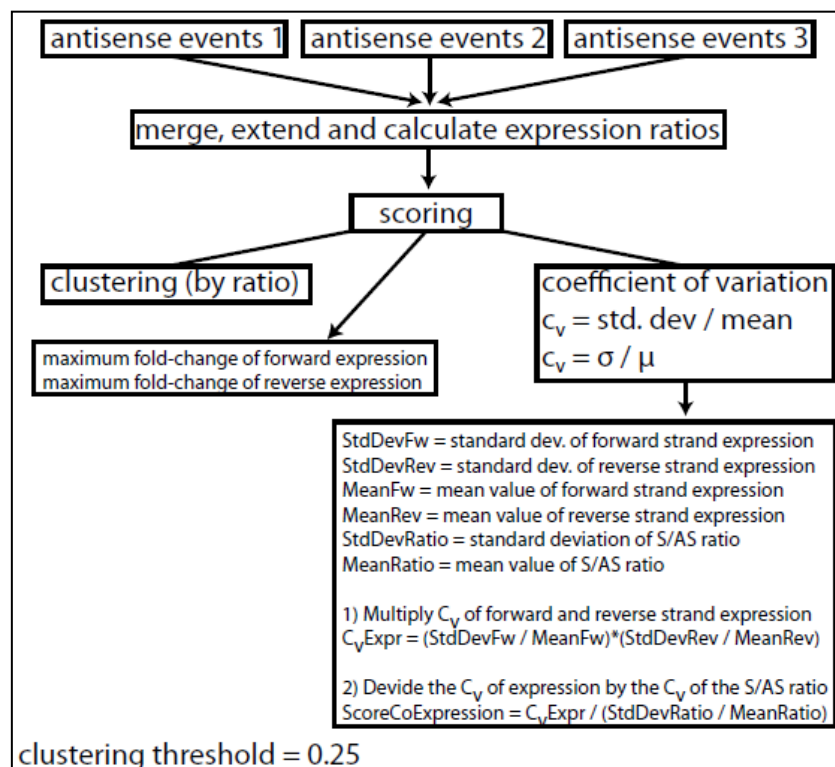


**Figure 6 Scoring algorithm for S/AS pairs.**

First the lists of antisense events detected in different samples are merged and overlapping events fused into single events. Sense-antisense ratios are calculated and expression values are normalized for total genomic coverage and event length. Antisense events are given scores for mutually exclusive expression. All scores, including the total score, are between 0 and 1. Score<sub>1</sub> represents the fraction of the event which was covered. Score<sub>2</sub> (0 to 1) marks the distance of the event's S/AS ratio to an equal expression on both strands. Score<sub>3</sub> calculates the maximum S/AS-ratio-distance between the sample with the lowest S/AS ratio to the sample with the highest S/AS ratio. Additionally a flag (either TRUE or FALSE) is generated, which tells the user if a sample with a very low and another with a very high S/AS ratio was detected. Score<sub>4</sub> represents the fraction of samples which achieved a normalized coverage of at least 1. The total score weights all four scores in slight favor of the expression ratios that are more important for the detection of differentially expressed S/AS pairs than the quality scores 1 and 4.

The chosen scoring algorithm was not designed to preferentially identify S/AS events exhibiting the exact same number of reads on both strands, as this will not commonly be the case in regulatory events. However, it is possible to filter for them using additional parameters which are also calculated during the scoring step for differential expressed S/AS pairs (Figure 7). Samples are grouped by their AS:S ratio for a given S/AS pair using a clustering threshold of 0.25 that was empirically identified in the testing data set. For an equally co-expressed S/AS pair only one group

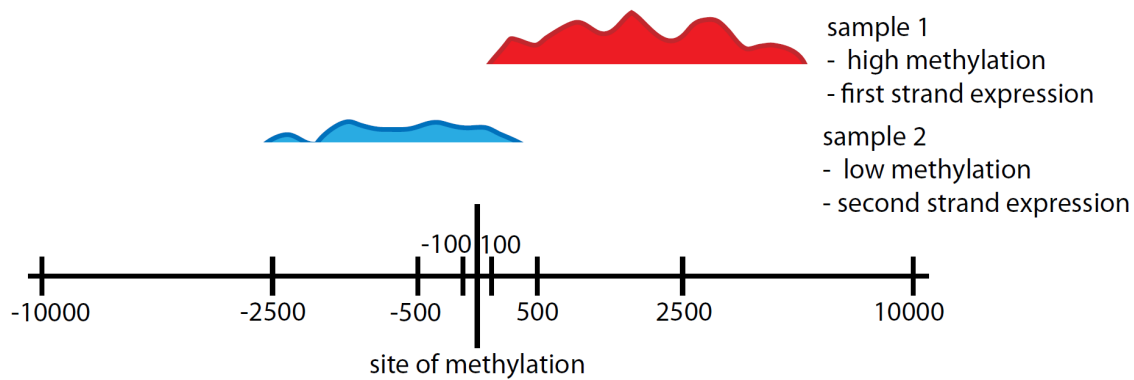
should be present. Also, the similarity score, used in the calculation of differential S/AS pairs, is supposed to be low (0 indicates a AS:S expression ratio of 0.5) and the maximum difference of the highest and lowest AS:S ratio should be very low as well. For identification of equally co-expressed S/AS pairs the fold-change of overall expression of the S/AS transcripts between samples needs to be high. *Janus* generates a score for the co-expression by calculating the coefficients of variation for the forward and reverse strand expression and AS:S ratio. The score grows with higher expression dissimilarities between samples while maintaining a similar AS:S expression ratio.



**Figure 7** Calculations helping to identify S/AS pairs which are showing an equal expression (Co-expressed S/AS pairs).

The Co-Expression score is a value for the dispersion of the coverage values in relation to the dispersion of the ratio. The value gets better/higher when the dispersion of coverage is high, while the dispersion of the ratio is low.

Methylation data can be incorporated in the results (list of antisense events) by specifying a tab-delimited file with methylation data. With this data, *Janus* can generate an additional file for the AS:S ratio in 100 bp, 500 bp, 2.5 kb and 10 kb range of the given site of methylation (Figure 8).



**Figure 8 Estimation of sense and antisense expression near methylated CpG islands**

The sense and antisense expression ratios are calculated within different ranges (100, 500, 2500 and 10000 bp) to given locations of methylation sites. Changes of the sense to antisense expression ratio in either range bracket in proximity of methylation sites with a different degree of methylation could hint to a control mechanism of S/AS transcription by methylation.

S/AS pairs that might derive exclusively from either homolog, i.e. one transcript is transcribed from the maternal chromosome while the antisense transcript is transcribed from the paternal chromosome, can be investigated based on a list of single nucleotide variants (SNVs) or without prior SNV information. VCF files can be processed, which results in tab-delimited files that are then used by Janus (see documentation for file formats). For all SNV positions located within detected S/AS events, Janus counts the number of reads for each allele, which might be evenly distributed on both strands or prefer one strand for the sense and antisense transcript each (Supplementary figure 3), and calculates a p-value based on Fisher's exact test for a 2x2 contingency table (first, second strand and allele A, B). Only single nucleotide variants are considered for this analysis. Read counts for all SNV loci that show at least one allele, without requiring reads on both strands, will be reported and a warning will be displayed at the end of each entry if the 'inferred' (=dominant allele per strand) alleles differ from the alleles specified in the SNV file. This type of analysis can be used to identify genes where one haplotype is preferentially expressed, possibly in a strand-specific manner, which can be a sign of regulatory variation in *cis*. However, *Janus* was not specifically designed to detect monoallelic expression. Without a specified SNV list, *Janus* will investigate all positions covered by TARs and search for allele differences between the two strands. By this method, only loci which are covered by forward and reverse reads showing two distinct alleles will be reported and Fisher's exact test will be applied using the dominant allele from each strand.

## 4. Results

### 4.1. General mapping and variant calling statistics

Five different genomic libraries and one transcriptome library from blood samples were sequenced per individual, with the addition of one transcriptome library of a colon biopsy of the child. A Robust detection of genetic variants is strongly dependent on a decent genomic coverage. For all three genomes an average coverage of 35 to 40x (for bases not designated N) was generated, which is comparable to the coverage attained by other whole genome sequencing studies (Table 2, Supplementary table 5-8).<sup>63,65,70</sup> The genomes of all three individuals are almost completely (>99.85%) covered by at least one read and more than 84% by at least 20 reads. More than 48 million reads (130 mio. total reads) were mapped per transcriptome library which is in the upper range of what many other studies using human samples generated and mapped.<sup>90-93</sup>

**Table 2 Per library numbers of generated sequences, mapped sequences and coverage depth across the genome.**

All individuals show a comparable number of generated sequences (>100 Gb) and >99.5% of the genomes are covered by at least one read. More than 80% are covered by 20x and more. Resulting in an average coverage per coverable base of 34 to 40x.

	generated sequences						
	lib_MP	lib_PE1	lib_PE2	lib_PE3	lib_EXM	lib_TRANS	lib_TRANS2
<b>mother</b>	920,348,453	939,401,482	930,584,259	1,046,748,964	205,889,144	131,457,938	-
<b>father</b>	969,530,350	900,078,995	1,074,093,601	882,420,484	210,193,250	151,026,450	-
<b>son</b>	891,794,740	908,963,316	901,806,883	1,258,255,210	200,649,856	127,103,413	149,329,469
	mapped sequences [N]						
	lib_MP	lib_PE1	lib_PE2	lib_PE3	lib_EXM	lib_TRANS	lib_TRANS2
<b>mother</b>	727,227,180	533,163,835	449,813,371	524,773,007	111,801,312	48,937,125	-
<b>father</b>	709,232,320	548,006,777	668,768,996	535,440,522	119,306,238	65,126,226	-
<b>son</b>	702,691,390	500,487,153	514,134,371	642,785,699	109,316,862	59,741,433	51,449,910
	mapped sequences [Gb]						
	lib_MP	lib_PE1	lib_PE2	lib_PE3	lib_EXM	lib_TRANS	lib_TRANS2
<b>mother</b>	36.3614	21.7825	17.7922	24.2766	5.0619	2.1965	-
<b>father</b>	35.4616	22.1349	27.1026	23.7637	5.3737	2.8693	-
<b>son</b>	35.1346	20.6031	21.0426	29.8446	4.9322	2.7352	2.2794
	total coverage across the genome (excluding Ns in the reference)						
	0x	1x	3x	10x	20x	average coverage per coverable base	
<b>mother</b>	0.14%	99.86%	99.47%	96.49%	84.44%	35.59	
<b>father</b>	0.12%	99.88%	99.55%	97.34%	89.58%	40.28	
<b>son</b>	0.13%	99.87%	99.48%	96.51%	84.81%	36.57	

MP = mate-pair library; PE = paired-end library; EXM = exome library; TRANS = transcriptome library from blood, TRANS2 = transcriptome library from biopsy sample.

The subsequent step was the identification of genetic variation in comparison to the putatively healthy reference genome (here hg18). More than 3.3 million single-nucleotide variants (SNVs) were identified per individual by the Bioscope SNV caller and the robustness of the calls was

assessed by comparison with a SNV chip, which achieved a concordance rate of >99% for SNVs called by Bioscope and >97% when including positions covered by the chip which did not result in a SNV call by Bioscope (Table 3, Supplementary table 9-10). More than 200k small structural variants were identified per individual (Table 3, Supplementary table 11), but only a small fraction is located within exonic regions, which is expected as these variants are likely to produce nonfunctional transcripts (frame-shift) or proteins. Both methods used for identification of large structural variants differed by a large margin and resulted in about 5,000 ISVs called by Bioscope and 15-113k ISVs called by the second method that was performed according to Korbel *et al.* 2009 (Table 3, Supplementary table 12).<sup>86</sup> Manual inspection of multiple ISVs showed that both approaches report a larger number of false positive ISVs, especially the zygosity did not seem to be correct in many cases. Thus for the following analysis the number of ISVs was reduced to large deletions with a minimum length of 3 kb which were called (at least overlapping) by both approaches (N=1185). These 1185 large deletions were manually inspected to further assess the most probable zygosity (Supplementary table 13).

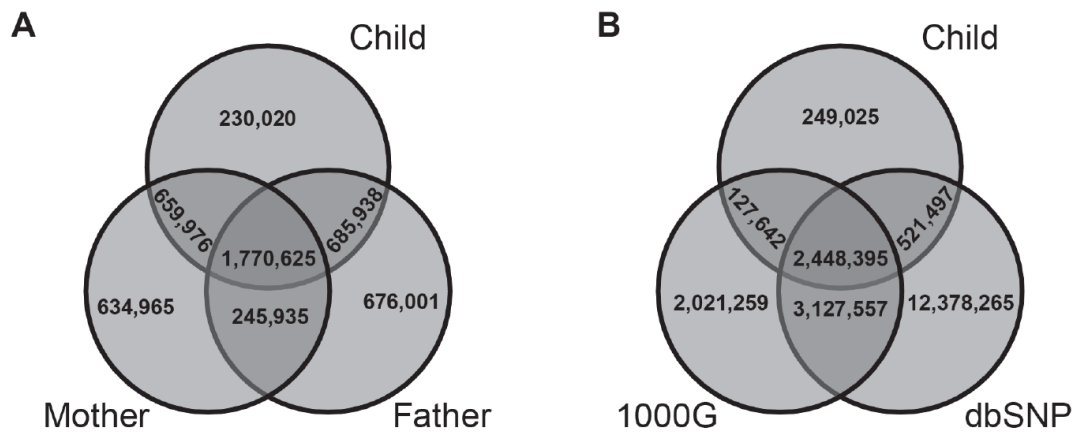
**Table 3 Variant detection statistic.**

Gene associated SNVs include intronic SNVs and SNVs in gene neighborhood. Large structural variants were called using two different approaches. The first approach used the Bioscope Large InDel caller and the second approach is based on Korbel *et al.* 2009.<sup>86</sup> The overlap of large deletions  $\geq 3,000$  bp between both approaches (1185 ISVs) was manually investigated.

single nucleotide variants						
SNP-type/location	mother		father		child	
	known	novel	known	novel	known	novel
rsID SNVs:	2,995,470	-	3,055,342	-	3,027,916	-
novel SNVs:	-	312,818	-	319,877	-	315,156
gene associated SNVs:	1,173,779	112,517	1,199,115	113,088	1,194,183	112,535
short structural variants						
SNP-type/location	known	novel	known	novel	known	novel
total sSVs detected:	145,477	48,722	170,158	49,395	149,450	43,328
no gene association:	88,918	31,585	104,652	32,332	91,224	28,300
exonic:	196	83	213	81	212	68
large structural variants						
Total ISVs detected:	Bioscope	Korbel	Bioscope	Korbel	Bioscope	Korbel
		5099	15892	5059	22763	4693

#### **4.2. Verification of *de novo* SNVs using Sanger sequencing**

Potential rare variants with major contributions to CD are expected to be unique to the child and thus cannot be detected in the parents or other healthy, unrelated individuals. About 230k SNVs were identified in the child, that were not detected in the parents and about 300k SNVs per individual that are not included in the dbSNP build 130. More than 249k of SNVs detected in the child are not included in the SNV data of 60 individuals of the 1000 Genome Project (Figure 9 A+B).

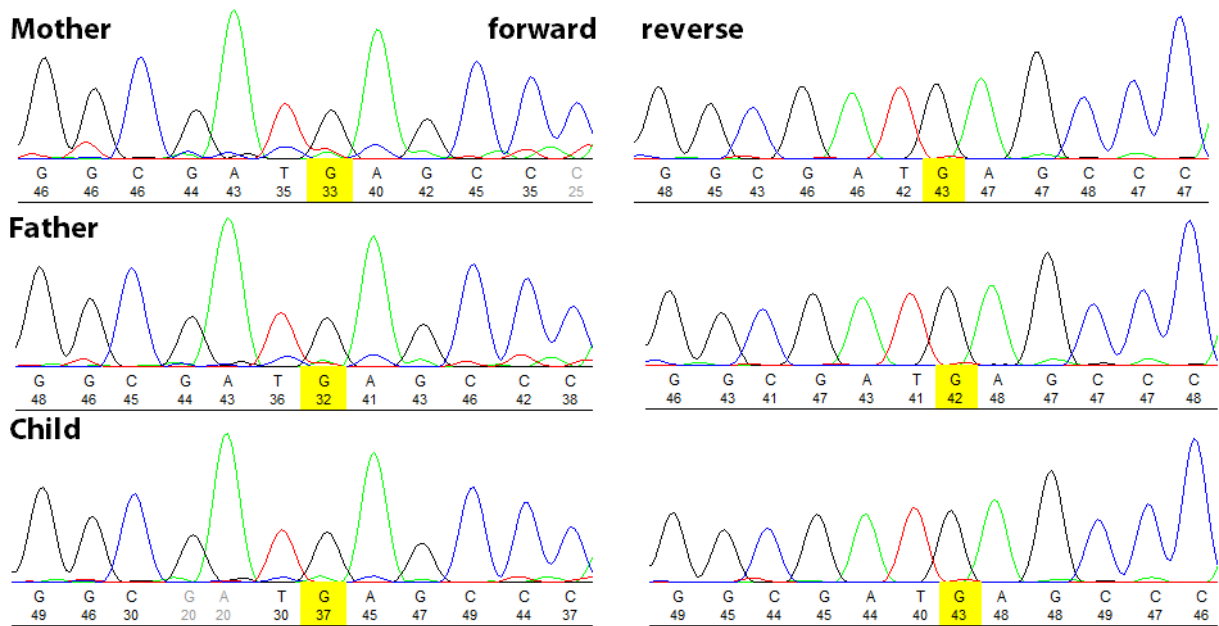


**Figure 9** Diagrams showing shared genetic variants between the family, dbSNP and the 1000 genome.

**A** Venn diagram showing the overlap of the child's SNVs with parents' SNVs. The child shows about 230k SNVs that were not detected in the parents. **B** Venn diagram showing the overlap of child's SNVs with dbSNP130 and 1000G SNVs. While most SNVs were also identified in the 1000G individuals or dbSNP about 250k SNVs remain that are yet novel.

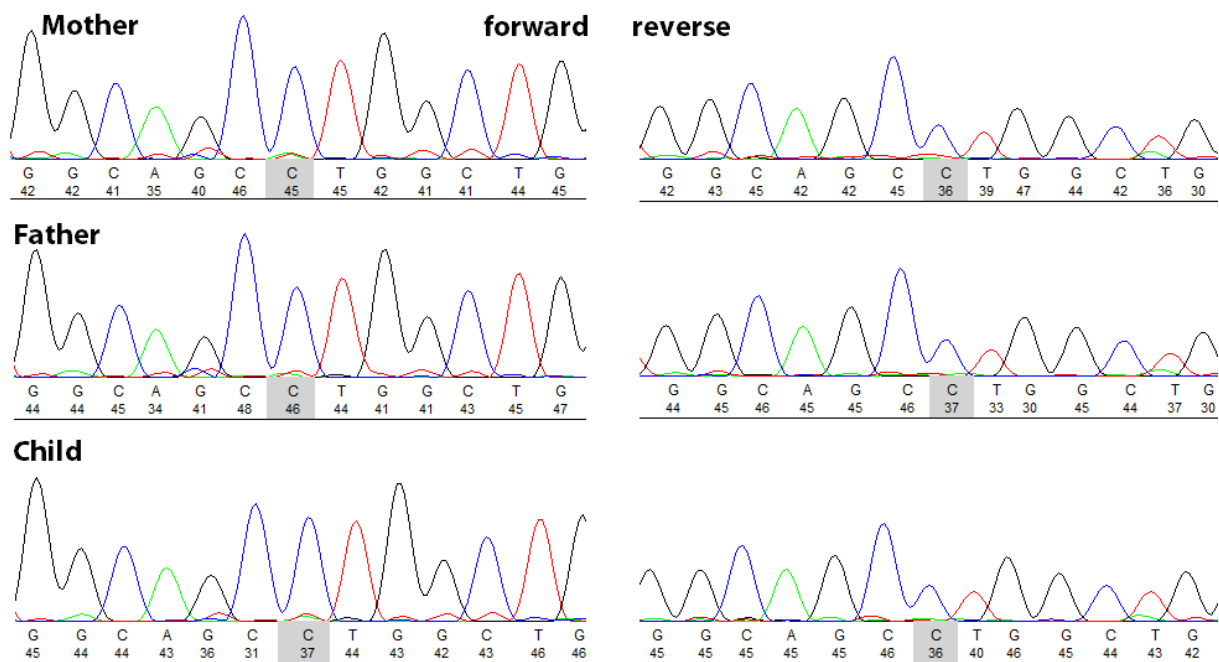
The SNVs unique to the child (potential *de novo* SNVs) were further investigated. More than 230k potential *de novo* SNVs (230k) were identified which exceeds the number of expected *de novo* SNVs by far. In humans, the mutation rate has been reported to be  $3.0 \times 10^{-8}$  mutations per nucleotide and generation.<sup>94</sup> The human genome has about  $3.0 \times 10^9$  million base pairs, which, in a diploid organism, is equivalent to 20 nucleotide exchanges per generation. Therefore, two algorithms (see methods) were applied to narrow the number of potential *de novo* SNVs. Both approaches did not result in convincing data for *de novo* SNVs in coding regions. Despite this, the overlap of weaker exonic *de novo* SNV candidates from both approaches was investigated (23 SNVs) and resequenced using Sanger sequencing. Additionally three SNVs with convincing data for a *de novo* variant were included that are located in noncoding regions (Table 4 and Supplementary table 14), and potential *de novo* SNVs in *IKZF1* (Figure 10), *CD19* (Figure 11), *MST1* (Figure 12) and *MUC2* that were called by only one algorithm. None of the coding SNVs could be confirmed, but one *de novo* SNV in the UTR of *NF1* (neurofibromin, Figure 13) and one intragenic SNV between *SYT4* (synaptotagmin) and *SETBP1* (SET-binding protein 1). Two of the investigated loci proved to be false negative calls in one of the parents (*MST1* and *FAM190A*), which could be due to low coverage in the respective region. No *de novo* variants in coding regions could be identified in the child. However, it is possible that a *de novo* mutation remained undetected by the SNV callers.





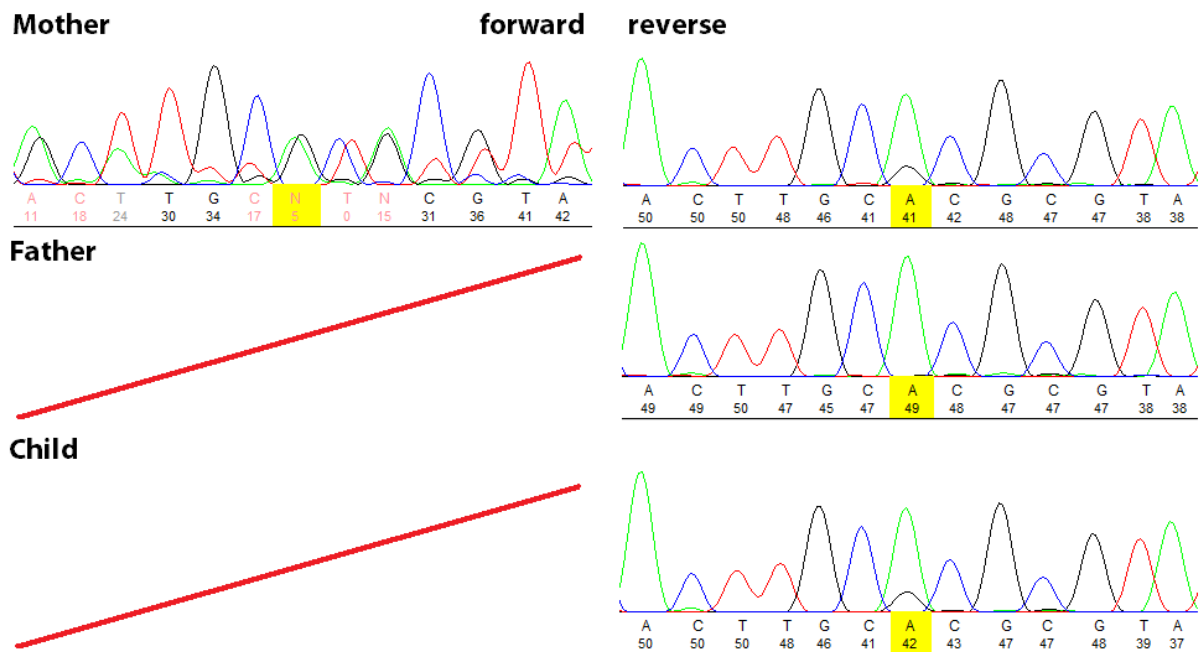
**Figure 10** Sanger sequencing result of *IKZF1* recessive SNV.

All three individuals show the reference G allele in all sequencing reactions.



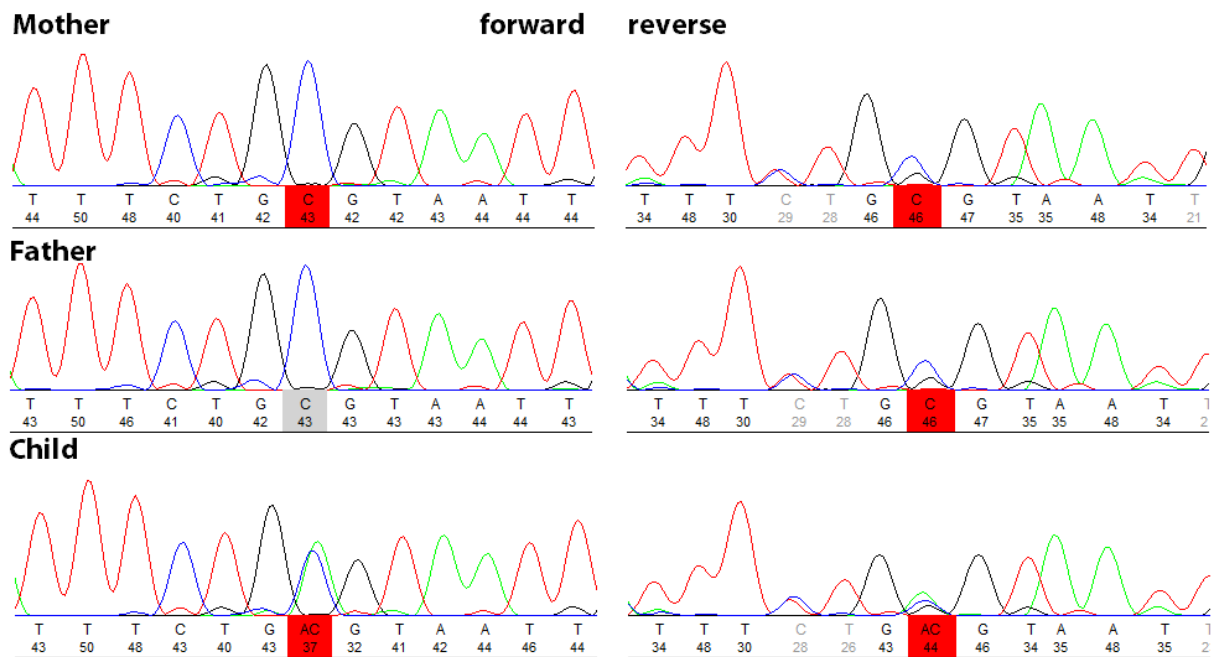
**Figure 11** Sanger sequencing result of *CD19* novel SNV.

All three individuals show the reference G allele in all sequencing reactions.



**Figure 12 Sanger sequencing result for *MST1* recessive SNV.**

The SNV was found heterozygous (supported by fw and rev reactions) in the mother and child (only rev. primer reaction) and homozygous for the reference allele in the father (only rev. primer reaction). Thus this SNV can be excluded as recessive SNV.



**Figure 13 Sanger sequencing result for *NF1* de novo SNV**

The SNV was not detected in the parents, but heterozygous in the child. The SNV is located in the UTR and thus its function, if any, remains unclear.

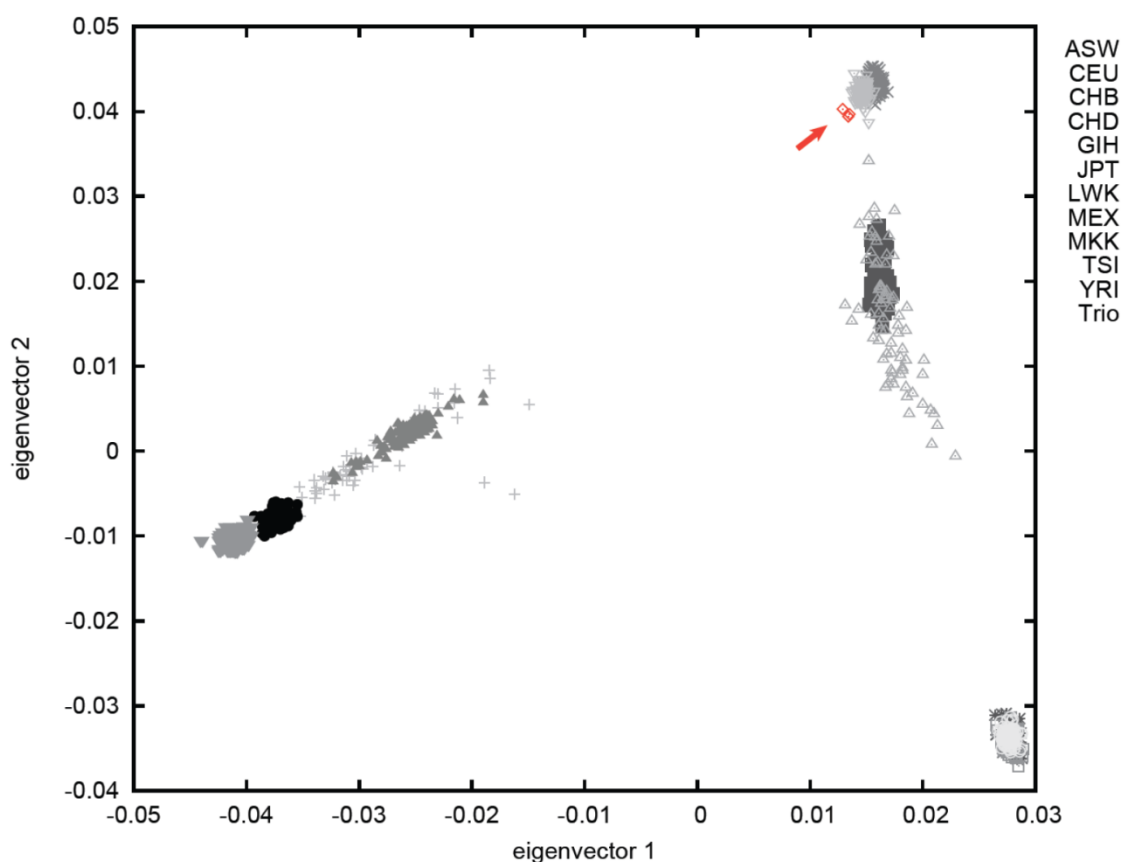
**Table 4 Verification of *de novo* SNVs using Sanger sequencing.**

Two non-coding SNVs were verified (*NF1* UTR, *SYT4-SETBP1* intergenic). Two SNVs (*FAM190A* and *MST1*) were detected to be heterozygous in one of the parents and thus represent false negative SNV calls. All remaining potential *de novo* SNVs show the reference allele and present false positive SNV calls.

gene	Sanger result	gene	Sanger result	gene	Sanger result
<i>FAM190A</i> (intron)	✗ (parent het)	<i>MST1</i>	✗ (parent het)	<i>IL18BP</i>	✗
<i>SYT4-SETBP1</i> (intergenic)	✓	<i>PAK6</i>	✗	<i>RASA3</i>	✗
<i>NF1</i> (UTR/exon)	✓	<i>CD19</i>	✗	<i>IKZF1</i>	✗
<i>AURKAIP1</i>	✗	<i>KIAA1522</i>	✗	<i>IRF5</i>	✗
<i>FUT4</i>	✗	<i>MUC2</i>	✗	<i>KLC3</i>	✗
<i>PNKP</i>	✗	<i>WNT9B</i>	✗	<i>MMP9</i>	✗
<i>KCNAB1</i>	✗	<i>LDOC1L</i>	✗	<i>OGFOD2</i>	✗
<i>HCN4</i>	✗	<i>ITGAV</i>	✗	<i>NOS3</i>	✗
<i>C19orf6</i>	✗	<i>SPZ1</i>	✗	<i>NUMB</i>	✗

### 4.3 Geographic origin determination based on genetic variants

As has been exemplified by *NOD2* as an ethnic specific CD risk gene, it is important to investigate the ethnic background of individuals used in genome analysis to assess the involvement of known risk factors. Principal component analysis of called SNVs positioned the trio closest to the European population (Figure 14). This data is further supported by mitochondrial variation (analysis was performed according to Röhl *et al.* 2001),<sup>95</sup> positioning the family in Eastern Europe (Supplementary figure 4).



**Figure 14 SNV-based principal component analysis (PCA) of the family trio with 11 HapMap populations.**

Individuals of African origin (left side), Asian origin (right side) and European origin (upper- right side) form distinct clusters. The family trio (red) is located close to other European populations.

#### 4.4. Genetic variants in genes associated with monogenic phenocopies of Crohn's disease

Multiple diseases are known to cause a Crohn's disease like phenotype but are based on a very limited number of genes as would be expected for Mendelian diseases. To investigate whether the child suffers from one of these phenocopies the corresponding genes were screened for genetic variants (Table 5, Supplementary tables 15-18).

**Table 5 Genetic variation in genes of monogenic disorders phenocopying Crohn's disease.**

No genetic variants with known association to disease phenotypes have been identified in the child. Several of the disease genes did not show any variant (WAS, X-linked Agammaglobulinemia, IPEX, Hyper-IgM syndrome, C1 esterase inhibitor deficiency). Some not disease associated variants were identified in genes for the remaining monogenic phenocopies of CD. All but one of them are included in dbSNP130. The novel SNV was not verified by Sanger sequencing.

disease	impaired function / associated genes	identified coding variants (SNVs, sSVs, ISVs) and additional information
<b>GLOBAL LYMPHOCYTE DYSFUNCTION</b>		
<b>Wiskott–Aldrich Syndrome</b>	The etiology is mutation of WASP, which encodes a signal transduction protein between transmembrane receptors and the actin cytoskeleton in hematopoietic-derived cells, resulting in multiple leukocyte deficits.	- no variants in <i>WASP</i> gene
<b>Common Variable Immunodeficiency (CVID)</b>	Decrease of at least two of IgA, IgM and IgG along with impaired antibody production following immunization, and propensity to infection (Mutations in <i>ICOS</i> , <i>CD19</i> , <i>TNFRSF13B/TACI</i> , <i>TNFRSF13C/BAFFR</i> ).	- nonassociated SNVs detected in <i>CD19</i> and <i>TNFRSF13B</i> : <ul style="list-style-type: none"> <li>o two <i>CD19</i> SNVs: rs2904880 and one novel SNV</li> <li>o three <i>TNFRSF13B</i> SNVs detected: rs11078355, rs34562254, rs8072293</li> </ul> - Sanger sequencing of novel <i>CD19</i> SNV negative
<b>X-linked Agammaglobulinemia</b>	Absence of B cells and concomitant inability to produce antibody (Mutations in <i>BTK</i> ).	- no variants in <i>BTK</i> detected - child has no agammaglobulinemia
<b>T-CELL DYSFUNCTION</b>		
<b>Immune Dysregulation, Polyendocrinopathy, Enteropathy, X-linked Syndrome (IPEX)</b>	T-Lymphocytes escape apoptosis and proliferate in an uncontrolled manner. (X-linked, caused by mutations in <i>FOXP3</i> ).	- no variants in <i>FOXP3</i>
<b>Hyper-IgM syndrome</b>	Patients possess normal to high concentrations of serum IgM, but markedly diminished IgG, IgE and IgA (X-linked, caused by mutations in <i>CD40LG</i> ).	- no variants in <i>CD40LG</i>
<b>COMPLEMENT DYSFUNCTION</b>		
<b>C1 esterase inhibitor deficiency</b>	A Series of patients developed small bowel enteritis. C1 esterase inhibitor deficiency is usually associated with hereditary angioedema (Mutations in <i>SERPING1</i> ).	- no variants in <i>SERPING1</i>
<b>DEFECTS IN INNATE IMMUNITY</b>		
<b>Blau syndrome</b>	Mutations in nucleotide binding site of <i>NOD2/CARD15</i> , possibly disrupting interactions with LPSs and NF- $\kappa$ B signaling (Mutations in <i>NOD2</i> ).	- nonassociated SNVs detected in <i>NOD2</i> <ul style="list-style-type: none"> <li>o three SNVs in <i>NOD2</i> detected: rs2066842, rs2066843, rs2066845.</li> </ul> - no insertions or deletions in coding regions of target genes detected
<b>Chronic granulomatous disease (CGD)</b>	Impaired killing (faulty $O_2^-$ production) (Mutations in <i>CYBA/p22phox</i> , <i>NCF1/p47phox</i> , <i>NCF2/p67phox</i> ).	- nonassociated SNVs: <ul style="list-style-type: none"> <li>o two coding <i>CYBA</i> SNVs detected: rs4673, rs8053867</li> <li>o one coding <i>NCF1</i> SNV detected: rs62475423</li> <li>o two coding <i>NCF2</i> SNVs detected: rs17849501, rs2274064</li> </ul> - NBT-test CGD negative
<b>CROHN'S DISEASE LIKE MONOGENIC SYNDROMES</b>		
<b>IL10Ra, IL10Rb, IL10</b>		- nonassociated SNVs: <ul style="list-style-type: none"> <li>o two SNVs in <i>IL10RA</i> detected: rs2229113, rs2256111</li> </ul>
<b>XIAP/BIRC4</b>		- nonassociated SNVs: <ul style="list-style-type: none"> <li>o one SNV in <i>XIAP</i> detected: rs5956583</li> </ul> - gene has been Sanger sequenced without finding.

No variants were observed in *BTK* (X-linked agammaglobulinemia), *FOXP3*, (Immune Dysregulation, Polyendocrinopathy Enteropathy, X-linked Syndrome (IPEX)), *WAS* (Wiskott–Aldrich Syndrome),

*CD40LG* (Hyper-IgM syndrome) or *SERPING1* (C1 esterase inhibitor deficiency), hence these diseases were excluded. Common Variable Immunodeficiency (CVID) was excluded as no causative variants could be identified, however *TNFRSF13B* showed three SNVs, two synonymous (rs11078355, rs8072293) and one missense SNV (rs34562254) that affects a conserved nucleotide and is not known to be associated with CVID. The two prediction tools are discordant considering the impact of the SNV, SIFT predicted it to be tolerated while PolyPhen predicted it to be probably damaging. No variants were found in *ICOS* or *TNFRSF13C*, but *CD19* showed one known (rs2904880) and one novel SNV following the recessive model. The known SNV has not been described as pathogenic and was predicted to be benign by both SNV prediction tools (SIFT and PolyPhen). The novel SNV was sequenced using Sanger sequencing and identified as false positive SNV (Figure 11). *NOD2* was screened for variants that are known to cause Blau Syndrome. Three SNVs were detected, but none of them is associated with Blau Syndrome. One of the SNVs was synonymous (rs2066843) while the other two are missense SNVs (rs2066842, rs2066845). The SNV rs2066842 is not described as pathogenic, while rs2066845 is associated with an increased susceptibility to Crohn's disease.<sup>96</sup> Several SNVs were identified in the genes responsible for chronic granulomatous disease (CGD) (*CYBA*, *NCF1* and *NCF2*). The gene for *CYBA* shows two SNVs (rs8053867, rs4673), whereas the first is synonymous and the second a missense SNV that is reported as *CYBA* polymorphism without known pathogenic associations. One missense SNV in *NCF1* (rs62475423) was detected that is not described as pathogenic but has been marked as \*suspected\* in the dbSNP database, as it might be a sequencing artifact. One synonymous SNV (rs17849501) and one nonsynonymous SNV (rs2274064) were detected in *NCF2*. rs2274064 is located in a site of acceleration and is not described as pathogenic nor predicted to be damaging. No known pathogenic variants were identified for CD-like monogenic syndromes involving *IL10* or *XIAP*. Two coding SNVs were identified in *IL10RA*. The first SNV (rs2256111) is synonymous and the second a missense SNV (rs2229113) predicted to be tolerated that has not been associated with disease. *XIAP* shows one coding SNV (rs5956583) that is not known to be pathogenic and that has been predicted by both, SIFT and PolyPhen, to be tolerated. The whole *XIAP* gene was sequenced using Sanger sequencing and no SNV could be confirmed. Thus none of the investigated monogenic syndromes is likely responsible for the child's phenotype.

#### **4.5. Genetic variants with known CD-association and other variants in the associated genes.**

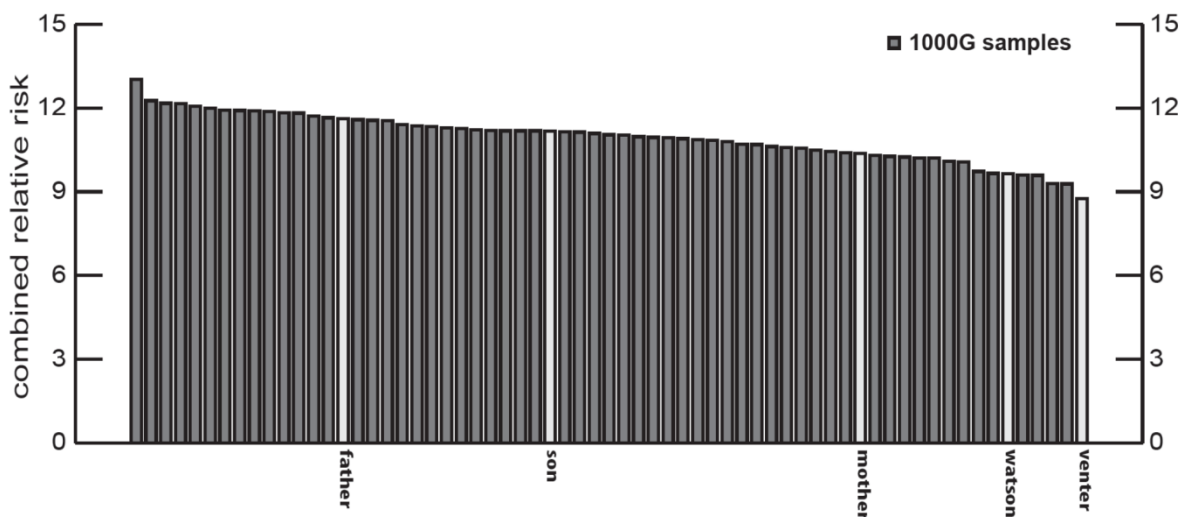
Known risk variants for Crohn's disease might be overrepresented in the child, compared to the healthy parents. To investigate this, CD associated risk variants reported by GWAS studies were identified in the child. Of 71 known CD risk loci<sup>12</sup> 25 were homozygous for the risk-allele in the child, 20 heterozygous and 26 homozygous for the non-risk allele (Table 6). These numbers are comparable to the parents, who showed 18 and 25 homozygous and 25 and 22 heterozygous loci.

**Table 6 Zygoty of the 71 meta-analysis Crohn's disease risk SNVs for the risk-allele.**

Both, father and child carry 25 SNVs homozygous for the risk allele, while the mother has slightly less (18). Instead the mother has few more SNVs heterozygous for the risk allele than the father (22) and child (20). No major differences concerning the known CD risk loci can be observed in the family trio.

	homozygous for risk allele	heterozygous	homozygous for non-risk allele
mother	18	25	28
father	25	22	24
son	25	20	26

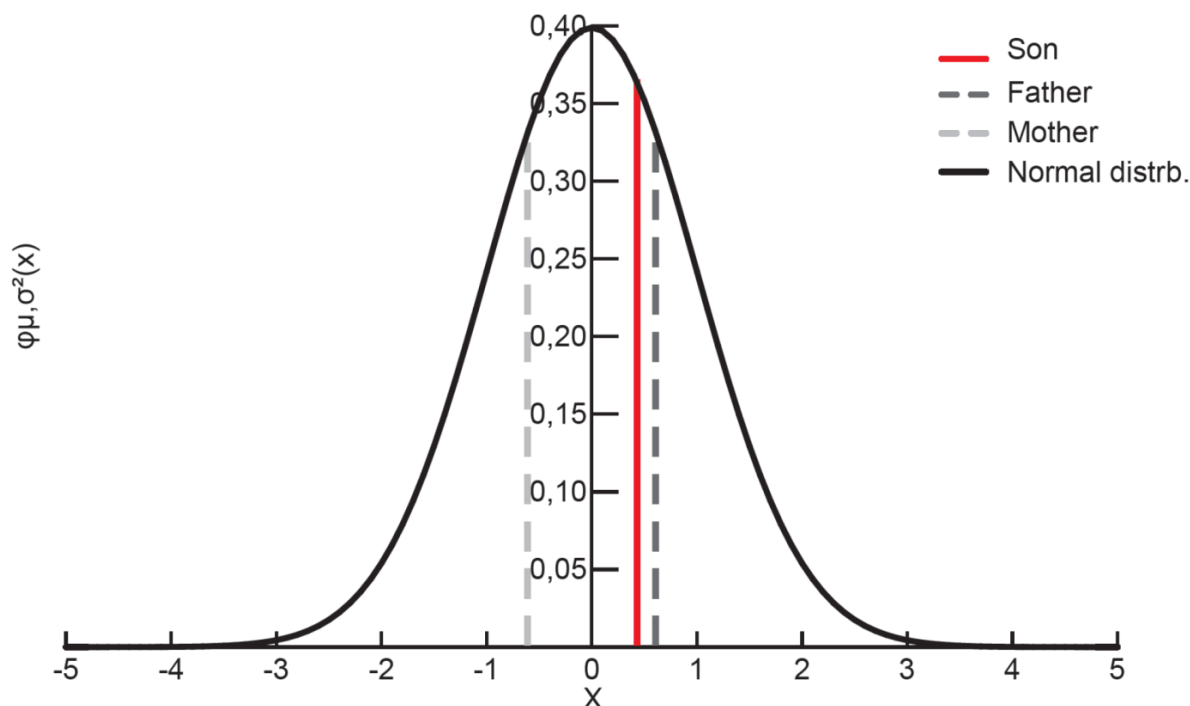
For all three individuals, for the Watson and Venter genomes and for 60 individuals of the 1000 Genome Project the combined risk for the 71 CD risk loci from meta-analysis was calculated (Figure 15). The combined risk calculated for the trio does not exceed the combined risk of other individuals. In the trio, the father shows the highest combined risk, located at the lower end of the upper third of individuals showing the highest combined risk. The son is located in the middle and the mother in the upper end of the lower third.



**Figure 15 Contribution of known CD risk loci to overall risk in comparison with other genomes.**

Combined CD risk of the family trio compared to 60 individuals from the 1000 Genome Project and the Watson and Venter genomes. The family trio does not show major differences in the combined risk than the other genomes. The child's risk lies within its mother's and father's risk.

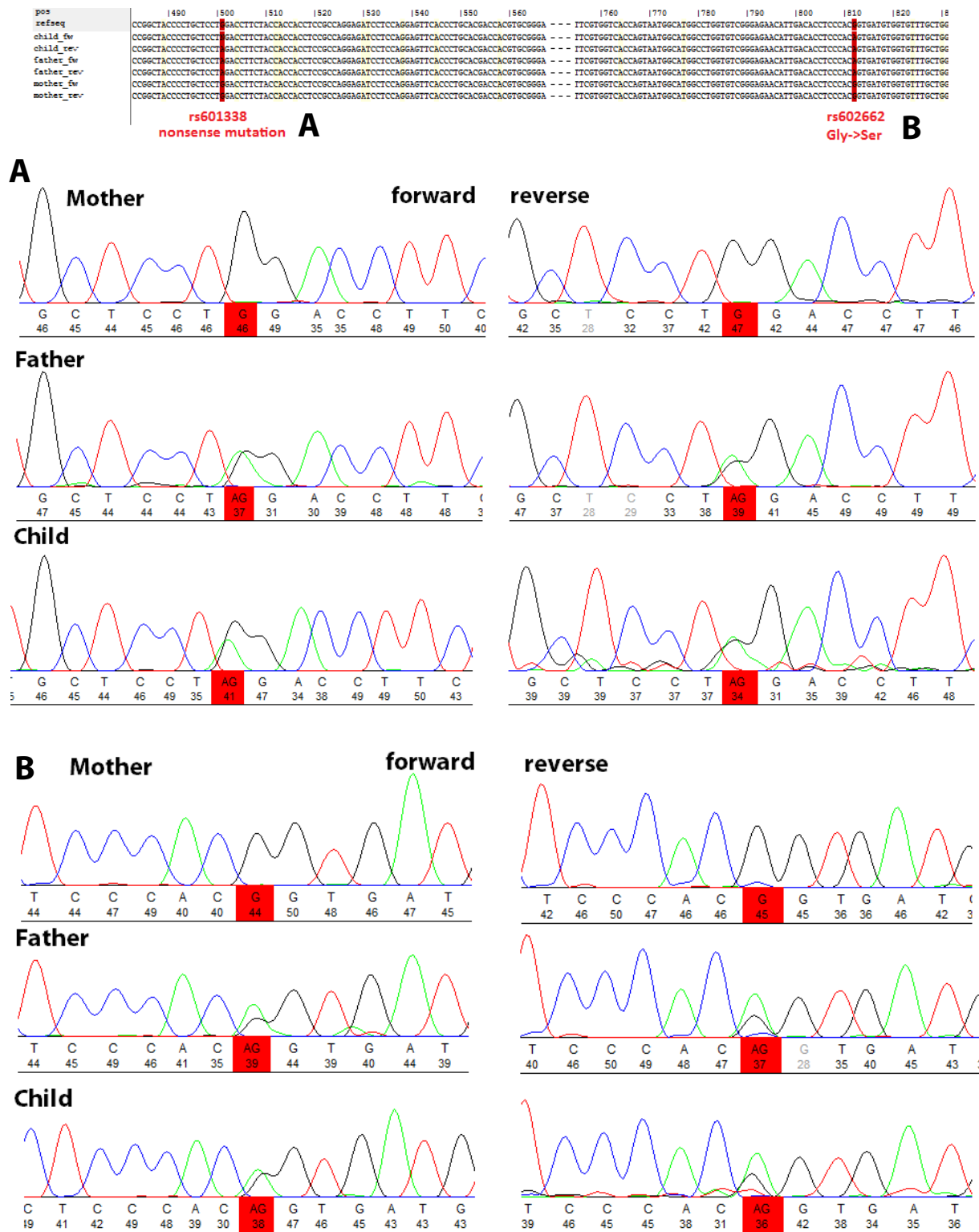
Assuming independent and additive contributions of the known risk alleles at the 71 CD risk loci Gaussian approximation was used to investigate whether the child received a higher CD burden than would be expected by chance (Figure 16, see methods). No significant enrichment of CD variants could be detected in the genome of the child.



**Figure 16** Gaussian approximation assuming independent and additive contributions of the various risk alleles at the 71 known CD-risk loci.

The black line represents the normal distribution for all genotypes possible determined by the parents' genomes. The child's combined risk is shown in red. It does not exhibit an unusual high inherited burden of known CD risk loci and received only slightly more known risk loci than expected, which are more than in the mother (shown in light grey) and fewer than in the father (shown in dark grey).

A larger number of genetic variants were detected within genes in the proximity of the known CD risk loci (Supplementary table 19-25). Three of these SNVs (in *IKZF1*, *CD19* and *MST1*) were already mentioned in the section dedicated to *de novo* variants and excluded as causative variants. One known CD risk conferring variant was detected in *NOD2*. Several additional variant were identified in the genes associated with CD risk loci. This includes a nonsense variant in the fucosyltransferase-2 gene (*FUT2*) that has been detected and verified in both, father and child (Figure 17). Sanger sequencing of the *FUT2* nonsense variant also revealed an additional missense SNV in the *FUT2* gene, which was again detected in the father and child only.



**Figure 17 Sanger sequencing result of *FUT2* nonsense SNV.**

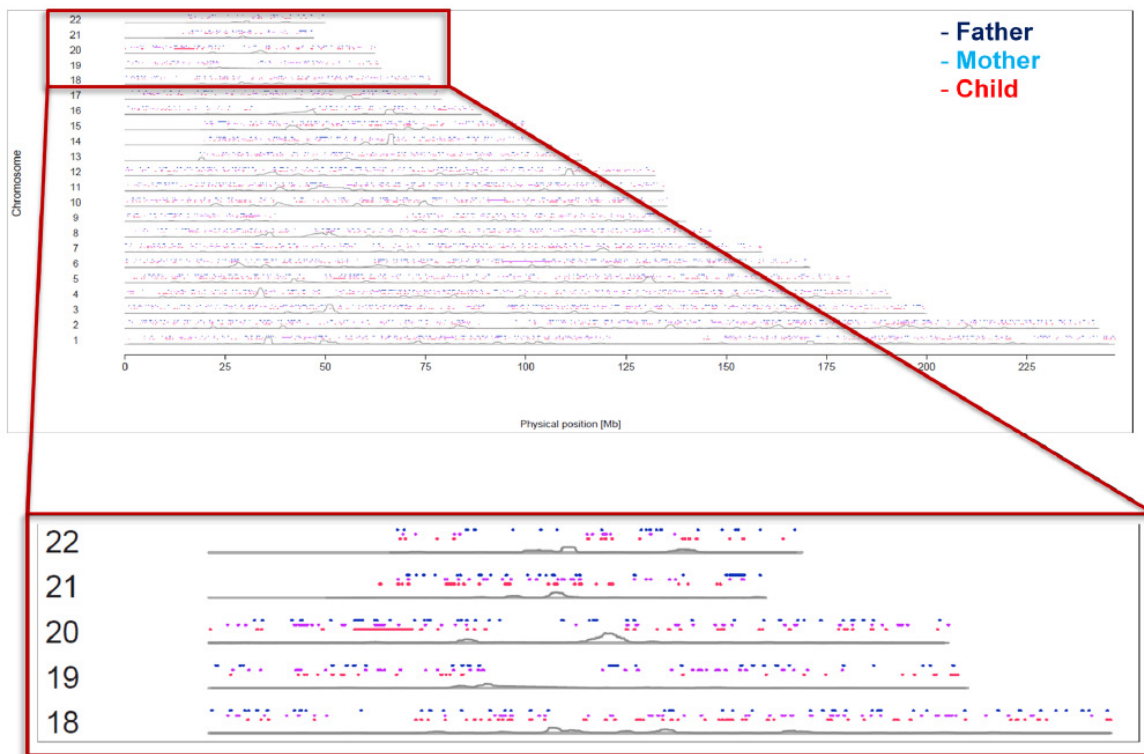
**A** The nonsense mutation could be confirmed and was detected heterozygous in the father and child and homozygous reference in the mother. **B** A second variant (missense) was identified in *FUT2*, which is heterozygous in both, father and child and homozygous reference in the mother.

Another SNV (rs11549656, P200L) affecting a conserved nucleotide was detected in the glutathione peroxidase gene *GPX1*. Also a variant was identified in the glucokinase regulatory protein (*GCKR*, rs1260326), that introduces a new proline. This variant has been strongly associated with lower



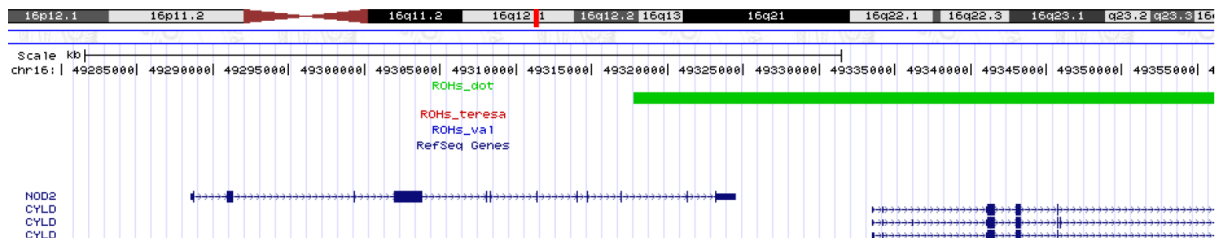
fasting glucose levels and fasting insulin levels and higher triglyceride levels that reduces type II diabetes risk in the French population.<sup>97</sup> Only one SNV that has been predicted to be damaging and follows the recessive model was identified. The SNV is located in a site of high acceleration in *SNAPC4* (small nuclear RNA-activating protein complex). Two small structural variants (sSVs) were observed. The first sSV is a 2-bp deletion in a member of the tumor necrosis factor superfamily (*TNFSF18*) that was not predicted to cause nonsense mediated decay, and was detected homozygous in the mother and child. The second sSV is a 1-basepair insertion in proximity of a splice-site in *DENND1B* (DENN/MADD domain-containing protein 1B) (Supplementary table 25). The variant was detected hemizygous in the father and child.

Regions (or runs) of homozygosity (ROHs) are long stretches without heterozygosity in the diploid state. Multiple studies have identified ROHs associated with other complex traits, such as schizophrenia and late onset of Alzheimer's disease and there is evidence that ROHs can be used to uncover hidden recessive variants in complex disease. This has been proven to be particularly useful in investigating autosomal recessive disorders in populations with a high prevalence of consanguinity. Additionally, larger differences in ROHs could be shown between cases and controls in various studies.<sup>98</sup> Here, the previously suggested GWAS-based definition in PLINK was adapted for identification of ROHs in all three individuals. The overlap of ROHs with the 71 published CD risk loci<sup>12</sup> was calculated by combining the length of ROHs overlapping any risk loci and dividing by the combined length of all ROHs detected. Using a variant-density adapted definition, a comparable number of ROHs was detected in each of the three genomes covering similar proportions of the genome (Figure 18).



**Figure 18** Regions of homozygosity identified in the family trio using a variant-density adapted definition. The child's genome contains 2,220 ROHs (mother: 2,267; father: 2,108) which cover about 222.3 Mb (mother: 234.8 Mb; father: 201.9 Mb). ROHs unique to the child cover 164.6 Mb in total.

The child's genome contained 2,220 ROHs (mother: 2,267; father: 2,108) that covered about 222.3 Mb (mother: 234.8 Mb; father: 201.9 Mb). ROHs unique to the child cover 164.6 Mb in total. These unique ROHs covered CD-associated genes such as *NOD2* (Figure 19), *IL18R1* and *ORMDL3* (Table 7).



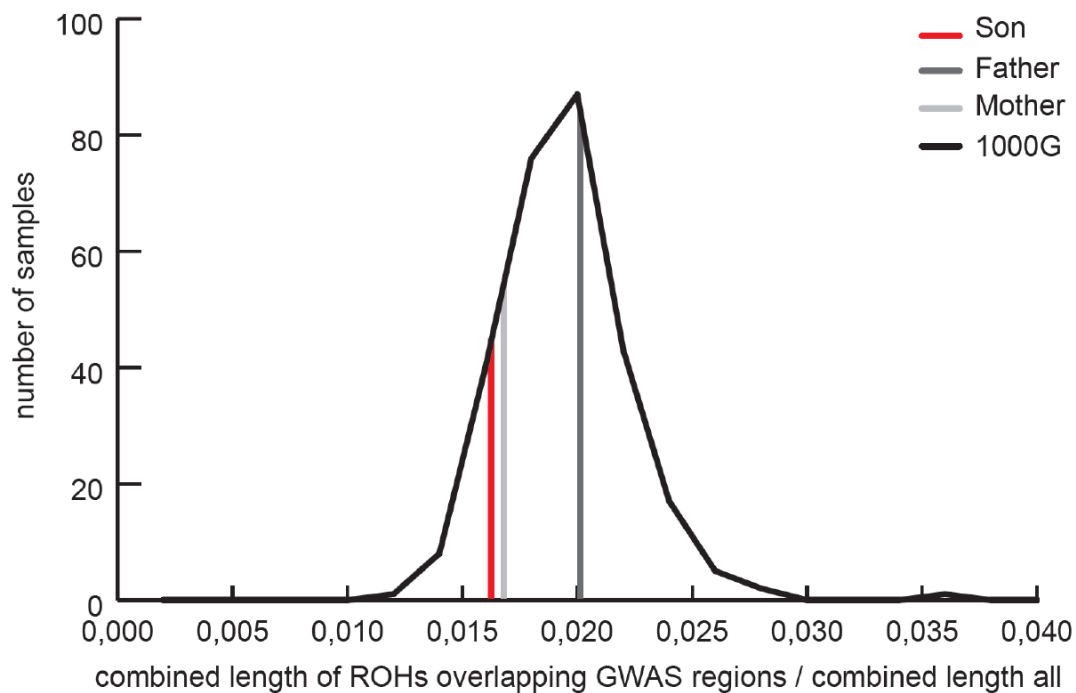
**Figure 19** Example of a region of homozygosity in the child. The region of homozygosity was only detected in the child (dot, green) and overlaps *NOD2*.

**Table 7 Regions of homozygosity overlapping Crohn's disease associated genes.**

A substantial number of ROHs are overlapping genes in regions associated with CD risk. The number of genes completely contained within ROHs is higher in the child (17) compared to the parents (11 and 10), while the number of genes partially inside ROHs is highest in the father (9) and equal in the mother and child (6). In summary, the child has more potential CD risk genes in ROHs than its parents.

RoH overlap with gene				RoH overlap with gene				RoH overlap with gene			
gene	mother	father	child	gene	mother	father	child	gene	mother	father	child
<i>APOB48R</i>	-	-	-	<i>ICAM1</i>	-	-	-	<i>PRDM1</i>	-	-	-
<i>ATG16L1</i>	complete	-	-	<i>ICAM3</i>	-	-	-	<i>PRDX5</i>	-	-	complete
<i>BACH2</i>	-	-	-	<i>ICOSLG</i>	-	complete	-	<i>PTGER4</i>	-	-	-
<i>BSN</i>	-	-	-	<i>IKZF1</i>	-	-	-	<i>PTPN2</i>	-	partial (50%?)	-
<i>C11orf30</i>	-	-	-	<i>IKZF3</i>	-	partial (80%?)	partial (90%?)	<i>PTPN22</i>	-	-	-
<i>C13orf31</i>	-	-	-	<i>IL10</i>	-	-	-	<i>RASIP1</i>	-	-	-
<i>C1orf106</i>	-	-	partial (90%?)	<i>IL12B</i>	complete	complete	-	<i>REL</i>	-	-	-
<i>C2orf74</i>	-	-	-	<i>IL18R1</i>	partial(90%?)	-	complete	<i>RTEL1</i>	-	-	-
<i>SLC22A23</i>	-	-	-	<i>IL18RAP</i>	partial (90%?)	-	complete	<i>SBNO2</i>	-	-	-
<i>CARD9</i>	-	-	-	<i>IL19</i>	-	-	-	<i>SCAMP3</i>	complete	complete	-
<i>CCL11</i>	-	-	complete	<i>IL1RL1</i>	complete	-	complete	<i>SH2B1</i>	-	-	-
<i>CCL2</i>	-	-	-	<i>IL23R</i>	-	partial (90%?)	-	<i>SLC22A4</i>	complete	-	-
<i>CCL7</i>	-	-	complete	<i>IL27</i>	-	-	-	<i>SLC22A5</i>	complete	-	-
<i>CCL8</i>	-	-	complete	<i>IL2RA</i>	-	-	-	<i>SMAD3</i>	-	-	-
<i>CCR6</i>	-	-	-	<i>IL3</i>	-	-	-	<i>SNAPC4</i>	-	-	-
<i>CD19</i>	-	-	-	<i>IRF1</i>	-	complete	-	<i>SP140</i>	-	-	-
<i>CD244</i>	partial (90%?)	-	-	<i>IRGM</i>	-	-	-	<i>STAT3</i>	partial (40%?)	-	-
<i>CDKAL1</i>	partial (10%?)	-	-	<i>ITLN1</i>	-	-	-	<i>TAGAP</i>	-	-	-
<i>CPEB4</i>	complete	complete	-	<i>JAK2</i>	-	-	-	<i>THADA</i>	-	partial (50%?)	-
<i>CREM</i>	-	-	-	<i>KIF21B</i>	-	-	-	<i>TMEM174</i>	-	-	-
<i>CSF2</i>	-	-	-	<i>LRRK2</i>	-	partial (10%?)	partial (9%?)	<i>TNF</i>	-	-	-
<i>DENND1B</i>	-	-	-	<i>LST1</i>	-	-	-	<i>TNFSF11</i>	-	complete	-
<i>DNMT3A</i>	-	-	-	<i>LTA</i>	-	-	-	<i>TNFSF15</i>	-	partial (20%?)	-
<i>EIF3C</i>	-	-	-	<i>LTB</i>	-	-	-	<i>TNFSF18</i>	-	-	-
<i>ERAP2</i>	-	partial (90%?)	-	<i>TAB1</i>	-	complete	-	<i>TNFSF4</i>	-	-	-
<i>ESRRA</i>	-	-	complete	<i>MCCD1</i>	-	-	-	<i>TNFSF8</i>	-	complete	complete
<i>FADS1</i>	-	-	-	<i>MLX</i>	-	-	-	<i>TYK2</i>	-	-	-
<i>FASLG</i>	-	-	-	<i>MST1</i>	-	-	-	<i>UBE2D1</i>	-	-	-
<i>FUT2</i>	-	-	-	<i>MTMR3</i>	-	-	-	<i>VAMP3</i>	-	-	-
<i>GALC</i>	complete	-	complete	<i>MUC1</i>	-	complete	-	<i>YDJC</i>	-	-	-
<i>GCKR</i>	-	-	-	<i>NCR3</i>	-	-	-	<i>ZFP36L1</i>	complete	-	complete
<i>GPR65</i>	complete	-	complete	<i>NDFIP1</i>	-	-	complete	<i>ZMIZ1</i>	-	-	-
<i>GPX1</i>	-	-	-	<i>NKX2-3</i>	complete	-	complete	<i>ZNF365</i>	-	-	-
<i>GPX4</i>	-	-	-	<i>NOD2</i>	-	-	partial (20%?)	<i>ZPBP2</i>	-	partial (10%?)	complete
<i>GSDMB</i>	-	-	complete	<i>ORMDL3</i>	-	-	complete	<i>MUC19</i>	-	complete	partial (80%?)
<i>HLA-DQA2</i>	-	-	-	<i>PLCL1</i>	-	-	-	<i>NELL1</i>	partial (10%?)	partial (10%?)	partial (9%?)
<i>total no overlap</i>	92	90	86	<i>total complete overlap</i>	11	10	17	<i>total partial overlap</i>	6	9	6

As there could be an over-representation of known CD risk loci in ROHs detected in the child in comparison with putative healthy individuals, the overlap of ROHs with CD risk loci was calculated for ROHs detected in 60 CEU individuals of the 1000 Genome Project using the same criteria for ROHs. The size of the overlapping regions of the family trio was compared to the overlap of the 60 1000 Genome individuals (Figure 20). No significant differences can be observed between the data sets.



**Figure 20** Overlap of regions of homozygosity with GWAS annotated CD risk regions of the family trio in with unrelated individuals.

Regions of homozygosity (ROHs) for 60 individuals of the 1000 Genome Project were detected with the same definition of a ROH as has been done for the family trio. Compared to the unrelated individuals the family trio does not show any major differences in the ROH distribution concerning CD risk regions. The child (shown in red) does not have a higher proportion of ROHs in CD risk regions than its parents.

#### **4.6. Genetic variants in genes associated with other inflammatory diseases**

Many risk genes associated with inflammatory traits are shared between multiple diseases. Therefore it is possible that genetic variants in genes involved in other inflammatory diseases could contribute to the child's phenotype. Based on the GWAS catalog, risk loci and genes associated with the following diseases were investigated: ankylosing spondylitis, asthma, celiac disease, inflammatory bowel disease (IBD, which is a distinct entry to ulcerative colitis and CD in the GWAS catalog), leprosy, multiple sclerosis (MS), psoriasis, rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), type 1 diabetes and ulcerative colitis (UC). Genetic variants identified in the associated genes are reported in Supplementary table 26-32. Nine genes carry variants in conserved regions and five of them were predicted to be damaging by both applied prediction

tools (*NOD2*, *MST1* (already excluded by Sanger sequencing), *ITGAX*, *IRF5* and *ERBB3*). Four of the detected variants are not included in dbSNP130 (*FRMD4B*, *IRF5*, *KIAA1109* and *LRRK2*) and might represent novel, rare variants. Five large SVs were identified in proximity of the investigated inflammatory genes, with only one SV involving the coding region of the growth hormone receptor (GHR). The genetic variants affecting genes involved in other inflammatory diseases could hint towards shared inflammation associated risk factors that could be involved in the pathogenesis of the presented case study.

#### 4.7. SNVs predicted to be damaging including all genes

Disease causing variants could occur in genes that have not been linked to the pathogenesis of inflammatory disease yet. Hence it is important to screen for genetic variants in all genes, independent on their described function. As this leads to a vast number of variants, further filtering is required. Here, the focus was put on potential damaging variants that contribute to compound heterozygosity, follow the recessive model of inheritance or are located in sites of high conservation. Among the SNVs following the recessive model, only 15 were identified in sites of conservation (Supplementary table 33). Ten of them show minor allele frequencies >0.2 (*BRWD1*, *C2orf73*, *GBE1*, *GOLGA4*, *LOXL4*, *NOM1*, *OBSL1*, both SNVs in *PCDHA1* and *PSDM9*). The remaining five were detected in *PCDHB7*, *MPHOSPH8*, *KRT76*, *KCNJ12* and *BEST2*. Besides *MPHOSPH8* and *BEST2* all SNVs are included in dbSNP130. The *BEST2* SNV has been verified by Sanger sequencing (Figure 21).

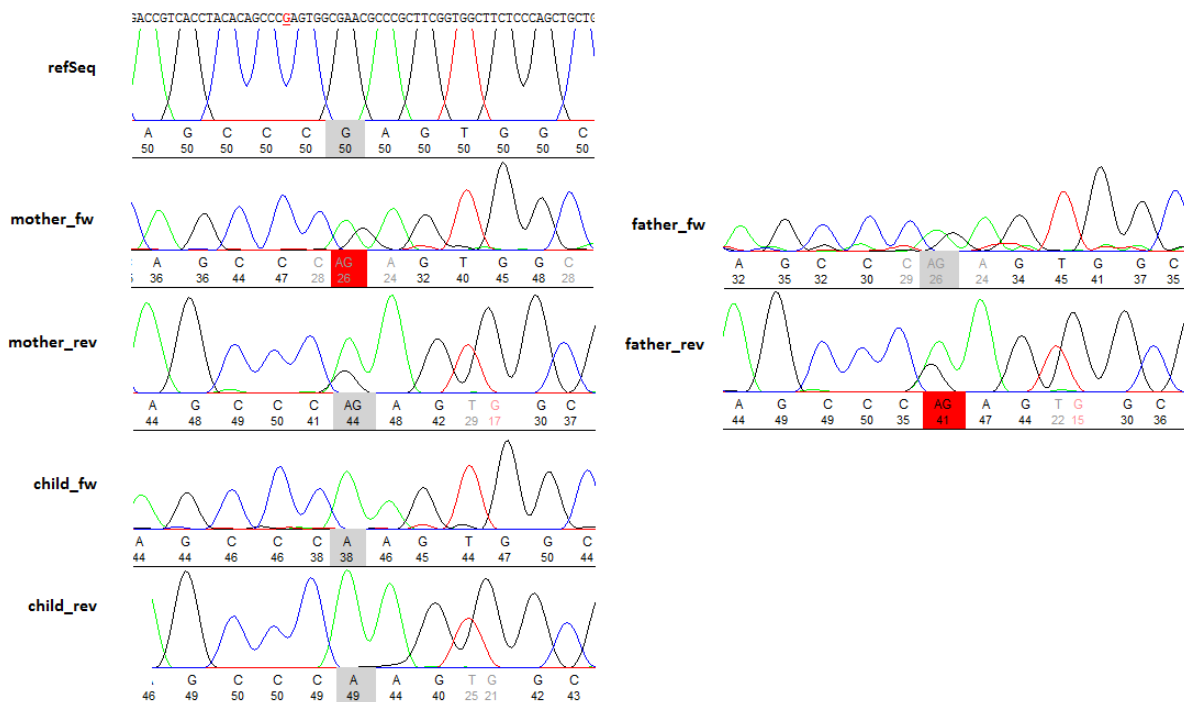


Figure 21 Verification of *BEST2* SNV by Sanger sequencing.

Both parents show a heterozygous genotype (A/G) while the child is homozygous for the alternative allele A.

Expression of *BEST2* could only be detected in the colon biopsy of the child, but not in the blood of any individual of the trio, while *MPHOSPH8* expression was detected in all samples (Supplementary table 34). Additionally the occurrence of the *BEST2* variant in the 1000 Genomes data set (1092 samples, Oct. 2011 release) was assessed. Only 84 individuals heterozygous for the variant and four homozygous individuals were identified, suggesting a very low allele frequency across the investigated human ethnicities. Encouraged by the low occurrence in the general population, the SNV was genotyped in large cohorts of Crohn's disease and ulcerative colitis patients that did not confirm the potential association of this SNV with disease risk (Table 8 and Table 9).

**Table 8 Plink Allelic Association Test with confidence interval and case/control-only.**

No significant p-value was detected for an association of the *BEST2* SNV with CD or UC. The odds ratio (OR) is only slightly higher (1.037) than for neutral risk association (1). A1 minor allele, A2; major allele; F\_A frequency in cases; F\_U frequency in controls; SE standard error; L95/U95 lower and upper bound of 95% confidence interval for odds ratio.

trait	CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR	SE	L95	U95
Crohn's Disease	19	BEST2_R8Q	12863429	2	0.07873	0.07616	1	0.1168	0.7325	1.037	0.1053	0.8433	1.274
Ulcerative Colitis	19	BEST2_R8Q	12863429	2	0.06671	0.07601	1	1.312	0.2520	0.8689	0.1227	0.6832	1.105

allele 1 = G, allele 2 = A

**Table 9 Plink Hardy-Weinberg calculation.**

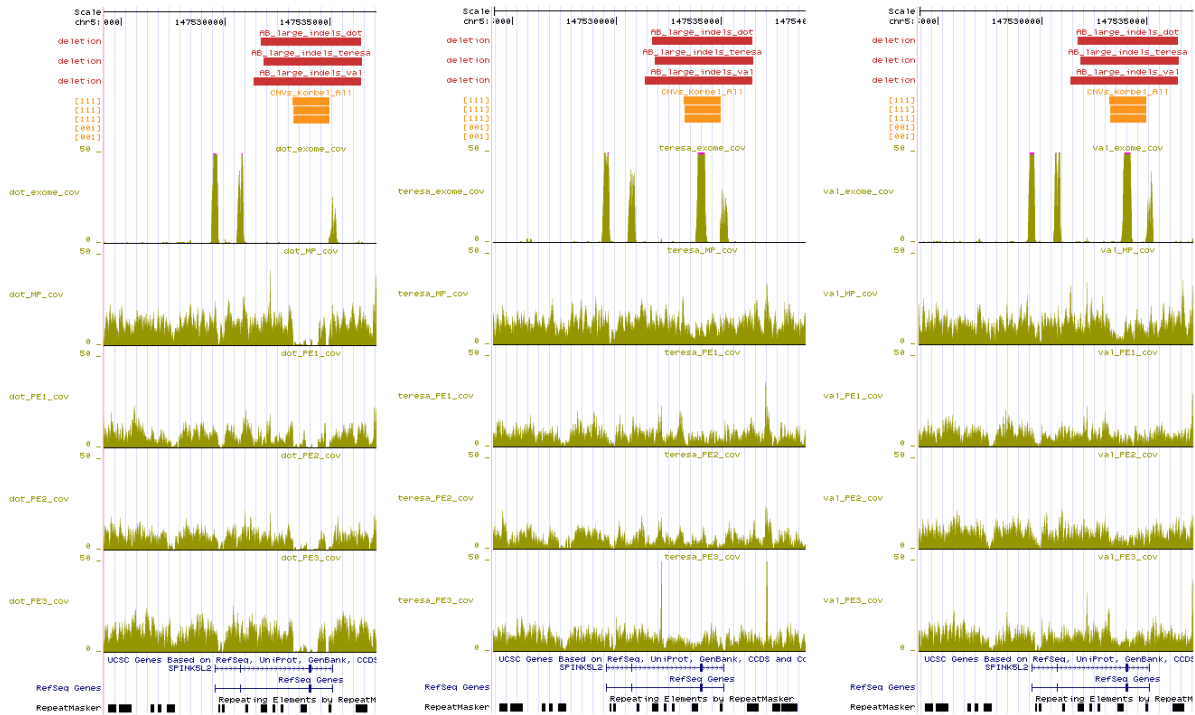
No significant differences concerning the observed heterozygosity between the cases and controls were detected. A1 minor allele; A2 major allele; GENO genotype; O(HET) observed heterozygosity, E(HET) expected heterozygosity. The number of individuals homozygous for this variant is generally very low (10 of 2537 = 0.39%), but indifferent between cases (0.38%) and controls (0.41%).

CHR	SNP	TEST	A1	A2	GENO*	O(HET)	E(HET)	P
19	BEST2_R8Q	ALL	2	1	10/370/2157	0.1458	0.1419	0.2062
19	BEST2_R8Q	AFF	2	1	4/159/903	0.1492	0.1444	0.3944
19	BEST2_R8Q	UNAFF	2	1	6/211/1251	0.1437	0.1404	0.4577

\*Genotype counts: 11/12/22

Several genes show SNVs that possibly contribute to compound heterozygosity, whereas only few of these SNVs include at least one SNV that was predicted to be damaging (Supplementary table 35) or were not predicted at all. The *HLA-A* gene and three mucin genes (*MUC2*, *MUC4* and *MUC16*) carry at least one SNV predicted to be damaging from either mother or father and additional SNVs from the other parent. For the remaining genes with compound heterozygosity only few have known functions (Supplementary table 36). Multiple genes that are likely involved in immune system processes were identified by screening for genetic variants in conserved exonic sites (Supplementary tables 37-40). Three large deletions were identified that probably follow the recessive model of inheritance affecting exons of *SPINK5L2/SPINK14*, *GUCY2GP* and *CCDC135* (Supplementary table 13). The *SPINK5L2* deletion is shown in Figure 22 and nicely demonstrates the loss of exon three in the child's genome, while the parents both still have the respective region, at

least in a hemizygous state. The data is supported by all sequenced libraries, which is a strong proof for the existence of this variant.

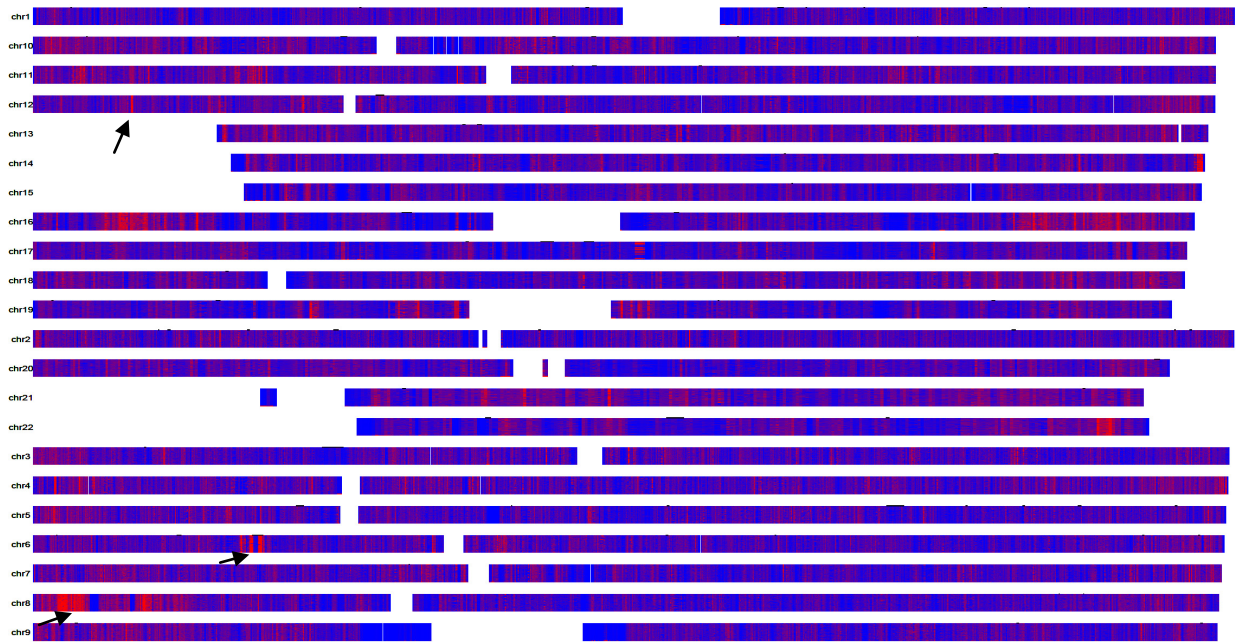


**Figure 22** Large deletion affecting the *SPINK5L2* exon 3.

The deletion occurs homozygous in the child (dot) and is either hemizygous in the parents (recessive model of inheritance) or *de novo* in the child. All five libraries (the exome library, long mate-pair library (MP), and the three paired-end (PE) libraries) in the child support the loss of the third exon.

#### 4.8. Identification of mutational hotspots

In an approach to identify mutational hotspots in the genomes of the family trio compared to 60 individuals of the 1000 Genome Project the number of SNVs in a given region for all individuals was plotted (Figure 23). The SNV density is reflected by a color gradient (red = high SNV density, blue = low SNV density).



**Figure 23** SNV density plot for 60 individuals (top 60 lines) from the 1000 Genome Project and the family trio (lowest three lines).

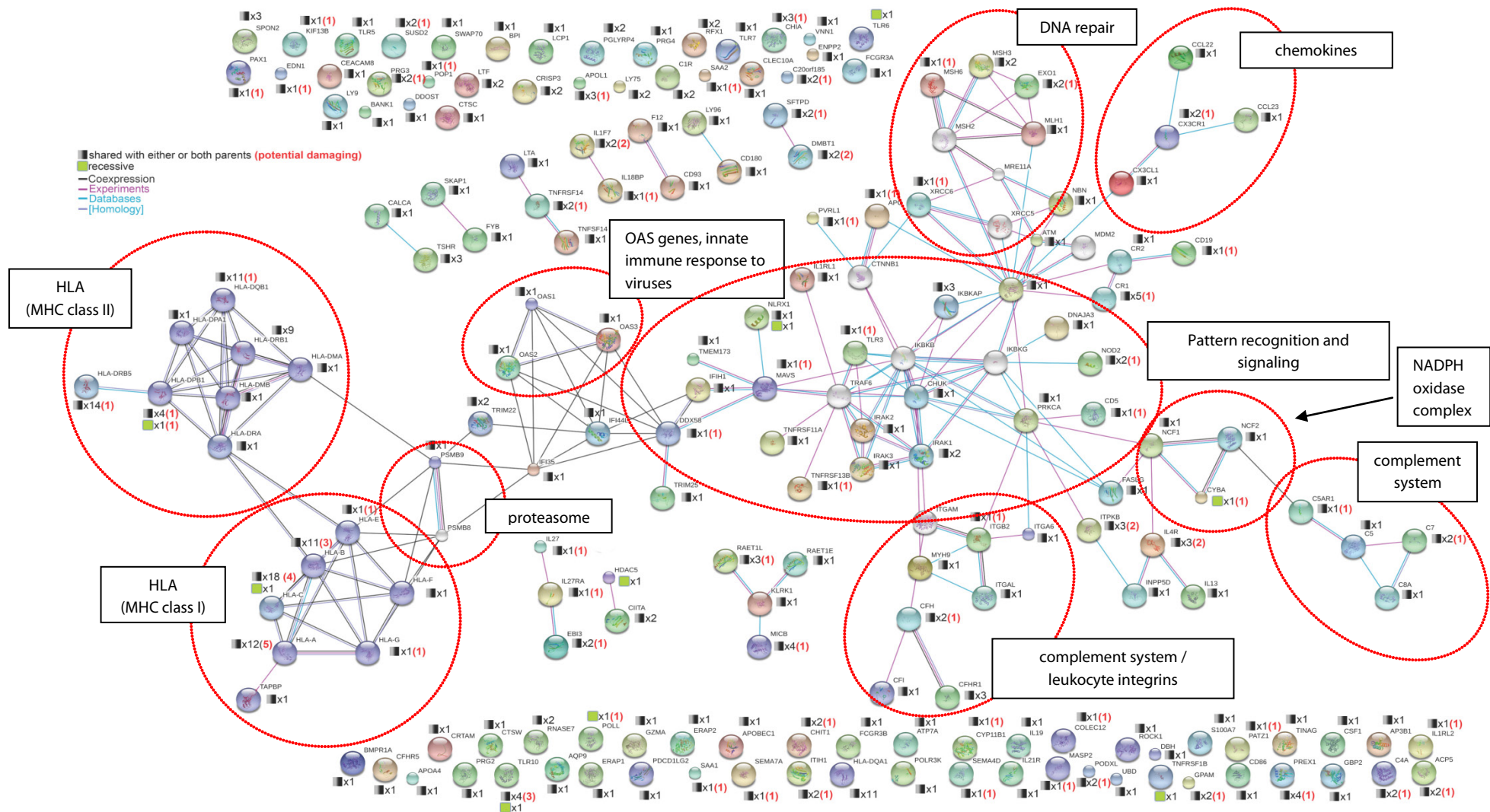
Mutational hotspots carrying many SNVs are shown in red, while regions with few SNVs are shown in blue. No obvious differences between the family trio and the 60 individuals are visible, except for the regions adjacent to the centromer and other repeat rich regions, which are probably caused by mapping problems. Multiple locations involved in immune processes can be identified as red spots (arrows). These locations include the highly polymorphic HLA region on chr6, a defensin containing region at the beginning of chr8 and a lectin containing region in chr12p. The black lines above the density plot for each chromosome represent CD risk regions identified by GWAS. A high resolution PDF file is digitally available on the accompanying CD (compact-disc).

Several known mutational hotspots could be identified using this approach, i.e. the MHC regions on chromosome 6, a defensin containing region on chromosome 8 and a lectin containing region on chromosome 12. However, no mutational hotspots unique to the trio could be identified, except for highly repetitive regions (i.e. the centromers) that are probably subject to problematic mapping.

#### **4.9. Pathway analysis of genes affected by missense variants in the child**

All genes with missense SNVs were filtered by the gene ontology term “immune system process” and used as input for a functional protein network generated with STRING (<http://string-db.org/>) (Figure 24).





**Figure 24** String network of identified missense SNVs in the child which are contained in the gene ontology term “immune system process”.

SNVs predicted to be damaging are shown in red and brackets. Recessive SNVs are marked with green boxes and SNVs shared with either parent in grey boxes. Only links by co-expression, experiments, data bases and homology were allowed, but not data mining, gene fusions and neighborhood. Ten white nodes (not included in the missense SNV list) were added for a better network. A high resolution copy of this picture is digitally available on the accompanying CD (compact-disc).

While many genes remained without connection, several clusters in antigen recognition and processing (HLA genes, MHC class I and II), proteasome function (*PSMB9*), innate immune responses to virus infections (2-prime,5-prime oligoadenylate synthetases (*OAS*) genes), DNA repair (*XRCC5*, *MSH3*, *MSH6*, *MLH1*, *EXO1*), chemokines (*CCL22*, *CCL23*, *CX3CR1*, *CX3CL1*), complement system (*C5AR1*, *C5*, *C7*, *C8A*; *CFI*, *CFHR1*, *CFH*; *CR1*, *CR2*), NADPH oxidase complex (*CYBA*, *NCF1*, *NCF2*) and pattern recognition and signaling associated pathways (*NOD2*, *TLR3*, *NLRX1*, *IL1RL1*, *TNFRSF11A*, *TNFRSF13B*, *IRAK1*, *IRAK2*, *IRAK3*, *IKBKAP*, *CHUK*) were identified. Only few genes carry putative recessive SNVs (*HLA-DPB1* (potential damaging SNV), *HLA-C*, *HDAC5*, *TLR6*, *TLR10*, *TNFRSF1B*, *POLL* (potential damaging SNV), *CYBA* (potential damaging SNV) and *NLRX1*. The number of non-recessive SNVs predicted to be damaging is much higher and these SNVs affect all above mentioned clusters, except *PSMB9* and the *OAS* genes.

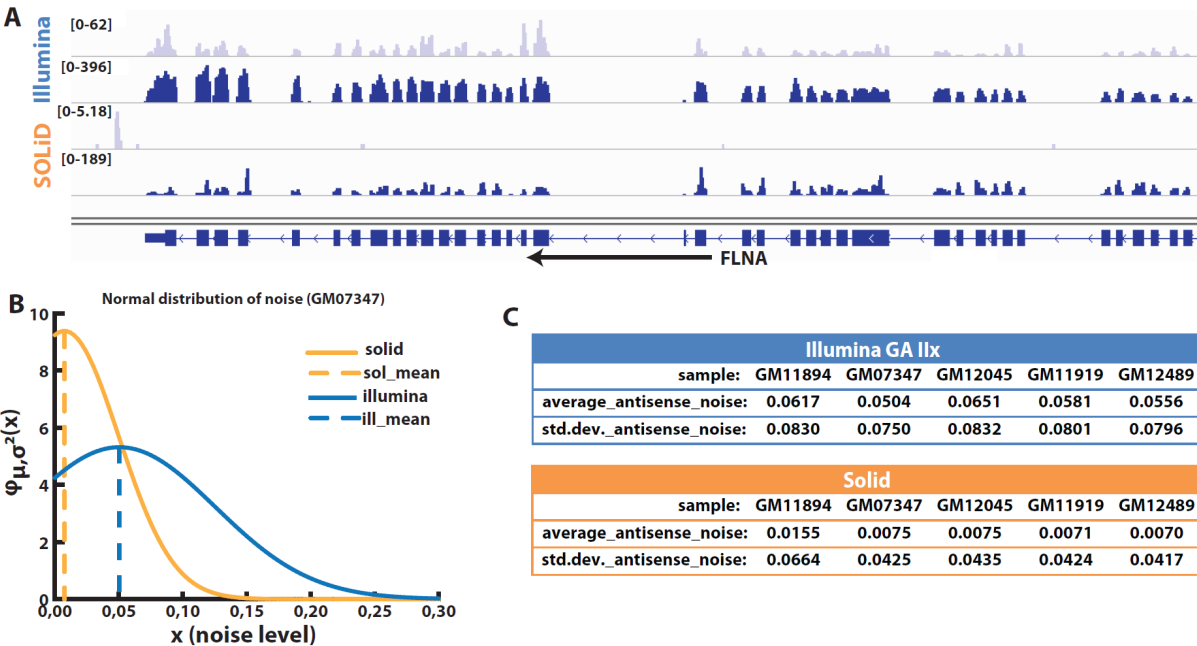
#### **4.10. Expression changes in the child compared to the parents and genetic variants in the respective genes.**

The transcriptomes of all three family members were analyzed to identify differentially expressed genes. Expression levels were determined with the widely used Cufflinks software<sup>89</sup> that are then reported as fragments per kilobase of exon model per million mapped fragments (FPKM). In a filtering step, genes were identified that show a 5-fold up- or downregulation in the child compared to the parents, leading to 75 upregulated and 12 downregulated genes. Among these are many genes involved in inflammation, i.e. several genes are associated with immune cell function and development (*CD177*, *C19orf59*, *ARG1*, *CST7*, *TNFSF13B*, *IL4*), and with pattern recognition and associated downstream signaling (*CLEC4D*, *HMGB2*, *TLR5*, *NLRC4*, *IRAK3*, *CR1*) (Supplementary table 41). Comparing regulated transcripts with genetic data, only a very limited number of genetic variants was observed in these genes, i.e. in *IL4R* and *MMP9* (Supplementary table 42). Yet, no obvious regulatory variant was detected that would functionally link expression differences with genetic variation.

#### **4.11. Strand specific transcriptome analysis**

Extending the transcriptome work done in the trio, a method for assessing sense/anti-sense transcription events was developed that allows integrating transcriptomic data with genetic and epigenetic information. As a benchmarking data set the expression of closely occurring sense/antisense (S/AS) transcript pairs was investigated in transcriptomes derived from lymphoblastoid (LBL) cell lines of putative healthy individuals (five individuals of the 1000 Genome project),<sup>99</sup> three classical Hodgkin lymphoma (cHL) cell lines (kindly provided by Ohle Ammerpohl and Reinert Siebert)<sup>100</sup> and transcriptomes of murine small intestine and colon which were generated in duplicates.<sup>101</sup> As no tool is publically available to assess S/AS transcription on the

genome-wide level using next generation sequencing technologies and S/AS expression may represent an important regulatory mechanism that may determine transcript levels beyond cis-linked transcription factor binding motifs I developed *Janus* to fill this gap. *Janus* was applied to the transcriptomes mentioned above to demonstrate its use in detection of S/AS pairs in general and identification of differentially expressed S/AS pairs. Data for five LBL cell lines was generated using two different technologies (SOLiD whole transcriptome analysis kit (WTAK)<sup>101</sup> and Illumina GA Iix (kindly provided by Ralf Sudbrak and Hans Lehrach).<sup>102</sup> Prior identification of S/AS pairs, comparison of the performance of both technologies revealed a higher background noise in the Illumina data, which manifests as expression in ‘exonic’ regions on the strand opposite to the template strand (the “antisense” strand). An example is given in Figure 25A, where the expression of *FLNA* (second strand) generates an echo on the opposite (first) strand in the Illumina data. The exon-intron structure is preserved on the first strand, which is highly unlikely, given conserved splice-site associated motifs. The SOLiD data on the contrary shows almost no expression on the first strand. Therefore, the noise level was investigated in all exonic regions by calculating the mean and standard deviation of AS noise across exons (Figure 25B and C).



**Figure 25 Antisense noise-level estimation.**

**A** Expression shown for the same transcript with Illumina data (top) and solid data (bottom). First strand expression is shown in light blue and second strand expression in dark blue and the transcript (*FLNA*) is located on the second strand. Other than the Illumina data, SOLiD data shows almost no expression on the first strand. The Illumina coverage on the first strand shows the same exon-intron pattern as the expression of the second (sense) strand. **B** Plotted mean and standard deviation for individual GM07347, the solid data shows a smaller mean and std. dev. than the Illumina data. **C** Mean and standard deviation of noise level calculation for SOLiD and Illumina GA Iix data in numbers for all five LBL cell lines. In the SOLiD data, about 0.7 to 1.6% of coverage in exonic elements considered for noise calculation is located on the antisense strand whereas this contributes to 5.0 to 6.5% of coverage in the Illumina Data.

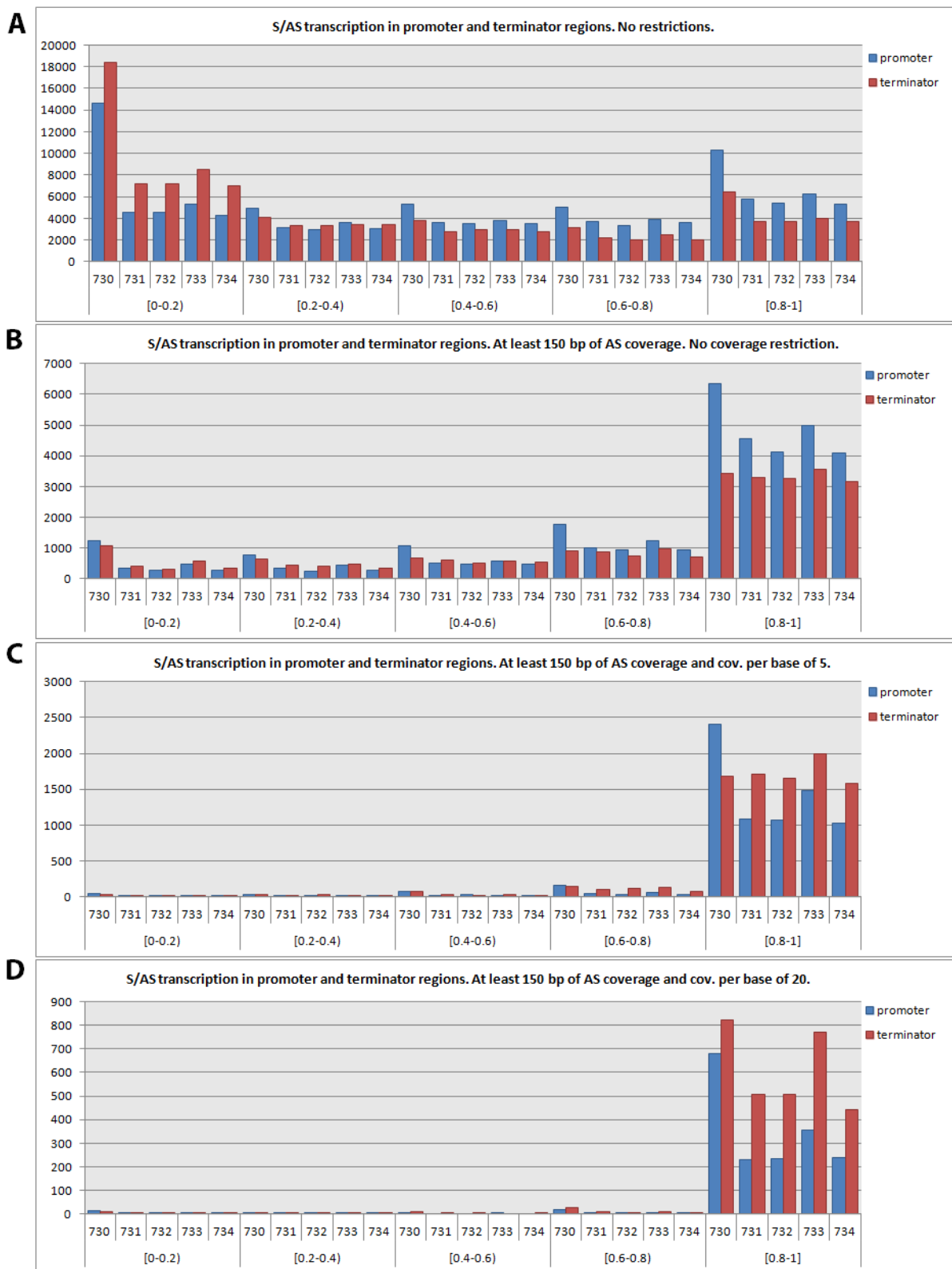
Previous publications mention a high abundance of antisense tags in both the promoter and terminator regions of genes.<sup>78,103,104</sup> Therefore, the occurrence of antisense transcription was investigated in promoter and terminator regions, here defined as 1 kb before and after the transcript start and end site. Using cutoff criteria of a minimum length of 150 bp and a mean coverage per base of five for the antisense transcripts, antisense expression was identified in 740 promoter and 980 terminator regions in the cHL (hg18 mapped data, 32.110 annotated transcripts) and 2100 promoter and 2400 terminator regions in the LBL data set (hg19 mapped data, 66.065 annotated transcripts) (Table 10).

**Table 10** Number of antisense events in promoter and terminator regions.

Promoter and terminator regions were defined as 1 kb before and after transcription start and end site. Valid antisense events required a minimum length of 150 bp and a minimum coverage per base of 5. The number of AS events is slightly higher in terminator regions and the hg19 gene annotation (LBL cell lines) yielded more events than the hg18 gene annotation used for the cHL cell lines.

Sample	AS events in promoter	AS events in terminator
GM11894	2726	1981
GM07374	1794	2461
GM12489	1812	2411
GM12045	2312	2767
GM11919	1812	2352
KM-H2	614	894
L-1236	1083	1215
U-H01	586	823

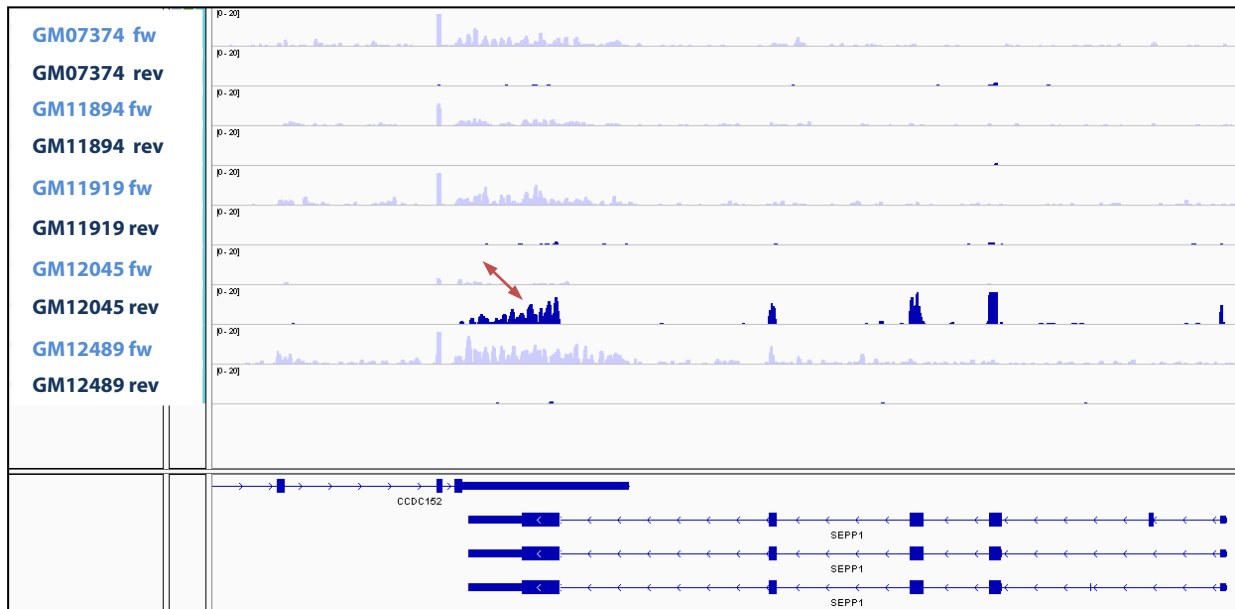
When considering antisense events with low coverage in the promoter and terminator regions, the amount of antisense tags in promoter regions is higher than in terminator regions, while this shifts towards more antisense tags in terminator regions when considering only higher covered antisense events (Figure 26).



**Figure 26** Antisense transcription in promoter and terminator regions of annotated transcripts for five samples (730-734 = GM11894, GM07374, GM12489, GM12045, GM11919).

Promoters were defined as up to 1 kb upstream of the transcription start and the terminator region up to 1 kb downstream from the transcription end. The sense to antisense ratio is shown as a value between 0 and 1, where 0 is pure sense and 1 pure antisense transcription. **A** Antisense to sense ratios without restrictions. **B** AS:S ratios in promoter and terminator regions which have at least 150 bp covered by antisense reads. **C-D** AS:S ratios in promoter and terminator regions which have at least 150 bp covered by antisense reads and a minimum AS coverage per covered AS base of 5 and 20 respectively. With increasing minimum AS coverage, the number of terminator regions increases relative to the number of promoter regions.

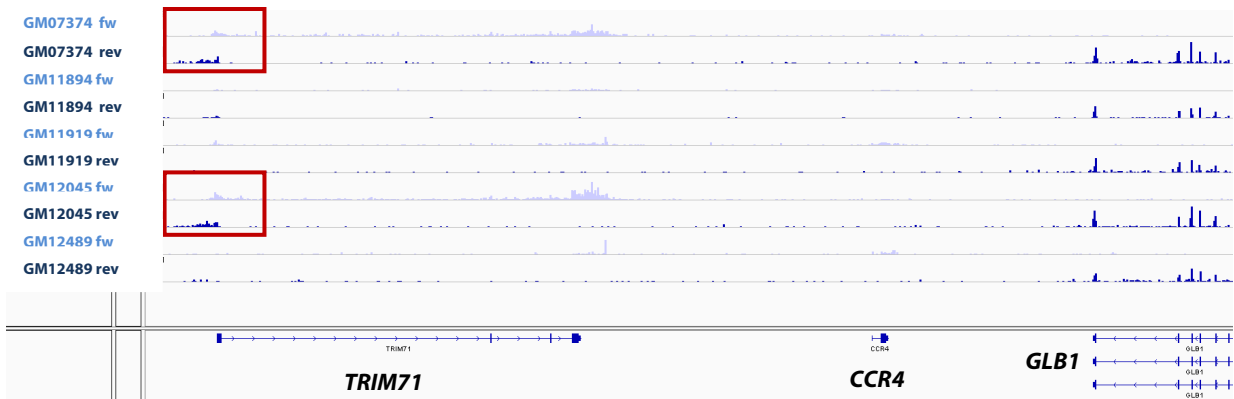
Among the huge amount of detected S/AS pairs, multiple pairs could be detected which show mutually exclusive expression of one transcript in some samples, and mutually exclusive expression of the other transcript in others. This kind of S/AS pairs was identified in all three data sets. Among the best hits in the LBL data set, a S/AS pair consisting of *SEPP1* and *CCDC152* was identified (Figure 27). Four individuals show expression of *CCDC152* but not of *SEPP1*, while for the remaining individual the opposite is true.



**Figure 27** Differentially expressed S/AS pair in LBL cell lines.

*SEPP1* and *CCDC152* lie in antisense orientation to another. Expression levels in LBL cell lines from five individuals are shown for the first strand (light blue) and second strand (dark blue). Either *CCDC152* or *SEPP1* show expression, but not both in the same individual.

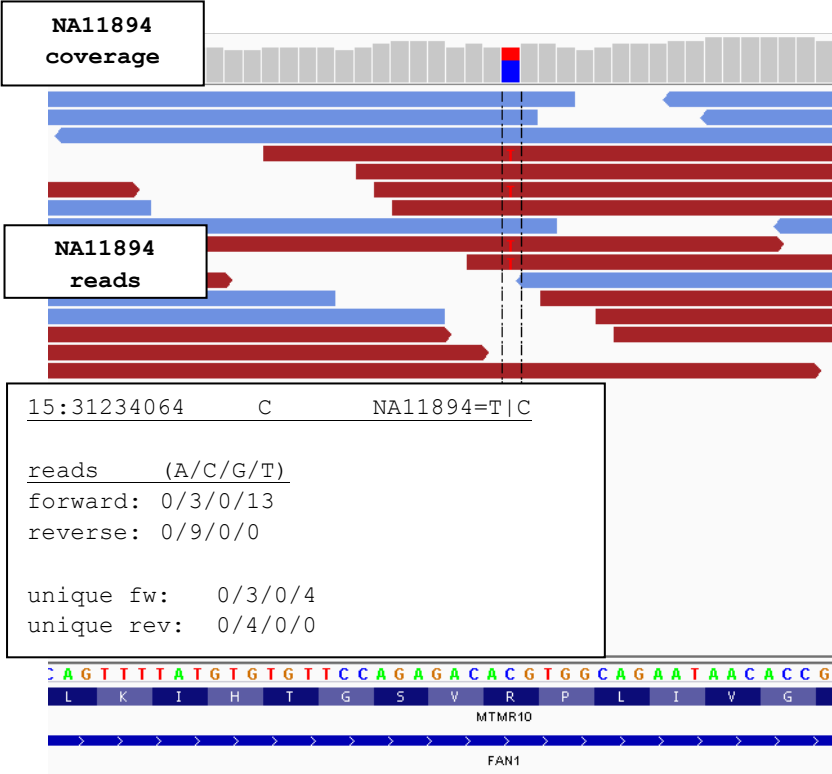
While the number of differentially expressed S/AS pairs was very low in the LBL data set, a slightly larger number of differentially expressed S/AS pairs could be detected in the cHL data set. Supplementary figure 5 shows an example for this data set. Two of the three cHL cell lines show expression of *CD59596* but no *FSTL5* expression, while in the remaining sample only *FSTL5* is expressed. Additionally differential S/AS expression does not only occur between individuals, but also between different tissues. Expression in antisense direction to *Mtus1* was detected in the two murine small intestine samples that could correspond to *B4300010123Rik* expression, but this expression is not detectable in the two murine colon samples. Also, expression of the 'long' *Mtus1* transcript seems to be slightly diminished in the small intestine samples (Supplementary figure 6). Besides differentially expressed S/AS pairs, S/AS pairs that show similar expression for both transcripts but differ in expression per sample were identified. An example of a yet unannotated transcript antisense to *TRIM71* is shown in Figure 28. *GLB1* shows a similar expression rate in all samples while *TRIM71* and the unknown antisense event show a positively correlated expression, which differs between samples.



**Figure 28** Co-expressed S/AS pair in LBL cell lines. The antisense event close to the 5'-end of *TRIM71* (left side, dark blue) is expressed at a similar rate as *TRIM71* (*uc003cff.2*, light blue) itself. In contrast, the expression of *GLB1* (right side) is more even between samples. Scale 0-15 for all tracks.

Additional co-expressed S/AS pairs were identified in the cHL data set (Supplementary figure 7) and in the different murine tissues (Supplementary figure 8). The antisense transcript in the cHL data is novel and no gene has been assigned to this position yet. The antisense transcript identified in the murine colon samples is overlapping with the *Tcf7l1* gene and is not detectable in the small intestine samples, where also expression of *Tcf7l1* is barely detectable. In this case the antisense transcript is also not included in current gene annotation databases.

By investigating known SNV positions (generated by the 1000 Genomes Consortium) in the LBL data set, some SNVs showing signs of strand specific allelic imbalance (Figure 29) were identified.



**Figure 29** IGV picture of SNV at chr15:31234064, only reads with unique start points shown. Reads corresponding to *FAN1* expression (red) show both, the C and T allele, while reads corresponding to *MTMR10* expression show exclusively the C allele.

The example shows expression of *FAN1* and *MTMR10*. Reads corresponding to *FAN1* expression show both SNV alleles (T/C) with slight favor of the T-allele, while the reads corresponding to *MTMR10* expression show the C allele only. Antisense transcripts are likely to have regulatory effects; either by influencing gene expression (i.e. by alteration of histone methylation) or by formation of RNA-RNA-duplexes (due to overlapping antisense transcripts). RNA-RNA-duplex formations involving N-myc have been suggested to cause retention of intron 1, linking antisense transcripts with splicing.<sup>105</sup> Further, double stranded RNA is a feature of multiple viral infections, and thus mechanisms exist to detect and destroy dsRNA fragments (i.e. siRNA can be involved in this process). Therefore, disturbances in the RNA regulatory machinery can cause disease phenotypes, thus it is necessary to gain more knowledge about antisense transcript function.



## 5. Discussion

This thesis focuses on the application of next-generation sequencing to unravel functional genetic variants that could be linked to inflammatory bowel disease using the example of a Falk-Rubinstein trio with early onset CD in the affected child. I present the first study aiming for the extraction of the full genetic variability in a Crohn disease trio and try to utilize this data together with genetical genomics analyses by RNAseq as a basis for personalized understanding of genetic variants that cause human disease. To assess the genetic disease determinants in the child's genome, especially variants that follow the recessive model of inheritance, the genomes of the child and its parents were completely sequenced using different library types. It is self-evident, that robust identification of genetic variants is dependent on decent genomic coverage, as sequencing technologies might introduce errors during the necessary amplification steps. Also it is important to show that the variant calling algorithms applied are reliable, i.e. by comparison with variant calling results of a different technology. Therefore, the sequencing and variant calling performance will be described first, before the detected genetic variations in the child are discussed.

### **5.1. Sequencing and variant calling performance**

Multiple libraries were sequenced per individual to achieve a deep and diverse genomic coverage. Further, applying different protocols for library preparation has an advantage over sequencing the same library repeatedly, as each protocol has its own weaknesses and strengths. Long mate-pair libraries for example can be used to link distant genomic regions, which is required for the detection of large structural variants, but require a large amount of input DNA, whereas paired-end data links more closely located genomic regions and only requires low amounts of DNA. The study demonstrates that most of the investigated genomic regions (>99.85%) are accessible with current next-generation sequencing technologies and that SNV detection is, in general, very robust, evidenced by the high concordance rate (>99%) between the SNV-Calls and state-of the art comparative technology (Illumina SNV chip). However, screening for potential *de novo* variants enriches false-positive results. In cases of low coverage or too low quality reads a SNV might not be detected in one of the parents or might show an incorrectly called zygosity, which ultimately leads to false positive *de novo* variants. As shown, it is possible to reduce the number of likely false-positive *de novo* variants dramatically by re-investigating the mapped sequences (independent of the SNV-Call). Two of the highest scoring filtered potential *de novo* variants could be verified, representing 2 of 20 (10%) of the expected *de novo* variants per generation. Based on the genetic variation of the family trio the geographical origin was determined as Central European, which is

required knowledge for further analysis as the effect of CD risk variants can differ between ethnicities. Also, this allows choosing of appropriate control populations, used in several of the applied analyses. While analogous technologies exist for SNVs detection, quality assessment of the larger structural variants, which easily span several kbs and have no precisely defined start and end positions, proves to be more difficult, as technologies for identification of very large numbers of SVs in a manageable way have not been developed. Hence for this type of variation it is highly important to verify candidates by conventional methods (i.e. PCR-amplification).

### **5.2. Investigation of *de novo* variants**

Variants that appear only in the child (*de novo* variants) could provide a reasonable explanation why the child is diseased while the parents do not show symptoms of CD. Two *de novo* SNVs could be verified, but present non-coding variants. One variant was detected in the UTR of *NF1* (neurofibromin) and one variant intragenic between *SYT4* (synaptotagmin) and *SETBP1* (SET-binding protein 1). It is very difficult to assess their impact on pathogenesis, as they do not influence protein function directly, but might mediate a change of expression. As this is very speculative, these SNVs are not considered as major drivers for CD pathogenesis in the presented case study. None of (the weakly scored) coding *de novo* variants could be verified. Assuming that no *de novo* SNVs were missed by the approach applied, *de novo* variants can be ruled out as causative variants for this case of CD. However, absolute certainty does not exist.

### **5.3. Genetic variants concerning monogenic phenocopies of Crohn's disease**

Multiple diseases are known to cause a Crohn's disease like phenotype but are based on a very limited number of genes as would be expected for Mendelian diseases. To exclude that the child has one of these phenocopies, the genes underlying these diseases were investigated. No disease associated SNVs could be identified, except for a SNV in *NOD2* that is however not associated with Blau Syndrome but confers a higher CD risk. Some uncertainty remains for the missense SNV (rs34562254) detected in *TNFRSF13B* that affects a conserved nucleotide and that was predicted discordantly (damaging/tolerated) by the applied prediction tools. Based on these findings, none of the investigated monogenic syndromes is likely responsible for the child's phenotype.

### **5.4. Crohn's disease associated risk variants**

As several Crohn's disease risk variants are known, another tempting hypothesis is that the child carries an extensive burden of risk variants which together would explain the disease phenotype. Taking this hypothesis into account it must be postulated that the cumulative risk is higher in the

child than in his parents or other healthy individuals. However, despite a high number of CD risk variants in all three individuals, the combined relative risk for known CD risk variants in the child (and also the parents) does not show differences to the combined relative risk of putative healthy individuals. Additionally, I demonstrated that the child did not receive more CD risk variants than would be expected for an 'average child' given the parents' genotypes. Apparently, the known risk loci do not suffice to explain the disease outbreak in the child, as the observed genetic burden for CD is similar to that of putative healthy individuals. This is not surprising, as the known CD risk conferring loci only attribute to about 20% of the genetic heritability of CD. Therefore the genes associated with variants conferring a higher CD-risk were screened for additional, novel variants. Among these variants, a nonsense SNV was detected in the fucosyltransferase gene (*FUT2*) that was verified using Sanger sequencing. The SNV is shared between the father and the son, and thus cannot be a major driver of the disease under the postulated recessive model. Another SNV was identified in the "macrophage stimulating 1" gene (*MST1*). At first, the SNV appeared as potential *de novo* variant, however Sanger sequencing revealed that the variant is heterozygous in both, mother and child while the father is homozygous for the reference allele. It could be speculated that the possibly damaging *FUT2* and *MST1* SNVs might exhibit a synergistic effect on CD pathogenesis.

The SNV (rs11549656, P200L) detected in a conserved region of *GPX1* is located very closely to a similar variant (P197L) that has been reported as polymorphism by Forsberg *et al.* (1999).<sup>106</sup> Proline is often found in turn motifs in proteins and has the ability to interrupt  $\alpha$ -helices and  $\beta$ -strands, thus replacement of prolines could lead to a conformational and functional change. The P200L variant is listed as probable pathogenic in dbSNP; though on a closer look no clinical link to pathogenesis exists. Also, Forsberg *et al.* did not link the P197L variant to pathogenesis. A strong link of this variant with CD is therefore rather unlikely. Another proline change was detected in the glucokinase regulatory protein (*GCKR*, rs1260326) that is located between two motifs which are thought to be directly involved in binding of phosphoesters. The SNV is strongly associated with lower fasting glucose levels and fasting insulin levels and higher triglyceride levels that reduces type II diabetes risk in the French population.<sup>97</sup> The protective link of the SNV with diabetes is interesting, but it is not likely that the SNV has an opposite effect in CD. SNVs following the recessive model represent targets of high interest under the assumption of a strong inherited genetic background of the disease. Only one such SNV could be identified in genes associated with CD risk loci. However, the SNV identified in *SNAPC4* is located in a site of high acceleration, which, although the SNV was predicted to be damaging by SIFT (with low confidence), disagrees with an important function of the residue. Also, *SNAPC4* is located in the same genetic risk region identified

by GWAS as *CARD15*, which is a much stronger candidate for Crohn's disease. Two small SVs were detected in CD risk loci associated genes (*TNFSF18*, *DENND1B*), but the variants are shared with the mother and father respectively. Summarized, no strong candidate variants were identified in genes associated with CD risk loci that are sufficient to explain the disease outbreak in the child. However, distinct inheritance of several of these variants could overcome an unknown threshold to develop the disease. As shown, the sSV in *TNFSF18* and the *MST1* missense SNV were (likely) inherited by the mother, while the *FUT2* nonsense variant and *DENND1B* variants are shared between the father and child.

### **5.5. Regions of homozygosity in relation to known Crohn's disease risk loci**

Regions (or runs) of homozygosity (ROHs) are long stretches without heterozygosity in the diploid state. Multiple studies have identified ROHs associated with other complex traits, such as schizophrenia and late onset of Alzheimer's disease and there is evidence that ROHs can be used to uncover hidden recessive variants in complex disease. This has been proven to be particularly useful in investigating autosomal recessive disorders in populations with a high prevalence of consanguinity. Additionally, larger differences in ROHs could be shown between cases and controls in various studies.<sup>98</sup> Although many ROHs were found to overlap with known CD risk loci, no over-representation of known CD risk loci in ROHs could be identified in the child's genome compared to healthy individuals. Therefore, it is unlikely that autozygosity in CD risk regions is responsible for CD in the presented case.

### **5.6. Genetic variants associated with other inflammatory diseases**

Many risk genes associated with inflammatory traits are shared between multiple diseases. It is likely that some genetic variants in genes involved in other inflammatory diseases could contribute to the child's phenotype. Risk loci and associated genes (based on the GWAS catalog) were investigated for the following diseases: ankylosing spondylitis, asthma, celiac disease, inflammatory bowel disease (IBD, which is a distinct entry to ulcerative colitis and CD in the GWAS catalog), leprosy, multiple sclerosis (MS), psoriasis, rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), type 1 diabetes and ulcerative colitis (UC). A larger number of missense SNVs in the HLA genes were predicted to be damaging, however the MHC regions are highly polymorphic which naturally leads to a large number of SNVs. It is difficult to assess if any of these variants have actually damaging effects. Although many SNVs were predicted to be damaging and occur at conserved positions, the alternative allele is the most common allele in the central European population for the variants in *CIITA*, *ERBB3* and *FRMD4B*, while the minor allele frequency

of the remaining SNVs is not very low (>0.2). The allele frequencies of some SNVs remain unknown with only six of them in sites of conservation. Among these are the *MST1* and *NOD2* SNVs (discussed above) and two SNVs identified in genes of regions associated with celiac disease (*KIAA1109*, *FRMD4B*). Information of the gene function for both genes is very limited, thus it remains unknown what (if any) role they could have in inflammatory diseases. The remaining two SNVs with unknown allele frequencies were identified in the gene for *interferon regulatory factor 5* (*IRF5*) and a phenol sulfotransferase gene (*SULT1A2*). The *IRF5* SNV is not included in dbSNP130, is located in a site of conservation and was predicted to be damaging by both applied prediction tools. *IRF5* has been associated with UC, RA and SLE and especially the association to UC makes this an interesting candidate gene for CD. However, this variant is not unique to the child and thus other factors are likely more important. The remaining SNV (*SULT1A2*) is located in a region associated with early onset of IBD, however other genes such as *IL27* and *EIF3C* are also located in this region and no direct link between *SULT1A2* and IBD has been made so far. Three short SVs were identified in exons or splice-junctions of *DENN1B*, *TNFSF18* (both discussed above) and *CSMD1*. *CSMD1* (*cub and sushi multiple domains 1*) is associated with MS and the 1-basepair deletion was predicted to cause nonsense mediated decay. *CSMD1* has been reported to inhibit classical complement pathway activation and might thus play a role in inflammation.<sup>107</sup> Only one large SV was identified to affect the coding region of a gene associated with the investigated inflammatory diseases. The affected gene is coding for the growth hormone receptor (*GHR*) that has been identified in a region associated with SLE. However, no direct link between the gene and the disease has been made yet and the deletion of exon 3 has been previously detected and described to increase the responsiveness to growth hormone.

This analysis revealed additional variants that could be involved in CD pathogenesis that are yet not clearly linked to CD etiopathogenesis. However, a more profound analysis of the variants (i.e. genetically modified cell and mouse models) or the verification of the variants in a larger case cohort are required. It is important to emphasize that all of the above variants are not unique to the child or follow the recessive model of inheritance, and therefore fail to explain a Mendelian cause of the child's disease.

### **5.7. Everything else - the remaining genetic variability**

Genetic regions with a higher density of variations in the family trio (or just the child), compared to putative healthy individuals could hint towards potential risk factors. However, in a heatmap no substantial differences could be observed in the SNV density across the whole genome, except for

regions close to the centromeres and some repeat regions, which is probably a result of mismapping.

Neither known CD risk variants nor other variants associated with inflammation suffice to explain the child's phenotype. Therefore, the analysis was extended to all genes that carry potential damaging genetic variants that 1) contribute to compound heterozygosity, 2) follow the recessive model of inheritance and/or 3) are located in sites of high conservation. Among the SNVs following the recessive model, only 15 were identified in sites of conservation. Ten of them show minor allele frequencies  $>0.2$  (*BRWD1*, *C2orf73*, *GBE1*, *GOLGA4*, *LOXL4*, *NOM1*, *OBSL1*, both SNVs in *PCDHA1* and *PSDM9*). The remaining five were detected in *PCDHB7*, *MPHOSPH8*, *KRT76*, *KCNJ12* and *BEST2*. Besides *MPHOSPH8* and *BEST2* all SNVs are included in dbSNP130. The SNV in *KRT76* was not further investigated as in a newer build of dbSNP (134) the minor allele frequency is reported to be  $>0.2$ . *KCNJ12* belongs to the *potassium channel, inwardly rectifying subfamily J*. The gene's function remains in great parts unknown. *PCDHB7* is encoding for *proto-cadherin-Beta 7* that might be involved in cell-cell adhesion, and thus could influence the structural integrity of the epithelial. The SNV in *MPHOSPH8* (M-phase phosphoprotein 8) has been included in the newer dbSNP build 134 (rs75390100) with a minor allele frequency of  $T=0.042$  (CEU). Expression of *MPHOSPH8* was similar in the blood samples of all three individuals with the child being lowest. Both applied SNV prediction tools designated this SNV as damaging. Data about *MPHOSPH8* is scarce, but it has been described to target the E-cadherin promoter and to recognize methyl-H3K9 marks, thus directing DNA methylation to repress tumor suppressor gene expression and, in turn, has an important function in epithelial-to-mesenchymal transition and metastasis.<sup>108</sup>

The SNV in *BEST2* has been included in the dbSNP build 134 (rs79300835) and shows a minor allele frequency of 0.042. The SNV was verified using Sanger sequencing and the occurrence in the 1000 Genomes data set (1092 samples, Oct. 2011 release) was investigated. Screening of the 1000 Genome project's individuals revealed only 84 heterozygous and four homozygous individuals for the variant, suggesting a very low allele frequency across investigated human ethnicities. PolyPhen predicted this SNV to be possibly damaging, which is in agreement with the low prevalence of homozygous individuals. The SNV is located in the first RFP (arg-phe-pro)-transmembrane domain (Pfam, PF01062.15) and the first 13 amino acids are highly conserved including the arginine at position 8 down to the teleost fishes. The SNV changes the arginine to a glutamine (R8Q). Arginine contains a positively charged side chain, while glutamine is uncharged, which could alter the properties or stability of the first transmembrane domain. Bestrophin-2 is localized at the basolateral membrane of mucin-secreting colonic goblet cells and described to mediate bicarbonate transport in the mouse colon.<sup>109</sup> Indeed considerable *BEST2* transcript levels could only

be detected in the RNAseq data of the colon-biopsy of the child, but not in any of the blood samples. Interestingly, Yu *et al.* could demonstrate that *Best2*<sup>-/-</sup> mice developed spontaneous intestinal inflammation and were highly sensitive to DSS-induced epithelial lesions and colitis when comparing them to wild-type mice. However, genotyping the SNV in large cohorts of Crohn's disease and ulcerative colitis patients did not confirm the potential association of this SNV with disease risk. Although a higher abundance of the SNV could not be demonstrated in our disease cohort compared to the controls, it is possible that larger studies or studies with a different ethnic background might find an association, but a high impact of *BEST2* in CD is rather unlikely.

Several genes show SNVs that possibly contribute to compound heterozygosity, whereas only few of these SNVs include at least one SNV that was predicted to be damaging or were not predicted at all. The *HLA-A* gene and three mucin genes (*MUC2*, *MUC4* and *MUC16*) carry at least one SNV predicted to be damaging from either mother or father and additional SNVs from the other parent. The possible involvement of these genes, influencing the mucus layer and antigen presentation, has been discussed in the introduction. For the remaining genes with compound heterozygosity only few have known functions. For example, decreased expression levels of *NLRP1* (*NLR family, pyrin domain-containing 7*) have been associated with Crohn's disease.<sup>110</sup> Faustin *et al.* suggested that *NLRP1* is a direct sensor of bacterial components in host defense against pathogens. Muramyl-dipeptide (MDP) leads to a 2-step oligomerization of *NLRP1* involving ribonucleotide triphosphates, which then activates *CASP1*.<sup>111</sup> Another variant was detected in *CCHCR1* (*coiled-coil alpha-helical rod protein 1*), that has been associated with abnormal keratinocyte proliferation, which is a key feature of psoriatic epidermis.<sup>112</sup> *SLC19A1* presents another gene with compound heterozygous SNVs. Chapkin *et al.* showed that mice compound heterozygous for the folate transporter *SLC19A1/RFC1* and folate-binding protein *FOLR1/FBP* (*FBP*(+/-) *RFC1*(+/-) mice) showed a higher number of aberrant crypt foci per centimeter colon compared to *FBP*(+/-) mice and controls.<sup>113</sup> However, the results concerning compound heterozygous variants have to be regarded with care as usually only one of the SNVs contributing to compound heterozygosity was predicted to be damaging. The remaining SNV(s) might not add any additional effect.

Three large deletions were identified that probably follow the recessive model of inheritance, which affect the exons of *SPINK5L2/SPINK14* (Kazal type serine protease inhibitor 5-like 2), *GUCY2GP* and *CCDC135*. Variants in the *SPINK5L2* related gene *SPINK5* (*LEKTI*) have been associated with Netherton syndrome, asthma and atopic dermatitis, thus representing an interesting target for CD pathogenesis. The identified deletion in *SPINK5L2* leads to the loss of exon three in the child's genome, while the parents both still have the respective region, at least in a hemizygous state. The data is supported by all sequenced libraries, which is a strong proof for the existence of this variant.

However, no expression of *SPINK5L2* was detected in the blood samples and colon biopsy of the child. It remains unclear how (or if) this variant is involved in CD pathogenesis. The remaining recessive deletion affects *GUCY2GP* that is likely a pseudogene and thus of little interest.

Both models, the recessive and compound heterozygous inheritance, do not yield sufficient evidence to explain the presented CD case. Therefore, additional variants were screened in exonic sites of conservation, as these locations likely have functional (or regulatory) consequences. Multiple genes were identified that are likely involved in immune system processes. Among the most interesting SNVs is one in *FURIN*, a gene that is involved in the function of regulatory T-cells. The SNV is not included in dbSNP and was predicted to be damaging by both prediction tools. *FURIN* deficient regulatory T-cells have been shown to be less protective in a transfer colitis model.<sup>114</sup> *FURIN* has also been shown to cleave pro-TGFB1 to produce biological active TGFB1.<sup>115</sup> Two SNVs have been identified in the toll-like receptor gene *TLR10*. One of these two SNVs shows a low minor allele frequency (0.025) in the central European population. The child was found homozygous for this variant and additionally both SNVs were predicted to be damaging by both applied prediction tools. Other interesting candidates include variants in *NLRP12* (inflammasome), *IL22RA2* (*interleukin-22 receptor, alpha 2*) that might be an antagonist of IL-22 in the regulation of inflammatory responses,<sup>116</sup> *SIGLEC1* (sialoadhesin) that is expressed on macrophages and facilitates antigen recognition of B-cells<sup>117</sup>, *FREM1/TILRR* (*FRAS1-related extracellular matrix protein 1*), *ITGAX* and *ITGA3* (integrins) and *MMP9* (*matrix-metalloproteinase 9*). *ITGAX* is a leukocyte surface molecule and has been associated with systemic lupus erythematosus, whereas *ITGA3* has been shown to promote *MMP9* production in human gastric carcinoma cell lines,<sup>118</sup> which in turn has been found to be upregulated in IBD and a mediator of tissue damage in colitis.<sup>119</sup> *TILRR* is a product of alternative splicing of the *FREM1* mRNA and is a co-receptor of IL-1RI. *TILRR* might regulate Toll-like receptor/interleukin 1 receptor signal transduction.<sup>120</sup>

All of the above mentioned variants could play a role in CD pathogenesis, but they cannot explain a strict Mendelian cause of the disease, as they are shared with the parents. Still, it might be worth pursuing these variants in cell/mouse models and diseased patients. As no evidence for a Mendelian disease-like genotype could be identified, it is likely that a large number of variants with smaller effects are responsible for CD in the presented case. A functional protein network based on all missense SNVs filtered by the gene ontology term "immune system process" was generated with STRING. While many genes remained without connection, several clusters could be identified including: antigen recognition and processing (HLA genes, MHC class I and II), proteasome function (*PSMB9*), innate immune responses to virus infections (2-prime,5-prime oligoadenylate synthetases (OAS) genes), DNA repair (*XRCC5*, *MSH3*, *MSH6*, *MLH1*, *EXO1*), chemokines (*CCL22*, *CCL23*, *CX3CR1*,



*CX3CL1*), complement system (*C5AR1*, *C5*, *C7*, *C8A*; *CFI*, *CFHR1*, *CFH*; *CR1*, *CR2*), NADPH oxidase complex (*CYBA*, *NCF1*, *NCF2*) and pattern recognition and signaling associated pathways (*NOD2*, *TLR3*, *NLRX1*, *IL1RL1*, *TNFRSF11A*, *TNFRSF13B*, *IRAK1*, *IRAK2*, *IRAK3*, *IKBKAP*, *CHUK*). Only few clusters include genes with putative recessive SNVs (*HLA-DPB1* (potential damaging SNV), *HLA-C*, *HDAC5*, *TLR6*, *TLR10*, *TNFRSF1B*, *POLL* (potential damaging SNV), *CYBA* (potential damaging SNV) and *NLRX1*. The number of non-recessive SNVs predicted to be damaging is much higher and these SNVs affect all above mentioned clusters, except *PSMB9* and the OAS genes. These findings highlight the complexity of the genetic findings. While any of these variants could contribute to CD pathogenesis, it is impossible to identify the true-positive associations in this study with very limited number of individuals (N=3) without further functional studies and/or studies using larger cohorts.

### **5.8. Differential expression and genetic variants in differentially expressed genes**

The transcriptomes of all three family members were investigated to identify differentially expressed genes. Expression levels were determined with the widely used Cufflinks software<sup>89</sup> that are reported as fragments per kilobase of exon model per million mapped fragments (FPKM). 75 genes were identified as 5-fold upregulated and 12 genes as 5-fold downregulated in the child compared to its parents. Among these are many genes involved in inflammation, i.e. several genes are associated with immune cell function and development (*CD177*, *C19orf59*, *ARG1*, *CST7*, *TNFSF13B*, *IL4*), and with pattern recognition and associated downstream signaling (*CLEC4D*, *HMGB2*, *TLR5*, *NLRC4*, *IRAK3*, *CR1*). Several of the identified genes have been shown to be differentially expressed in CD patients. Increased levels of *S100A12* (*S100 calcium binding protein A12*) in serum were found in patients with IBD compared to patients with irritable bowel syndrome (IBS) and can be used to discriminate both groups.<sup>121</sup> *S100A12* is a ligand of the multi-ligand binding receptor for advanced glycation end products (RAGE) that has been associated with a larger number of inflammatory and autoimmune diseases, including type 1 diabetes and rheumatoid arthritis. Binding of *S100A12* to RAGE has been shown to trigger cellular activation, with generation of key proinflammatory mediators.<sup>122</sup> Interestingly, Hoffman *et al.* also reported that blockade of *S100A12*/RAGE interaction i.e. by administration of soluble RAGE (sRAGE), that functions as a decoy molecule for the ligands efficiently, prevents inflammation in an IL-10 null mice colitis model. Additionally, the expression levels of *S100A8* and *S100A9* are increased in the child compared to the parents. Both molecules are ligands of RAGE that have been reported to induce pro-inflammatory responses associated with an increased vascular permeability due to a loss of endothelial cell-cell contacts and cell junction proteins.<sup>123</sup> Higher concentrations of *MMP9* (*matrix-metalloproteinase 9*) have been reported to be significantly increased in the sputum of CD patients compared to

controls,<sup>124</sup> which is in agreement with high MMP9 expression levels detected in the child. Arijis *et al.* tried to identify genes corresponding to a response to Infliximab (IFX) in CD patients. They identified 697 significant probe sets between Crohn's disease colitis responders and non-responders. The top five genes were *TNFAIP6*, *S100A8*, *IL11*, *GOS2* and *S100A9* that were sufficient to completely separate both groups.<sup>125</sup> Expression levels of all five genes were higher in non-responders than in responders. Three of these genes, *TNFAIP6*, *S100A8* and *S100A9* were found upregulated in the child, compared to its parents. The child was, among other things, treated with IFX and repeat colonoscopy after treatment still showed active disease after maximum therapy. Whether the higher expression of *TNFAIP6*, *S100A8* and *S100A9* in the child indeed counteracted IFX response remains speculative, but poses an interesting finding. Another interesting result is the higher expression of *TLR5* in the child, as a nonsense mutation in the TLR5 receptor, that most likely recognizes bacterial flagellin<sup>126</sup>, has been reported to protect persons of Jewish ethnicity against CD.<sup>127</sup> If a loss of TLR5 protects against CD, an increased expression of TLR5 might have opposite effects. Several of the higher expressed genes are involved in pathogen recognition and clearance. One example is the above mentioned *TLR5*, but also *CR1* (*complement factor 1*), *HMGB2* (*high mobility group box 2*), *IRAK3* (*interleukin receptor-associated kinase 3*) *BGR/CLEC7A/DECTIN1* (*C-type lectin domain family 7, member A*) and *PGLYRP1* (*peptidoglycan recognition protein 1*). *CR1* is present on red blood cells and binds to circulating complement-opsonized particles that are then transferred to macrophages in the liver and spleen for clearance.<sup>128</sup> *CLEC7A* was identified as a surface receptor of beta-glycans on macrophages, and thus functions as a pattern recognition receptor for fungi and some bacteria.<sup>129</sup> Lack of HMBGs has been reported to reduce activation of TLR3, -7 and -9, thus an increased level of HMGB2 might lead to an increased TLR response. *IRAK3* is upregulated in a cytokine dependent manner upon LPS exposure.<sup>130</sup> Upregulation of this gene raises further evidence for the high activity of the antimicrobial sensing pathways and downstream cascades in the child.

The genes detected to be differentially expressed in the child were screened for genetic variants that are predicted to be damaging. Five genes were identified, namely *CR1*, *VNN1*, *IL4R*, *EMR1* and *MMP9* that carry such variants. However, as these SNVs are not unique to the child it is unlikely that they are responsible for the expression changes. Also, non-exonic SNVs are likely more important as they could affect regulatory elements. It is however difficult to assess the effect of noncoding SNVs on the expression level, as this would require cloning or genetically modification of cell and/or animals for each variant. The coding SNVs are still noteworthy, as an increased level of potentially malfunctioning proteins might exhibit unexpected consequences in pathogenesis.

### **5.9. Strand specific transcriptome analysis**

Extending the work on the transcriptome analysis in the family example, expression of closely occurring sense/antisense (S/AS) transcript pairs was investigated in the transcriptomes derived from lymphoblastoid (LBL) cell lines of putative healthy individuals (five individuals of the 1000 Genome project),<sup>99</sup> three classical Hodgkin lymphoma (cHL) cell lines (kindly provided by Ohle Ammerpohl and Reinert Siebert)<sup>100</sup> and transcriptomes of murine small intestine and colon which were generated in duplicates<sup>101</sup>. As no tool was publically available to assess S/AS transcription on the genome-wide level using next generation sequencing technologies I developed *Janus* to fill this gap. *Janus* was applied to the transcriptomes mentioned above for detection of S/AS pairs in general and identification of differentially expressed S/AS pairs. Data for five LBL cell lines was generated using two different technologies (SOLiD whole transcriptome analysis kit (WTAK)<sup>101</sup> and Illumina GA IIx (kindly provided by Ralf Sudbrak and Hans Lehrach).<sup>102</sup> Prior identification of S/AS pairs, the performance of both technologies was compared which revealed a higher background noise in the Illumina data. This noise manifests as expression in 'exonic' regions on the strand opposite to the template strand (the "antisense" strand). An example is given in Figure 25A, where the expression of *FLNA* (second strand) generates an echo on the opposite (first) strand in the Illumina data. The exon-intron structure is preserved on the first strand, which is highly unlikely, given conserved splice-site associated motifs. The SOLiD data on the contrary shows almost no expression on the first strand. Thus the noise level was investigated in all exonic regions by calculating the mean and standard deviation of the noise across exons (Figure 25B and C). The results support the hypothesis that the Illumina library is less efficient in preserving the strand as is the SOLiD data. This gives a good explanation for the high antisense tag density in exonic regions observed by He *et al.* who used the Illumina GA, which could not be confirmed by Klevebring *et al.*, who used the SOLiD system.<sup>78,79</sup> It can thus be assumed that some of the differences between the two studies result from the obvious technical bias (background noise) of the two sequencing protocols. This underlines the importance to assess the 'noise' level, i.e. the error rate of the library preparation, before identifying antisense events. It should be added, that preliminary results using a novel protocol for strand specific Illumina libraries (ScriptSeq™ mRNA-Seq Library Preparation Kit) suggest that it is possible to generate Illumina libraries with a noise level comparable to the SOLiD data (data not shown). Previous publications mention a high abundance of antisense tags in both the promoter and terminator regions of genes.<sup>78,103,104</sup> Therefore, the occurrence of antisense transcription in promoter and terminator regions, here defined as 1 kb before and after the transcript start and end site, was investigated. Using cutoff criteria of a minimum length of 150 bp and a mean coverage per base of five for the antisense transcripts, identified antisense expression

in 740 promoter and 980 terminator regions in the cHL (hg18 mapped data, 32.110 annotated transcripts) and 2100 promoter and 2400 terminator regions in the LBL data set (hg19 mapped data, 66.065 annotated transcripts) (Table 10). Interestingly, when considering antisense events with low coverage in the promoter and terminator regions, the amount of antisense tags in promoter regions is higher than in terminator regions, while this shifts towards more antisense tags in terminator regions when considering only higher covered antisense events. This could be caused by a higher mismapping in promoter regions, maybe due to conserved motifs of transcription factor binding sites and TATA-like sequences, which generates a higher noise level. Another explanation is that there is more low level antisense transcription occurring in promoter regions and stronger antisense transcription in terminator regions. Also, it is possible that S/AS pairs in head-to-head conformation show lower expression than S/AS pairs in tail-to-tail conformation.

Among the huge amount of detected S/AS pairs, multiple pairs could be detected which show mutually exclusive expression of one transcript in some samples, and mutually exclusive expression of the other transcript in others. Examples with high confidence for this type of S/AS pair could be identified in all three data sets. This demonstrates that differential S/AS expression does not only occur between different tissues, but also between individuals.

Besides differentially expressed S/AS pairs I screened for S/AS pairs which show similar expression for both transcripts but differ in expression per sample. It could be argued this is caused by different library performance (i.e. this region was just unlucky in some samples) or a coverage bias might persist despite normalizing for the total coverage per sample. However, as shown in the example for the LBL data set (Figure 28), *GLB1* shows a similar expression rate in all samples while *TRIM71* and the unknown antisense event show a positively correlated expression, which differs between samples. It is unlikely that the observed co-expression is due to coverage effects. Katayama *et al.* suggested that long noncoding RNAs might be involved in S/AS pairs, reasoned by a higher number of S/AS pairs detected by random primed *cap analysis gene expression* (CAGE) compared to oligo-dT primed CAGE.<sup>103</sup> It could be speculated, that the identified AS event upstream of *TRIM71* might present such a noncoding RNA, as no exon-intron structure is visible. Interestingly, in the same paper Katayama *et al.* reported a frequent concordant regulation of S/AS pairs. Another study by Watanabe *et al.* raises further evidence of a high occurrence of positively correlated S/AS pairs.<sup>131</sup> The reason for this remains unclear, but it could be due to the fact, that transcription of one strand makes the other strand more accessible as well and palindromic sequences in regulatory regions, such as promoter regions might be used by both strands. In this case the observed transcript could be nonsense transcription, which ends by the first possible

encountered transcription end signal. This might explain why antisense transcripts in promoter regions are often described as short and long noncoding RNAs. However, it is not unlikely that these antisense transcripts have some function, as they do not occur in all promoter or terminator regions of every expressed gene. Multiple studies showed that noncoding RNAs might alter the chromatin conformation or lead to a local change of methylation if they are located within promoter regions.<sup>132-134</sup> The occurrence of co-expressed S/AS pairs was also shown for the cHL data set and in the different murine tissues.

Using the data available to us, no link between sense and antisense transcription and methylation of CpG islands could be made, except weakly for one case, which might be coincidental (Supplementary figure 9, Supplementary table 43). It is possible that the limited number of CpG islands interrogated by the chip (27k) prevented identification of CpG methylation conditional antisense expression. A larger data set or data based on histone modification might lead to new results, as a link to histone methylation has been shown before, i.e. in the case of the p21 antisense transcript.<sup>77</sup>

By investigating known SNV positions (generated by the 1000 Genomes Consortium),<sup>99</sup> some SNVs were identified in the LBL data set that show signs of strand specific allelic imbalance. While this is a very interesting finding, it remains unknown if this poses a mechanism to avoid potential inhibitory effects on transcription occurring on both strands simultaneously. Janus also identifies S/AS specific allelic imbalance for positions that are not known to be SNV positions. Yet, most of these positions can be traced back to regions that are difficult to map to (i.e. HLA), repeat regions, pseudogenes and genes with multiple copies, thus these results have to be considered with care. Janus proved to be very efficient in the identification of regions transcribed in antisense to other transcribed regions. Taken together, detection of differentially expressed and co-expressed S/AS pairs proved to be difficult but manageable.

## **5.10. Conclusions**

A large number of genetic variants have been identified in genes of the child that may potentially influence immune responses. However, no single variant could be identified that argues for a Mendelian cause of the child's phenotype. Additionally, examination of several monogenic diseases with a CD-like phenotype did not result in probably disease associated genetic variants and no *de novo* variants were identified in coding regions of the child. However, it cannot be ruled out that some variants were not detected by the SNV calling algorithms. Screening for genetic variants that have been previously associated to CD (GWAS studies) showed that many of these variants are present in the child. Yet, it could be shown that the family trio does not carry more of

these risk loci than other putative healthy individuals and that the child did not receive an unusual high inherited burden of CD risk loci than would be expected by random and independent inheritance of each locus. Further, no overrepresentation of CD risk loci in regions of homozygosity could be shown compared to putative healthy individuals. Therefore, the known CD risk loci were excluded as causative on their own, as more genetic variants must be involved. Investigation of genes involved in other inflammatory diseases revealed additional putative damaging genetic variants in potential CD relevant genes. However, no single variant could be identified that is likely causative for the phenotype. Investigated of all recessive, compound heterozygous and SNVs in phylogenetic conserved locations identified one recessive SNV in *BEST2* that was confirmed by Sanger sequencing. Although this SNV is associated with bicarbonate transport by goblet cells and thus represents a very interesting candidate for CD, no higher prevalence of the variant could be detected in CD or UC patients (N=1066) compared to controls (N=1468). Despite this, the variant is rare and only few homozygous individuals (N=10) were identified. Therefore this variant might underlie a higher selection pressure, suggesting an important impact on gene function. Although the variant alone is unlikely to cause a CD phenotype itself, it might add to a predisposition towards CD.

Investigation of large deletions did not yield a reliable recessive variant that effects exons and might be associated to CD. It must be kept in mind, that not all genetic variants might have been detected in all three individuals or that the zygosity of the variants was not assigned correctly. Assuming a no major causative variant was missed, the results indicate that the child's phenotype is, despite a strong genetic bias, probably not caused by a single genetic variant, but by a combination of several. It is enticing to speculate the variants in *FUT2* and *MST1* (and probably some more) that are shared with either parent might contribute to the pathogenesis by synergistic effects in the child that overcome an unknown threshold necessary to develop a pathogenic phenotype. With the current state of knowledge of CD pathogenesis, general gene function and available sequencing technologies it was not possible to explain the child's phenotype by a single causative variant. Despite this, this thesis provides a blue-print for the identification of potential rare risk variants in an otherwise overwhelming amount of genetic variants in a case of complex disease. Indeed, the applied approach was successful in the identification of a missense variant in the *BEST2* gene that was as a very good candidate for a Mendelian disease gene that upon validation proved to be rare, but also comparably common in the normal population.

Also in the context of next generation sequencing, in this thesis I developed *Janus* as a tool to investigate the nature of antisense transcription on the genome-wide level. As shown, the tool is very efficient in the identification of sense/antisense transcript pairs, is capable to detect S/AS pairs

that show mutually exclusive expression of one transcript in some samples and expression of the other in others, as well as S/AS pairs that are positively correlated. This tool might prove useful for the understanding S/AS expression in connection with disease phenotypes.

### **5.11. Perspective**

Identification of causative variants in complex diseases proved to be very complicated. A lack of functional information of many genes, the inaccessibility of some genetic regions by NGS and basically no prediction tools for non-coding variants limit the use of whole-genome sequencing in complex diseases. Future improvements of the sequencing technology (i.e. by higher read lengths) and genetic variant calling algorithms will overcome some of these problems, as they will allow the detection of variants in regions that are yet difficult to assess (i.e. the highly polymorphic genes of the MHC region). As the costs for genome sequencing with a decent coverage steadily drops, it becomes more feasible to sequence bigger families or ethnic groups that makes identification of common causative variants easier, which could also pinpoint to reoccurring non-coding variants in disease cohorts. Very soon the first studies will be published that are performing a GWAS-like analysis on exome data. This approach has a much better resolution than GWAS studies and will lead to the discovery of new disease genes. However, this is still not perfect, as exome sequencing will miss genetic variants in regulatory genetic elements and cannot assess epigenetic modifications. Once the costs for sequencing dropped enough, whole-genomes and epigenomes will take over and overcome these issues. The identification of genome-wide disease associated genetic variants at single nucleotide resolution will drastically increase our knowledge of the genetic basis of many diseases and might result in new therapeutic targets.

## **6. Summary (English)**

In this study, the genome and transcriptome of a family trio with a severe case of CD with very early onset was analyzed by whole-genome and transcriptome sequencing. The family history and the early onset (1.5 years), suggest a very strong genetic background of the disease. Genes involved in monogenic diseases presenting with a CD-like phenotype were investigated and excluded as no genetic variants linked to disease could be identified. Further, investigation of potential *de novo* single nucleotide variants (SNVs) did not result in any verified coding SNV. The occurrence of known CD-risk associated variants was investigated, but no higher occurrence of CD-risk associated was detected in the family trio compared to putative healthy individuals. Also, the child does not carry more of the known CD-risk loci than the parents. Several genetic variants were identified in the genes associated to CD risk loci (i.e. a nonsense SNV in *FUT2* and a missense SNV in *MST1*),

which proved to be shared with either mother or father. No sufficient evidence was detected that known CD risk variants and risk genes can explain the child's phenotype. Multiple genes involved in other inflammatory diseases were found to carry genetic variants, but none of them are likely to explain the CD phenotype in a Mendelian model. Therefore all genes were screened for variants that follow the recessive model of inheritance, are compound heterozygous or located in phylogenetically conserved locations. By this approach a recessive variant was detected in *BEST2*, a gene associated with bicarbonate transport by goblet cells in the mouse model, which proved to be rare, especially in the homozygous state, but not more uncommon than in healthy controls. In agreement with its suggested function, *BEST2* expression was not detectable in the transcriptomes derived from blood, but only in the transcriptome from a colon biopsy of the child. Additionally, a large deletion following the recessive model was detected, that affects exon three of *SPINK5L2*. However, no expression of *SPINK5L2* was observed in any of the transcriptomes. Although no single variant was identified that is likely causative for the child's phenotype, this thesis presents a blueprint for the identification of potentially disease associated genetic variants in a very limited number of genomes.

As the second main result of the thesis a tool, *Janus*, was developed that is capable of identifying transcriptional active regions (TARs) in antisense to other TARs. To the best of my knowledge, no tool to investigate sense and antisense expression on the genome-wide level using next generation sequencing data existed prior to this work. *Janus* shows that antisense transcription is a very common phenomenon and that, with *Janus*, sense/antisense pairs with differential expression affecting only one or both strands equally can be identified. *Janus* might be a useful tool to identify differences in antisense expression in case-control studies.

## 7. Summary (German)

Während dieser Arbeit wurde das gesamte Genom und Transkriptom einer Familie (gesunde Eltern, erkranktes Kind) sequenziert, in der ein schwerer Fall mit frühem Ausbruch (1,5 Jahre) von Morbus Crohn (Crohn's disease, CD) auftritt. Der familiäre Hintergrund und der frühe Ausbruch suggerieren einen starken genetischen Hintergrund der Erkrankung. In dieser Arbeit habe ich die Gene monogentischer Erkrankungen mit CD-ähnlichen Phänotypen untersucht, jedoch konnten keine krankheitsassoziierten genetische Varianten in den assoziierten Genen identifiziert werden, so dass diese Erkrankungen ausgeschlossen wurden. Potentielle *de novo* Einzel-Nukleotid-Varianten (*single nucleotide variants*, SNVs) die in kodierenden Bereichen liegen wurden identifiziert, konnten aber in einem nachfolgendem Schritt nicht verifiziert werden, so dass diese ebenfalls als Ursache auszuschließen sind. Auch eine Analyse der bekannten CD-assoziierten genetischen Varianten



zeigte keine Anreicherung dieser Varianten in den Genomen des Trios im Vergleich zu anderen putativ gesunden Individuen. Zudem habe ich gezeigt, dass das Kind, basierend auf den bekannten CD-Risiko loci, im Vergleich zu den Eltern kein höheres CD-Risiko trägt. Einige neue genetische Varianten konnten in Genen nachgewiesen werden, die aufgrund einer oder mehrerer anderen genetischen Varianten mit CD assoziiert sind. Zu diesen Varianten zählt eine Stop-Mutation in *FUT2*, sowie eine Variante in *MST1*, die zu einem Aminosäureaustausch führt. Beide Varianten wurden nicht exklusiv im Kind gefunden, sondern werden jeweils mit einem Elternteil geteilt. Es gibt keinen Hinweis darauf, dass die detektierten, bekannten CD-Risikoveränderungen ausreichend sind um den Phänotypen des Kindes zu erklären. Eine Analyse der Gene anderer entzündlicher Erkrankungen erbrachte ebenfalls keine Variante, die in einem Mendelschen Modell zu einem CD-Phänotyp führen könnte. Daher wurde die Analyse auf alle genetischen Varianten erweitert, die dem rezessiven Vererbungsmodell folgen oder zusammengesetzt heterozygot sind, sowie Varianten in phylogenetisch konservierten Regionen. Mit diesem Ansatz konnte eine rezessive Variante in *BEST2* identifiziert werden, einem Gen dessen Produkt mit dem Transport von Bikarbonat durch Becher-Zellen assoziiert ist. Diese Variante tritt selten, bzw. homozygot sehr selten auf, konnte aber im Vergleich mit gesunden Kontrollen nicht häufiger in Patienten mit entzündlichen Darmerkrankungen nachgewiesen werden. Anhand der analysierten Transkriptome konnte gezeigt werden, dass *BEST2* zwar in der Kolonbiopsie des Kindes exprimiert wird, allerdings nicht in den drei Transkriptomen aus Blut. Außerdem konnte eine große, rezessive Deletion in *SPINK5L2* nachgewiesen werden, die ein ganzes Exon des Gens betrifft. Allerdings konnte keinerlei Expression des Gens in einem der Transkriptome festgestellt werden. Obwohl der Phänotyp des Kindes an keiner einzelnen genetischen Variante festgemacht werden kann, wird mit dieser Arbeit ein methodisches Vorgehen gezeigt, wie potentielle krankheitsassoziierte genetische Varianten in einer sehr begrenzten Anzahl von Genomen gefunden werden können.

Ein weiterer Teil meiner Arbeit war die Entwicklung eines Programms, *Janus*, das der Identifizierung von transkriptionell aktiven Regionen (TARs) in *antisense* Richtung zu anderen TARs dient. Nach bestem Wissen ist dies das erste Programm um genomweit *sense* und *antisense* Expression mit Hilfe der *Next-Generation-Sequencing*-Technologie zu erfassen. Es wurde gezeigt, dass *antisense* Transkription sehr häufig auftritt, und dass *Janus sense/antisense*-Paare identifizieren kann, die eine differentielle Expression eines oder beider DNA-Stränge zeigen. *Janus* hat das Potential zu interessanten Entdeckungen bezüglich der Veränderung von *antisense* Transkription in Krankheitsstudien beizutragen.

## 8. References

1. Avery, O. T., MacLeod, C. M. & McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. *J Exp Med* **79**, 137–158 (1944).
2. WATSON, J. D. & CRICK, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
3. Green, E. D., Guyer, M. S. & Institute, N. H. G. R. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).
4. Feller, M. *et al.* Long-Term Antibiotic Treatment for Crohn’s Disease: Systematic Review and Meta-Analysis of Placebo-Controlled Trials. *Clin Infect Dis.* **50**, 473–480 (2010).
5. Török, H.-P. *et al.* Crohn’s disease is associated with a toll-like receptor-9 polymorphism. *Gastroenterology* **127**, 365–366 (2004).
6. Benchimol, E. I. *et al.* Epidemiology of pediatric inflammatory bowel disease: a systematic review of international trends. *Inflamm. Bowel Dis.* **17**, 423–439 (2011).
7. Grieci, T. & Bütter, A. The incidence of inflammatory bowel disease in the pediatric population of Southwestern Ontario. *J. Pediatr. Surg.* **44**, 977–980 (2009).
8. Shin, D. H. *et al.* Increasing incidence of inflammatory bowel disease among young men in Korea between 2003 and 2008. *Dig. Dis. Sci.* **56**, 1154–1159 (2011).
9. Hope, B. *et al.* Rapid rise in incidence of Irish paediatric inflammatory bowel disease. *Archives of disease in childhood* (2012).doi:10.1136/archdischild-2011-300651
10. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
11. Kaser, A. & Blumberg, R. S. Autophagy, Microbial Sensing, Endoplasmic Reticulum Stress, and Epithelial Function in Inflammatory Bowel Disease. *Gastroenterology* **140**, 1738–1747.e2 (2011).
12. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
13. Adeyanju, O. *et al.* Common NOD2 risk variants in African Americans with Crohn’s disease are due exclusively to recent Caucasian admixture. *Inflammatory Bowel Diseases* doi:10.1002/ibd.22944
14. Taurog, J. D. *et al.* The germfree state prevents development of gut and joint inflammatory disease in HLA-B27 transgenic rats. *J. Exp. Med.* **180**, 2359–2364 (1994).
15. Vaishnava, S., Behrendt, C. L., Ismail, A. S., Eckmann, L. & Hooper, L. V. Paneth cells directly sense gut commensals and maintain homeostasis at the intestinal host-microbial interface. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 20858–20863 (2008).
16. Umesaki, Y., Setoyama, H., Matsumoto, S. & Okada, Y. Expansion of alpha beta T-cell receptor-bearing intestinal intraepithelial lymphocytes after microbial colonization in germ-free mice and its independence from thymus. *Immunology* **79**, 32–37 (1993).
17. Fransen, K., Mitrovic, M., van Diemen, C. C. & Weersma, R. K. The quest for genetic risk factors for Crohn’s disease in the post-GWAS era. *Genome Med* **3**, 13 (2011).
18. Colonna, M. All roads lead to CARD9. *Nature Immunology* **8**, 554–555 (2007).
19. Yoshinaga, S. K. *et al.* Characterization of a new human B7-related protein: B7RP-1 is the ligand to the co-stimulatory protein ICOS. *Int. Immunol.* **12**, 1439–1447 (2000).
20. Okamura, H. *et al.* Cloning of a new cytokine that induces IFN-gamma production by T cells. *Nature* **378**, 88–91 (1995).
21. Mucida, D. *et al.* Reciprocal TH17 and regulatory T cell differentiation mediated by retinoic acid. *Science* **317**, 256–260 (2007).
22. Chen, C.-R., Kang, Y., Siegel, P. M. & Massagué, J. E2F4/5 and p107 as Smad cofactors linking the TGFbeta receptor to c-myc repression. *Cell* **110**, 19–32 (2002).

23. Zhang, Y., Feng, X., We, R. & Derynck, R. Receptor-associated Mad homologues synergize as effectors of the TGF-beta response. *Nature* **383**, 168–172 (1996).
24. Muise, A. M. *et al.* Polymorphisms in E-cadherin (CDH1) result in a mis-localised cytoplasmic protein that is associated with Crohn's disease. *Gut* **58**, 1121–1127 (2009).
25. Buisine, M. *et al.* Mucin gene expression in intestinal epithelial cells in Crohn's disease. *Gut* **49**, 544–551 (2001).
26. Gersemann, M., Stange, E. F. & Wehkamp, J. From intestinal stem cells to inflammatory bowel diseases. *World J Gastroenterol* **17**, 3198–3203 (2011).
27. Salzman, N. H. Paneth cell defensins and the regulation of the microbiome. *Gut Microbes* **1**, 401–406 (2010).
28. Wehkamp, J., Stange, E. F. & Fellermann, K. Defensin-immunology in inflammatory bowel disease. *Gastroenterol. Clin. Biol.* **33 Suppl 3**, S137–144 (2009).
29. French, A. T. *et al.* The expression of intelectin in sheep goblet cells and upregulation by interleukin-4. *Vet. Immunol. Immunopathol.* **120**, 41–46 (2007).
30. Rehman, A. *et al.* Nod2 is essential for temporal development of intestinal microbial communities. *Gut* **60**, 1354–1362 (2011).
31. Sanders, D. S. A. Mucosal integrity and barrier function in the pathogenesis of early lesions in Crohn's disease. *J Clin Pathol* **58**, 568–572 (2005).
32. Cantó, E. *et al.* MDP-Induced selective tolerance to TLR4 ligands: Impairment in NOD2 mutant Crohn's disease patients. *Inflammatory Bowel Diseases* **15**, 1686–1696 (2009).
33. Slade, J. D. *et al.* Inherited deficiency of second component of complement and HLA haplotype A10,B18 associated with inflammatory bowel disease. *Ann. Intern. Med.* **88**, 796–798 (1978).
34. Marks, D. J. B. *et al.* Inflammatory bowel diseases in patients with adaptive and complement immunodeficiency disorders. *Inflammatory Bowel Diseases* **16**, 1984–1992 (2010).
35. Bak-Romaniszyn, L. *et al.* Mannan-binding lectin deficiency in pediatric patients with inflammatory bowel disease. *Scand. J. Gastroenterol.* **46**, 1275–1278 (2011).
36. Janeway, C. A., Travers, P., Walport, M. & Shlomchik, M. J. *Immunobiology: The Immune System in Health and Disease*. (Garland Science: New York, 2001).
37. Villani, A.-C. *et al.* Common variants in the NLRP3 region contribute to Crohn's disease susceptibility. *Nat. Genet.* **41**, 71–76 (2009).
38. Lewis, G. J. *et al.* Genetic association between NLRP3 variants and Crohn's disease does not replicate in a large UK panel. *Inflamm. Bowel Dis.* **17**, 1387–1391 (2011).
39. Travassos, L. H. *et al.* Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nat. Immunol.* **11**, 55–62 (2010).
40. Singh, S. B., Davis, A. S., Taylor, G. A. & Deretic, V. Human IRGM induces autophagy to eliminate intracellular mycobacteria. *Science* **313**, 1438–1441 (2006).
41. Murray, R. Z., Kay, J. G., Sangermani, D. G. & Stow, J. L. A role for the phagosome in cytokine secretion. *Science* **310**, 1492–1495 (2005).
42. Alegre-Abarategui, J. *et al.* LRRK2 regulates autophagic activity and localizes to specific membrane microdomains in a novel human genomic reporter cellular model. *Hum. Mol. Genet.* **18**, 4022–4034 (2009).
43. Gardet, A. *et al.* LRRK2 is involved in the IFN-gamma response and host response to pathogens. *J. Immunol.* **185**, 5577–5585 (2010).
44. Kaser, A. *et al.* XBP1 links ER stress to intestinal inflammation and confers genetic risk for human inflammatory bowel disease. *Cell* **134**, 743–756 (2008).
45. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
46. Ebstein, F., Kloetzel, P.-M., Krüger, E. & Seifert, U. Emerging roles of immunoproteasomes beyond MHC class I antigen processing. *Cellular and Molecular Life Sciences* 1–16doi:10.1007/s00018-012-0938-0

47. Interferon gamma stimulation modulates the proteolytic activity and cleavage site preference of 20S mouse proteasomes. *J Exp Med* **179**, 901–909 (1994).
48. Fehling, H. J. *et al.* MHC class I expression in mice lacking the proteasome subunit LMP-7. *Science* **265**, 1234–1237 (1994).
49. Van Kaer, L. *et al.* Altered peptidase and viral-specific T cell response in LMP2 mutant mice. *Immunity* **1**, 533–541 (1994).
50. Kenny, E. E. *et al.* A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS Genet.* **8**, e1002559 (2012).
51. Fransen, K. *et al.* Analysis of SNPs with an Effect on Gene Expression Identifies UBE2L3 and BCL3 as Potential New Risk Genes for Crohn's Disease. *Hum. Mol. Genet.* **19**, 3482–3488 (2010).
52. Fernando, M. M. A. *et al.* Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis. *PLoS Genet* **4**, e1000024 (2008).
53. Zhang, H., Li, J., Xu, G. & Liu, Z. [Association between HLA-Cw polymorphism and inflammatory bowel disease]. *Zhonghua Nei Ke Za Zhi* **50**, 856–858 (2011).
54. Lowenthal, J. W., Zubler, R. H., Nabholz, M. & MacDonald, H. R. Similarities between interleukin-2 receptor number and affinity on activated B and T lymphocytes. *Nature* **315**, 669–672 (1985).
55. Ferlazzo, G. *et al.* The abundant NK cells in human secondary lymphoid tissues require activation to express killer cell Ig-like receptors and become cytolytic. *J. Immunol.* **172**, 1455–1462 (2004).
56. Mao, M. *et al.* T lymphocyte activation gene identification by coregulated expression on DNA microarrays. *Genomics* **83**, 989–999 (2004).
57. McGovern, D. P. B. *et al.* Fucosyltransferase 2 (FUT2) Non-Secretor Status Is Associated with Crohn's Disease. *Hum. Mol. Genet.* **19**, 3468–3476 (2010).
58. Ghoreschi, K. *et al.* Generation of pathogenic T(H)17 cells in the absence of TGF- $\beta$  signalling. *Nature* **467**, 967–971 (2010).
59. Oppmann, B. *et al.* Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12. *Immunity* **13**, 715–725 (2000).
60. Kahn, S. D. On the Future of Genomic Data. *Science* **331**, 728–729 (2011).
61. Bonetta, L. Whole-genome sequencing breaks the cost barrier. *Cell* **141**, 917–919 (2010).
62. Ashley, E. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
63. Baranzini, S. E. *et al.* Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* **464**, 1351–1356 (2010).
64. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
65. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).
66. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
67. Pelak, K. *et al.* The characterization of twenty sequenced human genomes. *PLoS Genet.* **6**, (2010).
68. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
69. Rios, J., Stein, E., Shendure, J., Hobbs, H. H. & Cohen, J. C. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum. Mol. Genet.* **19**, 4313–4318 (2010).
70. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).

71. Sobreira, N. L. M. *et al.* Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet.* **6**, e1000991 (2010).
72. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19096–19101 (2009).
73. Christodoulou, K. *et al.* Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes. *Gut* (2012).doi:10.1136/gutjnl-2011-301833
74. Hawkins, N. A. & Kearney, J. A. Confirmation of an epilepsy modifier locus on mouse chromosome 11 and candidate gene analysis by RNA-Seq. *Genes, Brain and Behavior* **11**, 452–460 (2012).
75. Wu, J. Q. *et al.* Transcriptome sequencing revealed significant alteration of cortical promoter usage and splicing in schizophrenia. *PLoS ONE* **7**, e36351 (2012).
76. Ng, K., Pullirsch, D., Leeb, M. & Wutz, A. Xist and the order of silencing. *EMBO Rep* **8**, 34–39 (2007).
77. Morris, K. V., Santoso, S., Turner, A.-M., Pastori, C. & Hawkins, P. G. Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet.* **4**, e1000258 (2008).
78. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322**, 1855–1857 (2008).
79. Klevebring, D., Bjursell, M., Emanuelsson, O. & Lundeberg, J. In-depth transcriptome analysis reveals novel TARs and prevalent antisense transcription in human cell lines. *PLoS ONE* **5**, e9762 (2010).
80. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
81. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
82. Pollard, K., Hubisz, M. & Siepel, A. Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res.* (2009).doi:10.1101/gr.097857.109
83. Notarangelo, L. D. *et al.* Primary immunodeficiencies: 2009 update. *J. Allergy Clin. Immunol.* **124**, 1161–1178 (2009).
84. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073–1081 (2009).
85. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
86. Korbel, J. O. *et al.* PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* **10**, R23 (2009).
87. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
88. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
89. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
90. Jäger, K., Islam, S., Zajac, P., Linnarsson, S. & Neuman, T. RNA-Seq Analysis Reveals Different Dynamics of Differentiation of Human Dermis- and Adipose-Derived Stromal Stem Cells. *PLoS ONE* **7**, e38833 (2012).
91. Baldwin, R. L. *et al.* Quantification of Transcriptome Responses of the Rumen Epithelium to Butyrate Infusion using RNA-seq Technology. *Gene Regul Syst Bio* **6**, 67–80 (2012).
92. Kim, J. *et al.* Transcriptome landscape of the human placenta. *BMC Genomics* **13**, 115 (2012).
93. Hackett, N. R. *et al.* RNA-Seq quantification of the human small airway epithelium transcriptome. *BMC Genomics* **13**, 82 (2012).

94. Xue, Y. *et al.* Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. *Curr Biol* **19**, 1453–1457 (2009).
95. Röhl, A., Brinkmann, B., Forster, L. & Forster, P. An annotated mtDNA database. *Int. J. Legal Med.* **115**, 29–39 (2001).
96. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
97. Vaxillaire, M. *et al.* The common P446L polymorphism in GCKR inversely modulates fasting glucose and triglyceride levels and reduces type 2 diabetes risk in the DESIR prospective general French population. *Diabetes* **57**, 2253–2257 (2008).
98. Ku, C., Naidoo, N., Teo, S. & Pawitan, Y. Regions of homozygosity and their impact on complex diseases and traits. *Human Genetics* **129**, 1–15 (2011).
99. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
100. Ammerpohl, O. *et al.* Array-based DNA methylation analysis in classical Hodgkin lymphoma reveals new insights into the mechanisms underlying silencing of B cell-specific genes. *Leukemia* **26**, 185–188 (2012).
101. Klostermeier, U. C. *et al.* A tissue-specific landscape of sense/antisense transcription in the mouse intestine. *BMC Genomics* **12**, 305 (2011).
102. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
103. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–1566 (2005).
104. Layer, J. H. & Weil, P. A. Ubiquitous antisense transcription in eukaryotes: novel regulatory mechanism or byproduct of opportunistic RNA polymerase? *F1000 Biol Rep* **1**, 33 (2009).
105. Krystal, G. W., Armstrong, B. C. & Battey, J. F. N-myc mRNA forms an RNA-RNA duplex with endogenous antisense transcripts. *Mol. Cell. Biol.* **10**, 4180–4191 (1990).
106. Forsberg, L., de Faire, U. & Morgenstern, R. Low yield of polymorphisms from EST blast searching: analysis of genes related to oxidative stress and verification of the P197L polymorphism in GPX1. *Hum. Mutat.* **13**, 294–300 (1999).
107. Kraus, D. M. *et al.* CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. *J. Immunol.* **176**, 4419–4430 (2006).
108. Kokura, K., Sun, L., Bedford, M. T. & Fang, J. Methyl-H3K9-binding protein MPP8 mediates E-cadherin gene silencing and promotes tumour cell motility and invasion. *EMBO J.* **29**, 3673–3687 (2010).
109. Yu, K., Lujan, R., Marmorstein, A., Gabriel, S. & Hartzell, H. C. Bestrophin-2 mediates bicarbonate transport by goblet cells in mouse colon. *J. Clin. Invest.* **120**, 1722–1735 (2010).
110. Glinisky, G. V. SNP-guided microRNA maps (MirMaps) of 16 common human disorders identify a clinically accessible therapy reversing transcriptional aberrations of nuclear import and inflammasome pathways. *Cell Cycle* **7**, 3564–3576 (2008).
111. Faustin, B. *et al.* Reconstituted NALP1 inflammasome reveals two-step mechanism of caspase-1 activation. *Mol. Cell* **25**, 713–724 (2007).
112. Tiala, I. *et al.* The PSORS1 locus gene CCHCR1 affects keratinocyte proliferation in transgenic mice. *Hum. Mol. Genet.* **17**, 1043–1051 (2008).
113. Chapkin, R. S. *et al.* Use of a novel genetic mouse model to investigate the role of folate in colitis-associated colon cancer. *J. Nutr. Biochem.* **20**, 649–655 (2009).
114. Pesu, M. *et al.* T-cell-expressed proprotein convertase furin is essential for maintenance of peripheral immune tolerance. *Nature* **455**, 246–250 (2008).
115. Dubois, C. M., Laprise, M. H., Blanchette, F., Gentry, L. E. & Leduc, R. Processing of transforming growth factor beta 1 precursor by human furin convertase. *J. Biol. Chem.* **270**, 10618–10624 (1995).

116. Xu, W. *et al.* A soluble class II cytokine receptor, IL-22RA2, is a naturally occurring IL-22 antagonist. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 9511–9516 (2001).
117. Junt, T. *et al.* Subcapsular sinus macrophages in lymph nodes clear lymph-borne viruses and present them to antiviral B cells. *Nature* **450**, 110–114 (2007).
118. Saito, Y. *et al.* Potentiation of cell invasion and matrix metalloproteinase production by alpha3beta1 integrin-mediated adhesion of gastric carcinoma cells to laminin-5. *Clin. Exp. Metastasis* **27**, 197–205 (2010).
119. Garg, P. *et al.* Matrix metalloproteinase-9-mediated tissue injury overrides the protective effect of matrix metalloproteinase-2 during colitis. *Am. J. Physiol. Gastrointest. Liver Physiol.* **296**, G175–184 (2009).
120. Zhang, X. *et al.* TILRR, a novel IL-1RI co-receptor, potentiates MyD88 recruitment to control Ras-dependent amplification of NF-kappaB. *J. Biol. Chem.* **285**, 7222–7232 (2010).
121. Manolakis, A. C. *et al.* Moderate performance of serum S100A12, in distinguishing inflammatory bowel disease from irritable bowel syndrome. *BMC Gastroenterol* **10**, 118 (2010).
122. Hofmann, M. A. *et al.* RAGE mediates a novel proinflammatory axis: a central cell surface receptor for S100/calgranulin polypeptides. *Cell* **97**, 889–901 (1999).
123. Foell, D., Wittkowski, H., Vogl, T. & Roth, J. S100 Proteins Expressed in Phagocytes: A Novel Group of Damage-Associated Molecular Pattern Molecules. *J Leukoc Biol* **81**, 28–37 (2007).
124. Fireman, E., Kraiem, Z., Sade, O., Greif, J. & Fireman, Z. Induced sputum-retrieved matrix metalloproteinase 9 and tissue metalloproteinase inhibitor 1 in granulomatous diseases. *Clin Exp Immunol* **130**, 331–337 (2002).
125. Arijs, I. *et al.* Predictive value of epithelial gene expression profiles for response to infliximab in Crohn's disease. *Inflamm. Bowel Dis.* **16**, 2090–2098 (2010).
126. Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
127. Gewirtz, A. T. *et al.* Dominant-negative TLR5 polymorphism reduces adaptive immune response to flagellin and negatively associates with Crohn's disease. *Am. J. Physiol. Gastrointest. Liver Physiol.* **290**, G1157–1163 (2006).
128. Glodek, A. M. *et al.* Ligation of complement receptor 1 increases erythrocyte membrane deformability. *Blood* **116**, 6063–6071 (2010).
129. Brown, G. D. & Gordon, S. Immune recognition. A new receptor for beta-glucans. *Nature* **413**, 36–37 (2001).
130. Shen, W., Stone, K., Jales, A., Leitenberg, D. & Ladisch, S. Inhibition of TLR activation and up-regulation of IL-1R-associated kinase-M expression by exogenous gangliosides. *J. Immunol.* **180**, 4425–4432 (2008).
131. Watanabe, Y. *et al.* Genome-wide analysis of expression modes and DNA methylation status at sense-antisense transcript loci in mouse. *Genomics* **96**, 333–341 (2010).
132. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339–346 (2012).
133. Imamura, T. *et al.* Non-coding RNA directed DNA demethylation of Sphk1 CpG island. *Biochem. Biophys. Res. Commun.* **322**, 593–600 (2004).
134. Murrell, A., Heeson, S. & Reik, W. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat. Genet.* **36**, 889–893 (2004).
135. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
136. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
137. Parkes, M. *et al.* Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat. Genet.* **39**, 830–832 (2007).

138. Franke, A. *et al.* Systematic association mapping identifies NELL1 as a novel IBD disease gene. *PLoS ONE* **2**, e691 (2007).
139. Franke, A. *et al.* Genome-wide association analysis in sarcoidosis and Crohn's disease unravels a common susceptibility locus on 10p12.2. *Gastroenterology* **135**, 1207–1215 (2008).
140. Shi, C.-S., Huang, N.-N., Harrison, K., Han, S.-B. & Kehrl, J. H. The Mitogen-Activated Protein Kinase Kinase Kinase Kinase GCKR Positively Regulates Canonical and Noncanonical Wnt Signaling in B Lymphocytes. *Mol Cell Biol* **26**, 6511–6521 (2006).
141. Belisle, J. A. *et al.* Identification of Siglec-9 as the receptor for MUC16 on human NK cells, B cells, and monocytes. *Mol. Cancer* **9**, 118 (2010).
142. Conze, T. *et al.* MUC2 mucin is a major carrier of the cancer-associated sialyl-Tn antigen in intestinal metaplasia and gastric carcinomas. *Glycobiology* **20**, 199–206 (2010).
143. Moehle, C. *et al.* Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. *J. Mol. Med.* **84**, 1055–1066 (2006).
144. Thiebault, K. *et al.* The netrin-1 receptors UNC5H are putative tumor suppressors controlling cell death commitment. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4173–4178 (2003).
145. Fukuda, T., Chen, K., Shi, X. & Wu, C. PINCH-1 is an obligate partner of integrin-linked kinase (ILK) functioning in cell shape modulation, motility, and survival. *J. Biol. Chem.* **278**, 51324–51333 (2003).
146. Shim, J., Lim, H., R Yates, J. & Karin, M. Nuclear export of NF90 is required for interleukin-2 mRNA stabilization. *Mol. Cell* **10**, 1331–1344 (2002).
147. Jiang, X. *et al.* Distinctive roles of PHAP proteins and prothymosin-alpha in a death regulatory pathway. *Science* **299**, 223–226 (2003).
148. Vucic, D., Stennicke, H. R., Pisabarro, M. T., Salvesen, G. S. & Dixit, V. M. ML-IAP, a novel inhibitor of apoptosis that is preferentially expressed in human melanomas. *Curr. Biol.* **10**, 1359–1366 (2000).
149. Sadikovic, B. *et al.* Identification of interactive networks of gene expression associated with osteosarcoma oncogenesis by integrated molecular profiling. *Hum. Mol. Genet.* **18**, 1962–1975 (2009).
150. Behrens, T. W. *et al.* Jaw1, A lymphoid-restricted membrane protein localized to the endoplasmic reticulum. *J. Immunol.* **153**, 682–690 (1994).
151. Schlesinger, T. K. *et al.* Apoptosis stimulated by the 91-kDa caspase cleavage MEKK1 fragment requires translocation to soluble cellular compartments. *J. Biol. Chem.* **277**, 10283–10291 (2002).
152. Oetke, C., Vinson, M. C., Jones, C. & Crocker, P. R. Sialoadhesin-deficient mice exhibit subtle changes in B- and T-cell populations and reduced immunoglobulin M levels. *Mol. Cell. Biol.* **26**, 1549–1557 (2006).
153. Grotgut, S. *et al.* Hepatocyte growth factor protects hepatoblastoma cells from chemotherapy-induced apoptosis by AKT activation. *Int. J. Oncol.* **36**, 1261–1267 (2010).
154. Saupe, S. *et al.* Molecular cloning of a human cDNA IGSF3 encoding an immunoglobulin-like membrane protein: expression and mapping to chromosome band 1p13. *Genomics* **52**, 305–311 (1998).
155. Minczuk, M., Mroczek, S., Pawlak, S. D. & Stepień, P. P. Human ATP-dependent RNA/DNA helicase hSuv3p interacts with the cofactor of survivin HBXIP. *FEBS J.* **272**, 5008–5019 (2005).
156. Suzuki, A. *et al.* ARK5 suppresses the cell death induced by nutrient starvation and death receptors via inhibition of caspase 8 activation, but not by chemotherapeutic agents or UV irradiation. *Oncogene* **22**, 6177–6182 (2003).
157. Abdgawad, M. *et al.* Elevated neutrophil membrane expression of proteinase 3 is dependent upon CD177 expression. *Clin. Exp. Immunol.* **161**, 89–97 (2010).



158. Rotondo, R. *et al.* IL-8 induces exocytosis of arginase 1 by neutrophil polymorphonuclears in nonsmall cell lung cancer. *International Journal of Cancer* **125**, 887–893 (2009).
159. Li, K. *et al.* Identification and expression of a new type II transmembrane protein in human mast cells. *Genomics* **86**, 68–75 (2005).
160. Kantari, C. *et al.* Proteinase 3, the Wegener Autoantigen, Is Externalized During Neutrophil Apoptosis: Evidence for a Functional Association with Phospholipid Scramblase 1 and Interference with Macrophage Phagocytosis. *Blood* **110**, 4086–4095 (2007).
161. Badola, S. *et al.* Correlation of serpin–protease expression by comparative analysis of real-time PCR profiling data. *Genomics* **88**, 173–184 (2006).
162. Hershey, G. K., Friedrich, M. F., Esswein, L. A., Thomas, M. L. & Chatila, T. A. The association of atopy with a gain-of-function mutation in the alpha subunit of the interleukin-4 receptor. *N. Engl. J. Med.* **337**, 1720–1725 (1997).
163. Moslemi, A.-R. *et al.* Glycogenin-1 deficiency and inactivated priming of glycogen synthesis. *N. Engl. J. Med.* **362**, 1203–1210 (2010).
164. Pouyet, L. *et al.* Epithelial vanin-1 controls inflammation-driven carcinogenesis in the colitis-associated colon cancer model. *Inflamm. Bowel Dis.* **16**, 96–104 (2010).
165. Martin, F. *et al.* Vanin-1(-/-) mice show decreased NSAID- and Schistosoma-induced intestinal inflammation associated with higher glutathione stores. *J. Clin. Invest.* **113**, 591–597 (2004).
166. Chong, K. W. Y. *et al.* Annexin A3 is associated with cell death in lactacystin-mediated neuronal injury. *Neuroscience Letters* **485**, 129–133 (2010).
167. Sutterwala, F. S. & Flavell, R. A. NLR4/IPAF: a CARD carrying member of the NLR family. *Clin Immunol* **130**, 2–6 (2009).
168. Poyet, J. L. *et al.* Identification of Ipaf, a human caspase-1-activating protein related to Apaf-1. *J. Biol. Chem.* **276**, 28309–28313 (2001).
169. Arce, I., Martínez-Muñoz, L., Roda-Navarro, P. & Fernández-Ruiz, E. The human C-type lectin CLECSF8 is a novel monocyte/macrophage endocytic receptor. *European Journal of Immunology* **34**, 210–220 (2004).
170. Langerholm, T. *et al.* Inhibitory properties of cystatin F and its localization in U937 promonocyte cells. *FEBS Journal* **272**, 1535–1545 (2005).
171. D'Sa-Eipper, C., Subramanian, T. & Chinnadurai, G. bfl-1, a bcl-2 homologue, suppresses p53-induced apoptosis and exhibits potent cooperative transforming activity. *Cancer Res.* **56**, 3879–3882 (1996).
172. Nath, S. K. *et al.* A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. *Nat. Genet.* **40**, 152–154 (2008).
173. Waetzig, G. H., Seeger, D., Rosenstiel, P., Nikolaus, S. & Schreiber, S. P38 Mitogen-Activated Protein Kinase Is Activated and Linked to TNF-A Signaling in Inflammatory Bowel Disease. *J Immunol* **168**, 5342–5351 (2002).
174. Kashyap, D. R. *et al.* Peptidoglycan recognition proteins kill bacteria by activating protein-sensing two-component systems. *Nat. Med.* **17**, 676–683 (2011).
175. Yanai, H. *et al.* HMGB proteins function as universal sentinels for nucleic-acid-mediated innate immune responses. *Nature* **462**, 99–103 (2009).
176. Eidelwein, A. *et al.* Gene Expression Profiles in Children With Newly Diagnosed Crohn's Disease: 203. *Journal of Pediatric Gastroenterology and Nutrition* **41**, (2005).
177. Pagès, G. *et al.* Defective thymocyte maturation in p44 MAP kinase (Erk 1) knockout mice. *Science* **286**, 1374–1377 (1999).
178. Yamazaki, T. *et al.* Essential immunoregulatory role for BCAP in B cell development and function. *J. Exp. Med.* **195**, 535–545 (2002).
179. Nicholson, D. W. *et al.* Identification and inhibition of the ICE/CED-3 protease necessary for mammalian apoptosis. *Nature* **376**, 37–43 (1995).

180. Tong, W.-H. & Rouault, T. A. Functions of mitochondrial ISCU and cytosolic ISCU in mammalian iron-sulfur cluster biogenesis and iron homeostasis. *Cell Metab.* **3**, 199–210 (2006).
181. Spiliotis, E. T., Kinoshita, M. & Nelson, W. J. A mitotic septin scaffold required for Mammalian chromosome congression and segregation. *Science* **307**, 1781–1785 (2005).
182. Bloss, T. A., Witze, E. S. & Rothman, J. H. Suppression of CED-3-independent apoptosis by mitochondrial betaNAC in *Caenorhabditis elegans*. *Nature* **424**, 1066–1071 (2003).

## **9. List of Figures**

Figure 1	Modified schematic for Crohn’s disease associated genes (red) and gene interactions according to (Fransen <i>et al.</i> 2011). <sup>17</sup> .....	10
Figure 2	Filtering step for SV characterization.....	27
Figure 3	<i>Janus</i> workflow.....	32
Figure 4	Estimation of antisense noise level.....	33
Figure 5	Antisense classes.....	34
Figure 6	Scoring algorithm for S/AS pairs.....	35
Figure 7	Calculations helping to identify S/AS pairs which are showing an equal expression (Co-expressed S/AS pairs).....	36
Figure 8	Estimation of sense and antisense expression near methylated CpG islands .....	37
Figure 9	Diagrams showing shared genetic variants between the family, dbSNP and the 1000 genome.....	40
Figure 10	Sanger sequencing result of <i>IKZF1</i> recessive SNV. ....	41
Figure 11	Sanger sequencing result of <i>CD19</i> novel SNV.....	41
Figure 12	Sanger sequencing result for <i>MST1</i> recessive SNV.....	42
Figure 13	Sanger sequencing result for <i>NF1</i> <i>de novo</i> SNV.....	42
Figure 14	SNV-based principal component analysis (PCA) of the family trio with 11 HapMap populations. ....	43
Figure 15	Contribution of known CD risk loci to overall risk in comparison with other genomes... 46	
Figure 16	Gaussian approximation assuming independent and additive contributions of the various risk alleles at the 71 known CD-risk loci.....	47
Figure 17	Sanger sequencing result of <i>FUT2</i> nonsense SNV.....	48
Figure 18	Regions of homozygosity identified in the family trio using a variant-density adapted definition.....	50
Figure 19	Example of a region of homozygosity in the child.....	50
Figure 20	Overlap of regions of homozygosity with GWAS annotated CD risk regions of the family trio in with unrelated individuals. ....	52
Figure 21	Verification of <i>BEST2</i> SNV by Sanger sequencing.....	53
Figure 22	Large deletion affecting the <i>SPINK5L2</i> exon 3. ....	55
Figure 23	SNV density plot for 60 individuals (top 60 lines) from the 1000 Genome Project and the family trio (lowest three lines). ....	56
Figure 24	String network of identified missense SNVs in the child which are contained in the gene ontology term “immune system process”.....	57

Figure 25	Antisense noise-level estimation.....	59
Figure 26	Antisense transcription in promoter and terminator regions of annotated transcripts for five samples (730-734 = GM11894, GM07374, GM12489, GM12045, GM11919).....	61
Figure 27	Differentially expressed S/AS pair in LBL cell lines.....	62
Figure 28	Coexpressed S/AS pair in LBL cell lines. ....	63
Figure 29	IGV picture of SNV at chr15:31234064.....	63

## **10. List of Tables**

Table 1	Overview of applied library and run chemistry.....	25
Table 2	Per library numbers of generated sequences, mapped sequences and coverage depth across the genome. ....	38
Table 3	Variant detection statistic.....	39
Table 4	Verification of <i>de novo</i> SNVs by Sanger sequencing. ....	43
Table 5	Genetic variation in genes of monogenic disorders phenocopying Crohn's disease. ....	44
Table 6	Zygosity of the 71 meta-analysis Crohn's disease risk SNVs for the risk-allele.....	46
Table 7	Regions of homozygosity overlapping Crohn's disease associated genes.....	51
Table 8	Plink Allelic Association Test with confidence interval and case/control-only.....	54
Table 9	Plink Hardy-Weinberg calculation.....	54
Table 10	Number of antisense events in promoter and terminator regions.....	60

## **11. List of abbreviations**

abbreviation	full name
AS	antisense
BAM	binary SAM (see SAM)
bp	base pair
CAGE	cap analysis gene expression
CD	Crohn's disease
CGD	chronic granulomatous disease
cHL	classical Hodgkin's lymphoma
CVID	common variable immunodeficiency
DAMP	damage-associated molecular patterns
DNA	deoxyribonucleic acid
DSS	dextran sodium sulfate
ELISA	enzyme-linked immunosorbent assay
ePCR	emulsion PCR (see PCR)
ER	endoplasmic reticulum
FPKM	Fragments Per Kilobase of exon per Million fragments mapped
gDNA	genomic DNA
GWAS	genome-wide association study
GTF	gene transfer format
GO	gene ontology
HapMap	haplotype map
HLA	human leukocyte antigen
IBD	inflammatory bowel disease
IBS	irritable bowel syndrome
IFX	Infliximab
InDel	insertion or deletion
IP	immunoproteasome
IPEX	immune dysregulation, polyendocrinopathy enteropathy, X-linked syndrome
LBL	lymphoblastoid cell line
LMP	low molecular weight polypeptides (as protein subunit) long mate-pair (in library generation)
LPS	lipopolysaccharide
LRR	leucine-rich-repeat
ISV	long structural variant
MAP	<i>Mycobacterium avium</i> subspecies <i>paratuberculosis</i>
MB	mannan-binding
MDP	muramyl dipeptide
MHC	major histocompatibility complex
mRNA	messenger RNA
miRNA	micro RNA
MS	multiple sclerosis
NGS	next generation sequencing
OMIM	online mendelian inheritance in man
OR	odds ratio
PAMP	pathogen associated molecular patterns
PCA	principal component analysis
PE	paired-end

PCR	polymerase-chain-reaction
PRR	pattern recognition receptors
RNA	ribonucleic acid
RA	rheumatoid arthritis
ROH	region of homozygosity
SAM	sequence alignment / map format
S/AS	sense-antisense pair
SLE	systemic lupus erythematosus
SNV	single nucleotide variant
SP	standard proteasome
sSV	short structural variant
SV	structural variant
TAR	transcriptional active region
UC	ulcerative colitis
UPR	unfolded protein response
UTR	untranslated region
VAT	variant annotation tool
WAS	Wiskott–Aldrich syndrome
WGS	whole genome sequencing
WTAK	whole transcriptome analysis kit
Illumina GA	Illumina Genome Analyzer

## 12. Statement of authorship

Hiermit erkläre ich, dass die vorliegende Arbeit - abgesehen von den Beratungen durch meine akademischen Lehrer - nach Inhalt und Form meine eigene Arbeit ist. Die Arbeit wurde bis jetzt weder vollständig noch in Teilen einer anderen Stelle im Rahmen eines Prüfungsverfahrens vorgelegt. Ferner erkläre ich, dass ich noch keine früheren Promotionsversuche unternommen habe.

Kiel, .....

.....

(Matthias Barann)

## 13. Curriculum vitae

<b>First- and Last Name:</b>	Matthias Barann
<b>Location:</b>	Gravelottestraße 5 24116 Kiel
<b>Date of birth:</b>	21.10.1983
<b>Nationality:</b>	German
<b>School education:</b>	
08/1990 – 08/1994	Elementary school Ritterkamp, Bremen
08/1994 – 08/1996	Schulzentrum an der Lerchenstraße, Bremen (orientation stage)
08/1996 – 08/2000	Schulzentrum an der Lerchenstraße, Bremen (grammar school)
08/2000 – 08/2003	Schulzentrum des Sekundarbereichs II an der Bördestraße, Bremen (grammar school)
<b>Course of studies:</b>	
10/2003 – 01/2008	University Bremen, Biology Diploma Major Subject: Genetics / Cell- and Molecular Biology 1. Minor Subject: Microbiology 2. Minor Subject: Computer Science
04/2008 - 12/2008	Research associate at the Institute for Cell Biology and Immunology, University of Stuttgart
08/2009 – 09/2012	Research associate at the Institute for Clinical Molecular Biology Christian-Albrechts-University in Kiel



## 14. Publications

Schmid MW, Schmidt A, Klostermeier UC, **Barann M**, Rosenstiel P, Grossniklaus U. *A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing* PLoS One. 2012;7(1):e29685. Epub 2012 Jan 26.

Klostermeier UC<sup>†</sup>, **Barann M**<sup>†</sup>, Wittig M, Häslér R, Franke A, Gavrilova O, Kreck B, Sina C, Schilhabel MB, Schreiber S, Rosenstiel P. *A tissue-specific landscape of sense/antisense transcription in the mouse intestine* BMC Genomics. 2011 Jun 10;12:305. doi: 10.1186/1471-2164-12-305.

Autran D, Baroux C, Raissig MT, Lenormand T, Wittig M, Grob S, Steimer A, **Barann M**, Klostermeier UC, Leblanc O, Vielle-Calzada JP, Rosenstiel P, Grimanelli D, Grossniklaus U. *Maternal epigenetic pathways control parental contributions to Arabidopsis early embryogenesis* Cell. 2011 May 27;145(5):707-19.

Schramm A, Köster J, Marschall T, Heilmann M, Fielitz K, **Barann M**, Esser D, Rosenstiel P, Rahmann S, Eggert A, Schulte J. *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* International Journal of Cancer.

## 15. Submitted Manuscripts

Rosenstiel R<sup>†</sup>, **Barann M**<sup>†</sup>, Klostermeier UC, Sheth V, Ellinghaus D, Rausch T, Korbel J, Nothnagel M, Krawczak M, Gilissen C, Veltman J, Forster M, Stade B, McLaughlin S, Lee CC, Fritscher-Ravens A, Franke A<sup>†</sup>, Schreiber S<sup>†</sup>. *Whole Genome Sequence of a Crohn disease trio*

**Barann M**, Esser D, Klostermeier UC, Ammerpohl O, Siebert R, Sudbrak R, Lehrach H, Schreiber S, Rosenstiel R. *Janus – Investigating the two faces of transcription*

Forster M, Forster P, Elsharawy A, Hemmrich G, Kreck B, Wittig M, Thomsen I, Stade B, **Barann M**, Ellinghaus D, Petersen BS, May S, Melum E, Schilhabel MB, Keller A, Schreiber S, Rosenstiel P, Franke A. *From next-generation sequencing alignments to accurate comparison and validation of single nucleotide variants: the pibase software*

---

<sup>†</sup> shared authorship

Kreck B, Richter J, Ammerpohl O, **Barann M**, Esser D, Pedersen B, Vater I, Penas EMM, Chung C, Seisenberger S, Boyd VL, Drexler H, MacLeod R, Hummel M, Krueger F, Häsler R, Rosenstiel P, Franke A<sup>†</sup>, Siebert R<sup>†</sup>. *Generation and Analysis of a Base-Pair Resolution DNA Methylome from an Endemic Burkitt Lymphoma obtained by complete Bisulfite Sequencing*

---

<sup>†</sup> shared authorship

## 16. Acknowledgements

Ich danke Herrn Professor Dr. Stefan Schreiber für seine Förderung und Unterstützung, sowie Herrn Professor Dr. Philip Rosenstiel für die gute Betreuung und Diskussion meiner Arbeit. Zudem danke ich der Bruhn-Stiftung, im Besonderen Hans Dietrich und Annegret Bruhn, für die Auszeichnung mit dem Bruhn-Förderpreis. Desweiteren danke ich Ulrich C. Klostermeier für die Unterstützung bei der RNA und DNA Präparation für die Sequenzierung, David Ellinghaus für die Erstellung der ‚principle component‘ Analyse, Michael und Peter Forster für ihren Beitrag zur Analyse der mitochondrialen Varianten für die Bestimmung des geographischen Ursprungs der Familie, Tobias Rausch und Jan Korbel für den Input zur CNV Analyse, den helfenden Händen von Christian Gilissen und Joris A. Veltman bei der Detektion von *de novo* Varianten, Michael Nothnagel und Michael Krawczak für die Diskussion der statistischen Modelle, Melanie Schlapkohl für die Unterstützung bei der Verifikation einiger genetischer Varianten, sowie den MitarbeiterInnen der beiden Sequenzierlabore. Ich möchte auch meinen Kollegen aus der Zellbiologie danken, die ausschlaggebend für die angenehme Arbeitsatmosphäre waren.

## 17. Supplementary Material

### List of supplementary tables

Supplementary table 1	Crohn's disease risk variants with known odds ratios described in the GWAS catalog. ....	102
Supplementary table 2	Potential immune functions of genes associated with CD risk. ....	103
Supplementary table 3	Expression values by RNA-Seq. ....	105
Supplementary table 4	Primer sequences used for Sanger sequencing of SNVs. ....	105
Supplementary table 5	Summary of mapping statistics and raw data output for son. ....	109
Supplementary table 6	Summary of mapping statistics and raw data output for mother. ....	110
Supplementary table 7	Summary of mapping statistics and raw data output for father. ....	111
Supplementary table 8	Summary of achieved coverage. ....	112
Supplementary table 9	SNV calling statistics. ....	113
Supplementary table 10	Bioscope SNV correlation with Illumina iScan 1M human Omniquad chip. ....	114
Supplementary table 11	small structural variation (sSV) calling statistics. ....	115
Supplementary table 12	large structural variation (ISV) calling statistics. ....	115
Supplementary table 13	Manually reviewed ISVs. ....	116
Supplementary table 14	Sanger sequencing of potential <i>de novo</i> SNVs. ....	116
Supplementary table 15	Overview of noncoding SNVs detected in the child in genes associated with monogenic phenocopies of Crohn disease. ....	117
Supplementary table 16	Nonsynonymous SNVs detected in the child in genes associated with monogenic phenocopies of Crohn's disease. ....	117
Supplementary table 17	Noncoding SNVs detected in the child in conserved regions in genes associated monogenic phenocopies of Crohn disease. ....	118
Supplementary table 18	small structural variations detected in the child in genes associated with monogenic phenocopies of Crohn disease. ....	118
Supplementary table 19	Nonsynonymous SNVs predicted to be damaging associated with genes in proximity of GWAS Crohn's disease SNVs. ....	119
Supplementary table 20	Exonic and splice-junction small SVs in genes in proximity of GWAS Crohn's disease SNVs. ....	119
Supplementary table 21	Detected CD associated SNVs (meta-analysis). ....	120
Supplementary table 22	Known Crohn's Disease associated SNVs detected (GWAS catalog). ....	120
Supplementary table 23	Overview of noncoding SNVs associated with Genes in proximity of GWAS CD SNVs in the child. ....	120
Supplementary table 24	Noncoding SNVs in conserved regions of Crohn's Disease associated genes. ....	120
Supplementary table 25	Overview of noncoding small structural variations in genes associated with Crohn's Disease. ....	120
Supplementary table 26	Nonsynonymous SNVs in genes involved in other inflammatory diseases, which are predicted to be damaging. ....	120
Supplementary table 27	Exonic and splice-junction sSVs in genes associated with other inflammatory diseases. ....	121
Supplementary table 28	Large structural variations in proximity of genes involved in other inflammatory disease. ....	121

Supplementary table 29	Overview of noncoding SNVs in genes involved in other inflammatory diseases.....	122
Supplementary table 30	Noncoding SNVs in genes associated with inflammatory diseases following the recessive model (also including one non-inflammatory disease: Parkinson).....	122
Supplementary table 31	Overview of noncoding SNVs in genes associated with other inflammatory diseases in the child in locations of high conservation...122	
Supplementary table 32	Small structural variations in genes associated with inflammatory diseases in the GWAS catalog.....	122
Supplementary table 33	Nonsynonymous SNVs detected in the child following the recessive model of inheritance which are predicted to be damaging.....	123
Supplementary table 34	Exonic SNVs in other genes contributing to compound heterozygosity with at least one SNV predicted to be damaging.....	125
Supplementary table 35	Functional associations of genes listed in Error! Reference source not found..	128
Supplementary table 36	Excerpt of potential damaging exonic SNVs in conserved regions of other functional interesting genes in the child.....	129
Supplementary table 37	Genetic variants in conserved coding regions of genes. ....	130
Supplementary table 38	Functional associations of genes listed in Error! Reference source not found..	131
Supplementary table 39	sSVs in coding regions of other genes following the recessive model..	132
Supplementary table 40	Expression values calculated with Cufflinks (FPKM) for <i>MPHOSPH8</i> and <i>BEST2</i> in the transcriptomes from blood and colon biopsy (only for the child). ....	125
Supplementary table 41	Summary of genes 5-fold up- or downregulated in the child compared to parents. ....	133
Supplementary table 42	Exonic missense SNVs detected in genes differentially expressed in the child compared to the parents.....	135
Supplementary table 43	Average beta methylation values for ZSCAN16 (ZNF435) promoter region. ....	140

#### List of supplementary figures

Supplementary figure 1	Detection of antisense events. ....	107
Supplementary figure 2	Examples of possible sense/antisense pairs.....	108
Supplementary figure 3	Assessment of chromosome and allele specific expression.....	108
Supplementary figure 4	Origin determination by mitochondrial variations based on Röhl <i>et al.</i> 2001.....	116
Supplementary figure 5	Differentially expressed S/AS pair in cHL cell lines.....	136
Supplementary figure 6	Differentially expressed antisense event in <i>Mtus1</i> in murine tissues. ....	137
Supplementary figure 7	Coexpressed S/AS pair in cHL cell lines.....	138
Supplementary figure 8	Coexpressed S/AS example in <i>Tcf7l1</i> in murine tissues.....	139
Supplementary figure 9	Differentially expressed antisense event in ZSCAN16 in cHL cell lines. ....	140

**Supplementary table 1 Crohn's disease risk variants with known odds ratios described in the GWAS catalog.**

Genes are ordered by highest odds ratio and genes mapping directly to SNV positions are bold. For multiple entries of the same SNV, the most complete or most recent entry was chosen. Lower and upper limits of the 95% confidence interval are shown in brackets. GWAS catalog data was accessed 05/01/2012. NR = not reported. The odds ratio is a measure of the probability of an event to occur in two situations. An odds ratio of 1 means both events are equally likely, while a higher OR means the event is more likely in group 1 (here in CD patients).

#	gene(s)	variant	OR	reference
1	<b>NOD2</b>	rs2066847-C	3.99 [NR]	45
	<b>NOD2</b> , <i>CYLD</i> , <i>SNX20</i> , <i>NKD1</i>	rs2076756-G	1.66 [1.48-1.88]	50
	<b>NOD2</b>	rs17221417-G	1.29 [1.13-1.46]	135
2	<b>IL23R</b>	rs11209026-G	2.66 [2.36-3.00]	12
	<b>IL23R</b>	rs11465804-?	1.89 [1.47-2.44]	57
	<b>IL23R</b>	rs11805303-T	1.39 [1.22-1.58]	136
3	<i>MUC19</i> , <i>LRRK2</i>	rs11564258-A	1.74 [1.55-1.95]	12
	<i>LRRK2</i> , <i>MUC19</i>	rs11175593-T	1.54 [NR]	45
4	<i>IRGM</i>	rs1000113-T	1.54 [1.31-1.82]	136
	<i>IRGM</i>	rs13361189-?	1.38 [1.15-1.66]	137
	<i>IRGM</i>	rs7714584-G	1.37 [1.28-1.47]	12
	<i>IRGM</i>	rs11747270-G	1.33 [NR]	45
5	<b>SLC6A1</b>	rs7705924-G	1.48 [1.17-1.87]	50
6	<i>IL12B</i>	rs10045431-?	1.45 [1.27-1.64]	57
7	<i>HLA-F</i> , <i>MOG</i> , <i>HLA-G</i> , <i>GABBR1</i> , <i>HLA-H</i> , <i>UBD</i> , <i>HLA-A</i>	rs9258260-T	1.45 [1.21-1.68]	50
8	<b>PSMB10</b>	rs11574514-A	1.44 [1.35-1.52]	50
9	<i>PTGER4</i>	rs1992660-?	1.42 [1.24-1.67]	138
	<i>PTGER4</i>	rs9292777-T	1.37 [1.28-1.48]	50
	<i>PTGER4</i>	rs11742570-C	1.33 [1.27-1.39]	12
	<i>PTGER4</i>	rs4613763-C	1.32 [NR]	45
10	<i>CCR6</i> , <b>FGFR10P</b> , <i>RNASE2</i>	rs2301436-?	1.37 [1.22-1.53]	57
11	<i>IBD5</i>	rs2188962-?	1.36 [1.21-1.52]	57
12	<i>PTPN2</i>	rs2542151-G	1.35 [NR]	45
13	<b>ATG16L1</b>	rs3792109-A	1.34 [1.29-1.40]	12
	<b>ATG16L1</b> , <i>SAG</i> , <i>DGKD</i> , <i>INPP5D</i> , <i>USP40</i>	rs2241880-G	1.32 [1.24-1.41]	50
	<b>ATG16L1</b>	rs3828309-G	1.25 [NR]	45
	<b>ATG16L1</b>	rs10210302-T	1.19 [1.01-1.41]	136
14	<b>NKX2-3</b> , <i>SLC25A28</i> , <i>GOT1</i> , <i>ENTPD7</i> , <i>CNNM1</i> , <i>COX15</i> , <i>CUTC</i>	rs11190141-C	1.34 [1.25-1.43]	50
	<i>NKX2-3</i>	rs4409764-T	1.22 [1.17-1.27]	12
	<b>NKX2-3</b>	rs11190140-T	1.20 [NR]	45
	<i>NKX2-3</i>	rs10883365-G	1.20 [1.03-1.39]	136
15	<b>PTPN22</b>	rs2476601-G	1.31 [NR]	45
16	<b>NELL1</b>	rs1793004-?	1.30 [1.12-1.52]	138
17	<i>TCERG1L</i>	rs10734105-G	1.27 [1.10-1.43]	50
18	<i>ZNF365</i>	rs10995271-C	1.25 [NR]	45
	<i>ZNF365</i>	rs10761659-G	1.23 [1.18-1.29]	12
	<b>ZNF365</b> , <i>ERG2</i> , <i>ADO</i>	rs7076156-G	1.19 [1.10-1.30]	50
19	<i>SLC22A4</i> , <i>SLC22A5</i> , <i>IRF1</i> , <i>IL3</i>	rs12521868-T	1.23 [1.18-1.28]	12
	<i>IL3</i> , <i>ACSL6</i> , <i>P4HA2</i> , <i>PDLIM4</i> , <i>SLC22A4</i>	rs3091338-T	1.23 [1.08-1.42]	50
20	<i>MAP3K7IP1</i>	rs2413583-C	1.23 [1.17-1.29]	12
21	<i>GALC</i> , <b>GPR65</b>	rs8005161-T	1.23 [1.16-1.31]	12
22	<b>C10orf67</b>	rs1398024-A	1.23 [1.04-1.45]	139
23	<b>TNFSF15</b>	rs4263839-G	1.22 [NR]	45
	<b>TNFSF15</b> , <i>TNFSF8</i>	rs3810936-C	1.21 [1.15-1.27]	12
24	<b>MST1</b> , <i>GPX1</i> , <i>BSN</i>	rs3197999-A	1.22 [1.16-1.27]	12
	<i>BSN</i> , <b>MST1</b>	rs9858542-A	1.09 [0.96-1.24]	136
25	<b>CDKAL1</b>	rs6908425-C	1.21 [NR]	45
26	<b>CCL2</b> , <b>CCL7</b>	rs3091315-A	1.20 [1.14-1.26]	12
	<b>CCL7</b> , <b>CCL2</b> , <i>CCL11</i> , <i>CCL8</i> , <i>CCL13</i> , <i>CCL1</i>	rs3091316-G	1.14 [1.03-1.27]	50
27	<b>ZMIZ1</b>	rs1250550-G	1.19 [1.15-1.23]	12
28	<b>IL18RAP</b> , <i>IL12RL2</i> , <i>IL18R1</i> , <i>IL1RL1</i>	rs2058660-G	1.19 [1.14-1.26]	12
29	<i>UBE2D1</i>	rs1819658-C	1.19 [1.13-1.25]	12
30	<b>STAT3</b>	rs744166-A	1.18 [NR]	45
31	<b>CARD9</b> , <i>SNAPC4</i>	rs4077515-T	1.18 [1.13-1.22]	12
32	<b>C11orf30</b>	rs7927894-T	1.16 [NR]	45
33	<i>TMEM17</i> , <i>EHBP1</i> , <i>CPAMD8</i> , <i>AK3</i>	rs6545946-C	1.16 [1.06-1.27]	50
34	<i>SLC43A3</i> , <i>PRG2</i> , <i>PRG3</i>	rs11229030-C	1.15 [1.10-1.39]	50
35	<b>GCKR</b>	rs780093-T	1.15 [1.10-1.21]	12
36	<b>C8orf84</b> , <i>TERF1</i> , <i>RPL7</i> , <i>RDH10</i> , <i>KCNB2</i>	rs12677663-T	1.15 [1.04-1.28]	50
37	<b>ITLN1</b>	rs2274910-C	1.14 [NR]	45
38	<b>THADA</b>	rs10495903-T	1.14 [1.09-1.20]	12

#	gene(s)	variant	OR	reference
39	<i>C2orf74,REL</i>	rs10181042-T	1.14 [1.09-1.19]	<sup>12</sup>
40	<i>ICOSLG</i>	rs762421-G	1.13 [NR]	<sup>45</sup>
41	<b>SCAMP3</b> , <i>MUC1</i>	rs3180018-A*	1.13 [1.06-1.19]	<sup>12</sup>
42	<i>JAK2</i>	rs10758669-C	1.12 [NR]	<sup>45</sup>
43	<i>ORMDL3</i>	rs2872507-A	1.12 [NR]	<sup>45</sup>
44	<b>SP140</b>	rs7423615-T	1.12 [1.07-1.18]	<sup>12</sup>
45	<i>IL10, IL19</i>	rs3024505-T	1.12 [1.07-1.17]	<sup>12</sup>
46	<b>SMAD3</b>	rs17293632-T	1.12 [1.07-1.16]	<sup>12</sup>
47	<b>TYK2</b> , <i>ICAM1, ICAM3</i>	rs12720356-G	1.12 [1.06-1.19]	<sup>12</sup>
48	<i>RTEL1, TNFRSF6B, SLC2A4RG</i>	rs4809330-G	1.12 [1.06-1.18]	<sup>12</sup>
49	<b>IL2RA</b>	rs12722489-C	1.11 [1.05-1.16]	<sup>12</sup>
50	<i>YDJC</i>	rs181359-T	1.10 [1.06-1.15]	<sup>12</sup>
51	<i>PRDX5, ESRRA</i>	rs694739-A	1.10 [1.05-1.16]	<sup>12</sup>
52	<i>TNFSF11</i>	rs2062305-G	1.10 [1.05-1.15]	<sup>12</sup>
53	<i>TAGAP</i>	rs212388-G	1.10 [1.05-1.14]	<sup>12</sup>
54	<i>MTMR3</i>	rs713875-C	1.08 [1.04-1.13]	<sup>12</sup>
55	<b>CPEB4</b> <i>CPEB4</i>	rs359457-T rs359457-T	1.08 [1.04-1.12] 1.08 [1.04-1.12]	<sup>12</sup> <sup>12</sup>
56	<b>FADS1</b>	rs102275-C	1.08 [1.04-1.12]	<sup>12</sup>
57	<i>ZFP36L1</i>	rs4902642-G	1.07 [1.11-1.04]	<sup>12</sup>
58	<b>FUT2</b> , <i>RASIP1</i>	rs281379-A	1.07 [1.04-1.11]	<sup>12</sup>
59	<i>IL27, SH2B1, EIF3C, LAT, CD19</i>	rs151181-G	1.07 [1.03-1.12]	<sup>12</sup>
60	<b>BACH2</b>	rs1847472-G	1.07 [1.03-1.11]	<sup>12</sup>
61	<b>DNMT3A</b>	rs13428812-G	1.06 [1.03-1.10]	<sup>12</sup>
62	<i>NDFIP1</i>	rs11167764-C	1.06 [1.02-1.11]	<sup>12</sup>
63	<b>PLCL1</b>	rs6738825-A	1.06 [1.02-1.11]	<sup>12</sup>
64	<b>ERAP2, LRAP</b>	rs2549794-C	1.05 [1.02-1.09]	<sup>12</sup>
65	<i>VAMP3</i>	rs2797685-A	1.05 [1.01-1.10]	<sup>12</sup>
66	<b>PUS10</b> , <i>PEX13, REL, KIAA1841, C2orf74, PAPOLG, USP34</i>	rs13003464-G	1.05 [1.00-1.40]	<sup>50</sup>
67	<b>DENND1B</b>	rs1998598-G	1.04 [1.00-1.09]	<sup>12</sup>

**Supplementary table 2 Potential immune functions of genes associated with CD risk.**

The functional description is mainly from OMIM and RefSeq.

gene(s)	function
<b>NOD2</b>	- intracellular receptor for bacterial products - leads to NF- $\kappa$ B activation - induces autophagy in dendritic cells - interacts with MHC II antigen presentation machinery
<b>IL23R</b>	- associated with JAK2 and STAT3 - IL-12 stimulation induces production of IL-17 and IFNG by lymphoid cells expressing Thy1, stem cell antigen-1, retinoic acid-related orphan receptor (Ror)- $\gamma$ -t; (RORC), and IL23R
<i>MUC19, LRRK2</i>	<i>MUC19</i> - gel-forming mucin <i>LRRK2</i> - might be involved in the endosomal-autophagic pathway
<i>IRGM</i>	- gamma-interferon (IFNG)-induced GTP-binding proteins - control of intracellular pathogens - plays a role in autophagy
<b>SLCO6A1</b>	NA
<i>IL12B</i>	- produced by macrophages, induced by bacterial lipoproteins via TLRs - part of IL12B/p19 complex (=IL23), high expression of IL23 in dendritic cells - IL23 enhances IFNG secretion by memory T-cells in IL-2 dependent manner - induces production of IL22 in T-cells
<i>HLA-F, MOG, HLA-G, GABBR1, HLA-H, UBD, HLA-A</i>	HLAs = human leukocyte antigen, involved in MHC presentation of antigens
<b>PSMB10</b>	- part of the immunoproteasome - processing of MHC class I peptides
<i>PTGER4</i>	- Involved in migration of Langerhans' cells to lymph nodes
<i>CCR6</i>	- CC chemokine receptor - might interact with beta-defensins and promote adaptive immune responses by recruiting dendritic and T cells to the site of microbial invasion
<i>IBD5</i>	NA
<i>PTPN2</i>	- potential tumor suppressor gene - expression might modify beta-cell responses to dsRNA - might counter dsRNA induced apoptosis
<b>ATG16L1</b>	- component of the autophagy complex - required for autophagy in dendritic cells - ATG16L1 deficient macrophages produce high amounts of IL-1 $\beta$ and IL-18

gene(s)	function
<b>NKX2-3</b>	- induces expression of <i>MADCAM1</i> , which is probably involved in homing of lymphocytes and macrophages - homozygous KO mice lack <i>MADCAM1</i> expression
<b>PTPN22</b>	- might impair antigen dependent response of B- and T-cell receptors
<b>NELL1</b>	NA
<b>TCERG1L</b>	NA
<b>ZNF365</b>	NA
<b>SLC22A4/SLC22A5</b>	NA
<b>MAP3K7IP1</b>	- may interact in the signaling between TFGBR and TAK1
<b>GPR65</b>	- might be involved in elimination of auto reactive immature thymocytes
<b>C10orf67</b>	NA
<b>TNFSF15</b>	- might induce apoptosis of endothelial cells by activation of JNK, p38 MAPK and caspases - inducible by IL-1 and TNF - induces NFkB expression in death receptor 3 expressing cells - enhances IL2RA and IL2RB expression of T-cells, and thus IL-2 dependent proliferation and secretion of IFNG and GMCSF.
<b>MST1</b>	- delayed macrophage activation in KO mice
<b>CDKAL1</b>	NA
<b>CCL7, CCL2</b>	<i>CCL7</i> - promotes inflammatory cellular response via induction of TGFb1 - cleaved CCL7 acts as chemokine antagonist and dampens inflammation <i>CCL2</i> - regulated by NFkB - attracts leukocytes to sites of inflammation
<b>ZMIZ1</b>	NA
<b>IL18RAP</b>	- co-expression with IL18R1 required for IL-18 dependent activation of NFkB and JNK
<b>UBE2D1</b>	- might induce ubiquitination of p53 by E6AP
<b>STAT3</b>	- acute phase response gene - activates gp130, which is required for IL-6 dependent T-cell recruitment
<b>CARD9</b>	- interacts with CARD domain of BCL10 and leads to NFkB activation - key transducer of dectin-1 (PRR of myeloid phagocytes) signaling - synergistic effect with NOD2
<b>C11orf30</b>	NA
<b>TMEM17, EHP1, CPAMD8, AK3</b>	<i>TMEM17, EHP1, AK3</i> - NA <i>CPAMD8</i> - member of complement component 3 family, involved in innate immunity and damage control
<b>SLC43A3, PRG2, PRG3</b>	<i>SLC43A3</i> - NA <i>PRG2, PRG3</i> - cytotoxic polypeptide - microbicidal activity against gram-negative and gram-positive bacteria and fungi
<b>GCKR</b>	facilitates both canonical and noncanonical Wnt signaling in B lymphocytes <sup>140</sup>
<b>C8orf84</b>	NA
<b>ITLN1</b>	NA
<b>THADA</b>	NA
<b>C2orf74,REL</b>	NA
<b>ICOSLG</b>	- enhances T-cell proliferation and production of IL-2 dependent cytokines - increased expression on B-cells and monocytes and decreased expression in dendritic cells upon TNF- $\alpha$ stimulation - increased surface expression on dendritic cells stimulated by GMCSF
<b>SCAMP3, MUC1</b>	<i>SCAMP3</i> -NA <i>MUC1</i> - transmembrane mucin - physical barrier with antibacterial and anti-adhesive properties
<b>JAK2</b>	- involved in cytokine receptor signaling - phosphorylates STAT1-5
<b>ORMDL3</b>	- alters endoplasmic reticulum (ER)-mediated calcium homeostasis and facilitates the unfolded-protein response
<b>SP140</b>	- induced by IFNG
<b>IL10, IL19</b>	<i>IL10</i> - secreted in large amounts by regulatory T-cells (Tr1) after stimulation with CD3 and CD46 antibodies in presents of IL-2 or anti-CD28 - has anti-inflammatory properties - suppresses proliferation of bystander T helper cells <i>IL19</i> - Member of IL10 family of cytokines
<b>SMAD3</b>	- involved in response to TGFBR stimulation



gene(s)	function
<b>TYK2, ICAM1, ICAM3</b>	<i>TYK2</i> - involved in STAT1-3 and STAT5 phosphorylation and IL-23 dependent STAT4 phosphorylation - might be crucial for IL-12 dependent differentiation of Th1 cells and repression of Th2 cells <i>ICAM1, ICAM3</i> - ligand for lymphocyte function-associated antigens - provides adhesion signals in immune reactions - arrests leukocytes to blood vessels at sites of inflammation
<i>RTEL1, TNFRSF6B, SLC2A4RG</i>	<i>RTEL1, SLC2A4RG</i> -NA <i>TNFRSF6B</i> - binds to FASL and prevents apoptosis
<b>IL2RA</b>	- co-expressed with CD4 on some T regulatory cells - IL2RA (CD25)/CD4 positive T regulatory cells prevent autoimmunity - microbial activation of TLR4 mediates maturation of dendritic cells, which overcomes the suppressive effects of IL2RA(CD25)/CD4 positive T regulatory cells
<i>YDJC</i>	NA
<i>PRDX5, ESRRA</i>	<i>PRDX5</i> protective role against oxidative stress <i>ESRRA</i> - NA
<i>TNFSF11</i>	- regulator of dendritic cell function - activates anti-apoptotic serine/threonine kinase PKB
<i>TAGAP</i>	- coregulated with IL-2
<i>MTMR3</i>	NA
<b>CPEB4</b>	NA
<b>FADS1</b>	NA
<i>ZFP36L1</i>	target of IL4 signaling for B-cell survival
<b>FUT2</b>	- involved in antigen (Lewis antigens) expression in digestive organs
<i>IL27, SH2B1, EIF3C, LAT, CD19</i>	<i>IL27</i> - expressed by antigen presenting cells - triggers expansion of antigen-specific naive CD4-positive T cells and promotes polarization towards a Th1 phenotype with expression of gamma-interferon - synergistic effect with IL-12 <i>SH2B1, EIF3C</i> - NA <i>LAT</i> - may play an important role as a molecule downstream of PTKs, which are activated earliest upon T-cell stimulation <i>CD19</i> - reduces threshold for B-cell antigen receptor activation
<b>BACH2</b>	- expressed in B-cells - regulator of expression
<b>DNMT3A</b>	NA
<i>NDPIP1</i>	NA
<b>PLCL1</b>	NA
<b>ERAP2/LRAP</b>	Component of MHC class I antigen presenting pathway
<i>VAMP3</i>	- involved in phagocytosis
<b>PUS10</b>	- transduces apoptotic signal downstream of BID to mitochondria
<b>DENND1B</b>	- high expression in dendritic and natural killer cells

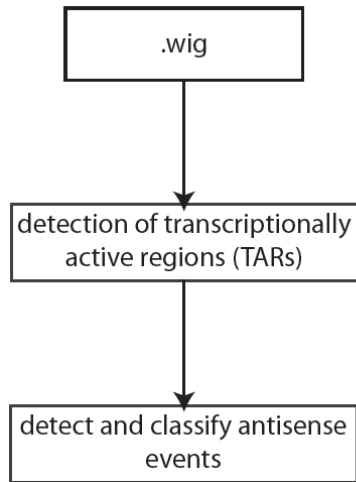
Digital only table:

**Supplementary table 3** Expression values by RNA-Seq.

**Supplementary table 4** Primer sequences used for Sanger sequencing of SNVs.

primer name	primer sequence	target gene
<b>mbBEST2fw</b>	GCAACTCAGGCAAGGGTCAG	<i>BEST2</i>
<b>mbBEST2rev</b>	CTCTGCCACAGGAGTTTGT	<i>BEST2</i>
<b>mbDN1fw</b>	AGCATGACACATGTTGTGTGA	<i>FAM190A (intron)</i>
<b>mbDN1rev</b>	TGGATGAATCTCTGGGCT	<i>FAM190A (intron)</i>
<b>mbDN2fw</b>	GACTGAAGTGGCTGCTTCTGC	<i>SYT4-SETBP1 (intergenic)</i>
<b>mbDN2rev</b>	CAAGCCCAACATCCTGAGG	<i>SYT4-SETBP1 (intergenic)</i>
<b>mbDN3fw</b>	ACGGGTTTACCAAGGGTGTAG	<i>NF1 (exonic, UTR)</i>
<b>mbDN3rev</b>	GATGGGAGTACTGGCATAACAGG	<i>NF1 (exonic, UTR)</i>
<b>mbDN4fw</b>	GAGATGGACAGGTGCTCTTTG	<i>MUC2</i>

<b>mbDN4rev</b>	AGGTCGGCTTCTGCTGCT	<i>MUC2</i>
<b>mbDN5fw</b>	CGAAGGTCTCTAAGCAGGCG	<i>MST1</i>
<b>mbDN5rev</b>	GAAAGCGGGTTTGGTCC	<i>MST1</i>
<b>mbDN6fw</b>	CTTGAAGACCTGAAGCTG	<i>PLVAP (intron)</i>
<b>mbDN6rev</b>	CAGGCTGGTCTCAAATCC	<i>PLVAP (intron)</i>
<b>mbCD19fw</b>	CATGCCACACCTCTCTCC	<i>CD19</i>
<b>mbCD19rev</b>	ATAGAGACAAAGACCCGAGAG	<i>CD19</i>
<b>mbCD19_2fw</b>	ACAGAGGAGATAACGCTGTG	<i>CD19</i>
<b>mbCD19_2rev</b>	AATGCAGAGACCCAGGA	<i>CD19</i>
<b>mbIKZF1fw</b>	GCATCCAGCAGAGAAGCA	<i>IKZF1</i>
<b>mbIKZF1rev</b>	TCCCACACACAGAGGAC	<i>IKZF1</i>
<b>mb_KCNA4fw</b>	GGTCAGAGCAATGAGGGAAG	<i>KCNA4</i>
<b>mb_KCNA4rev</b>	GAGAGGCTTGCTCACTCCAG	<i>KCNA4</i>
<b>mb_ITGAVfw</b>	TTCTCTCGGGACTCTGCTA	<i>ITGAV</i>
<b>mb_ITGAVrev</b>	CTCGAAATCAATCCCAATG	<i>ITGAV</i>
<b>mb_AURKAIP1fw</b>	GTATCCAGTCTGGCAGGAA	<i>AURKAIP1</i>
<b>mb_AURKAIP1rev</b>	GGCCCTTTACAGCACATC	<i>AURKAIP1</i>
<b>mb_FUT4fw</b>	GCCTACGGAGAGGCTCAGG	<i>FUT4</i>
<b>mb_FUT4rev</b>	GCGACTCGAAGTTCATCCA	<i>FUT4</i>
<b>mb_PNKPfw</b>	GGTCCGTAACCTGGGTCAG	<i>PNKP</i>
<b>mb_PNKPprev</b>	AGCAGTTAATGGTGGGAAA	<i>PNKP</i>
<b>mb_MMP9fw</b>	GGGAAGATGCTGTGTCA	<i>MMP9</i>
<b>mb_MMP9rev</b>	CCTCACCTCGGTACTGGAAG	<i>MMP9</i>
<b>mb_HCN4fw</b>	CTAGATGACGGGATCTGGA	<i>HCN4</i>
<b>mb_HCN4rev</b>	ACACCATCAGCTGGCGTAG	<i>HCN4</i>
<b>mb_TLL10fw</b>	GGATTGTCCAGGGTATTGGA	<i>TLL10</i>
<b>mb_TLL10rev</b>	GTGTGGCAGAGGCTCTTTCT	<i>TLL10</i>
<b>mb_C19orf6fw</b>	GGTCAGCTTCTCCTCCT	<i>C19orf6</i>
<b>mb_C19orf6rev</b>	TGTTTGTGCTCTTCGCTCTG	<i>C19orf6</i>
<b>mb_PAK6fw</b>	CCCTCTGACCACTTCGGATA	<i>PAK6</i>
<b>mb_PAK6rev</b>	CAACACCTGTGCTCACCA	<i>PAK6</i>
<b>mb_KIAA1522fw</b>	CTCAATCCTCCGACACCATT	<i>KIAA1522</i>
<b>mb_KIAA1522rev</b>	AGCTTACGCAGGGACACACT	<i>KIAA1522</i>
<b>mb_HEG1fw</b>	CGTGGTACTGTGGTGACAGAC	<i>HEG1</i>
<b>mb_HEG1rev</b>	CAACAGCTGTGCTGTGAACC	<i>HEG1</i>
<b>mb_RASA3fw</b>	CACCGAACGCCTTATTGTT	<i>RASA3</i>
<b>mb_RASA3rev</b>	GCTGTCTCGAGTGATCTCC	<i>RASA3</i>
<b>mb_WNT9Bfw</b>	GAGGACTCACCCAGCTTCTG	<i>WNT9B</i>
<b>mb_WNT9Brev</b>	TAGGCCTAGTGCTTGCAAGGT	<i>WNT9B</i>
<b>mb_LDOC1Lfw</b>	CTCTCTGCACAGCCCTATT	<i>LDOC1L</i>
<b>mb_LDOC1Lrev</b>	GTGTGGCCTTCCTGTGTCT	<i>LDOC1L</i>
<b>mb_SPZ1fw</b>	CAGGAAGCAGAACAACACTGGAG	<i>SPZ1</i>
<b>mb_SPZ1rev</b>	CATGCTTAGCCGACACTTCA	<i>SPZ1</i>
<b>mb_C2CD4Afw</b>	AGCAGAGAGTGCCAGATGT	<i>C2CD4A</i>
<b>mb_C2CD4Arev</b>	GGATGCAGAACTCAGGGATG	<i>C2CD4A</i>
<b>mb_PTRFfw</b>	GCCTTCTGAAGTCGTCCAC	<i>PTRF</i>
<b>mb_PTRFrev</b>	CCAAACTGAGCATCAGCAA	<i>PTRF</i>
<b>mb_SPTBN4fw</b>	TGCAGCTGCTCAAGAAACAC	<i>SPTBN4</i>
<b>mb_SPTBN4rev</b>	TGGACCATTCCTGTGAAGT	<i>SPTBN4</i>
<b>mb_IRF5fw</b>	ACAGTGACCGACCTGGAGAT	<i>IRF5</i>
<b>mb_IRF5rev</b>	CTTGCACTGGGGATGTCT	<i>IRF5</i>
<b>mb_KLC3fw</b>	ATAGTCCAGGGGCCAGTTA	<i>KLC3</i>
<b>mb_KLC3rev</b>	GCGTACTGGATCACGAGGTT	<i>KLC3</i>
<b>mb_OGFOD2fw</b>	GTTGGAACCTTGCTGCTGG	<i>OGFOD2</i>
<b>mb_OGFOD2rev</b>	GTAGGCTGCGGGGATTTTAT	<i>OGFOD2</i>
<b>mb_NOS3fw</b>	CAGCTCGCCTCTAGCTCCTA	<i>NOS3</i>
<b>mb_NOS3rev</b>	ACCTCCAGTCTTTCACACG	<i>NOS3</i>
<b>mb_NUMBfw</b>	ACAGGCTGAGAGGTGAGGAA	<i>NUMB</i>
<b>mb_NUMBrev</b>	CGACCGATGGTTAGAAGAGG	<i>NUMB</i>
<b>mb_IL18BPfw</b>	CTGCACAGCACTTCTCTC	<i>IL18BP</i>
<b>mb_IL18BPprev</b>	CTACACCTCCTGTCCCAAG	<i>IL18BP</i>



1. Find transcriptional active regions (TARs)

2. Divide TARs in known (exon) and novel

3. Classify TARs. If one TAR falls within several classes, the highest class (i.e. **AS1a**) will be selected.

**- antisense type I (AS1)**

Transcription of known TAR overlapping either: a known TAR (**AS1a**), a novel intragenic TAR (**AS1b**), or novel TAR in gene proximity (<=10kb) (**AS1c**) or a novel intergenic element (**AS1d**) on the opposite strand.

**- antisense type II (AS2)**

Transcription of novel intragenic TAR, types a-d as above.

**- antisense type III (AS3)**

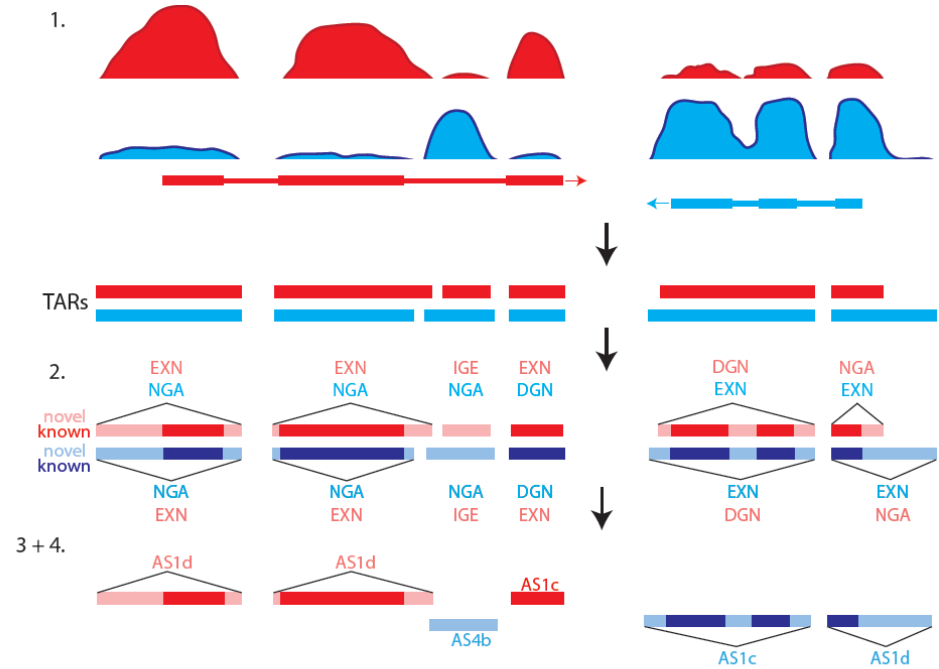
Transcription of novel TAR in 10kb proximity of a known TAR, types a-d as above.

**- antisense type IV (AS4)**

Transcription of novel TAR, types a-d as above.

4. Calculate antisense coverage to sense coverage ratio. Estimate difference to noise (calculate z-score). Reject events with z-score < 1.96.

$$z\text{-score for event with antisense:sense ratio } x = z = \frac{x - \mu}{\sigma}$$

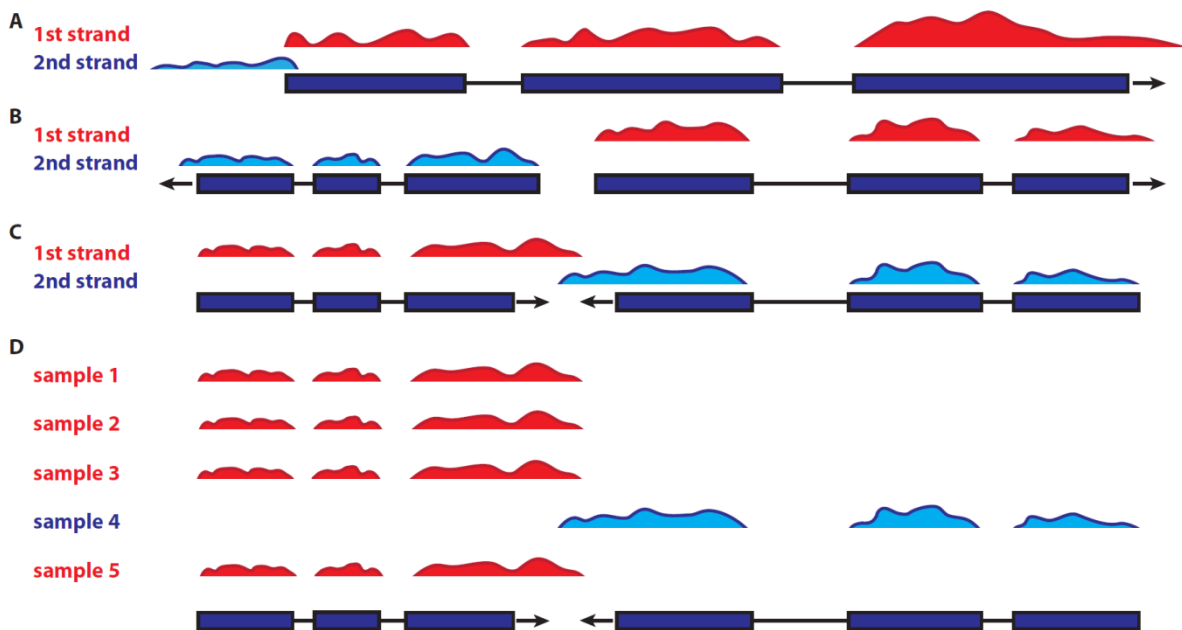


class	type(s)
exonic	EXN, ELU, ELD, DGE, IGE, RIN
intragenic	IGE
neighbourhood	UGN, DGN
intergenic	NGA

EXN = exonic  
 UGE/DGE = upstream/downstream gene extension  
 ELU/ELD = exon linked upstream/downstream  
 UGN/DGN = upstream/downstream gene neighbourhood  
 NGA = not gene associated

**Supplementary figure 1 Detection of antisense events.**

1. The wig files for the forward and reverse strand are searched for continuous blocks of coverage (transcriptionally active regions = TARs) 2. Each TAR is assigned a category, depending on its location relative to annotated transcripts. 3. Antisense events are classified by the category of the sense and antisense TAR. For a better overview of different S/AS pair classes see Figure X. 4. The antisense to sense coverage ratio is calculated and compared to the antisense noise level, which has been calculated before. If the score does not pass the threshold (z-score >= 1.96), the antisense event is rejected.



**Supplementary figure 2 Examples of possible sense/antisense pairs.**

**A-C** Schematic representation of sense antisense pairs which often show positively correlated expression patterns. **A** An short antisense transcript is located in the promoter region of the sense transcript. **B** S/AS pair in head-to-head conformation. **C** S/AS pair in tail-to-tail conformation. **D** S/AS pair showing mutually exclusive expression of one transcript.

**A)**

**maternal**

ATCGTGCAGAGCTAGCTAGCGATC

TAGCACGCTCGATCGATCGCTAG

**paternal**

ATCGTGCAGAGTTAGCTAGCGATC

TAGCACGCTCAATCGATCGCTAG

**B)**

	A	C	G	T
forward reads	-	12	-	5
reverse reads	6	-	15	-

**C)**

	A	C	G	T
forward reads	-	31	-	-
reverse reads	12	-	-	-

**Supplementary figure 3 Assessment of chromosome and allele specific expression.**

At positions in which the chromosomes inherited by the mother and father differ (**A**) it is possible to determine which chromosome shows transcriptional activity. At these positions it is possible, that there is equal expression on either or both chromosomes, which results in equal amounts of reads for both alleles on both strands (**B**), or it could be possible, that the sense and antisense events are transcribed mutually exclusive from either chromosome, in which case all forward and reverse reads would show only one allele respectively (**C**).

**Supplementary table 5 Summary of mapping statistics and raw data output for son.**

		MP genome	PE1 genome	PE2 genome	PE3 genome	PE exome	PE trans	PE trans biopsy
<b>raw data</b>	forward tags [N]:	445951311	456920738	453283622	629127605	100324928	74271146	75535115
>	forward tags [Gb]:	22.2976	22.846	22.6642	31.4564	5.0162	3.7136	3.7768
>	reverse tags [N]:	445843429	452042578	448523261	629127605	100324928	52832267	73794354
>	reverse tags [Gb]:	22.2922	11.3011	11.2131	22.0195	3.5114	1.3208	1.8449
>	total tags [N]:	891794740	908963316	901806883	1258255210	200649856	127103413	149329469
>	total tags [Gb]:	44.5897	34.1471	33.8773	53.4758	8.5276	5.0344	5.6216
>	x coverage.:	14.8632	11.3824	11.2924	17.8253	2.8425	1.6781	1.8739
<b>mapped data</b>	mapped forward tags [N]:	350124057	323635646	327569904	489808400	73743689	49667174	39726141
>	mapped forward tags [Gb]:	17.5062	16.1818	16.3785	24.4904	3.6872	2.4834	1.9863
>	% of total fw-tags:	78.51	70.83	72.27	77.86	73.5	66.87	52.59
>	uniquely mapped forward tags [N]:	292020440	254222360	258836268	397984051	61847298	39070035	29720947
>	uniquely mapped forward tags [Gb]:	14.601	12.7111	12.9418	19.8992	3.0924	1.9535	1.486
>	% of mapped fw-tags:	83.4	78.55	79.02	81.25	83.87	78.66	74.81
>	number of start points for uniquely fw-mapped tags:	246193228	228435757	238773079	336325047	35972149	8513193	5900407
>	mapped reverse tags [N]:	352567333	176851507	186564467	152977299	35573173	10074259	11723769
>	mapped reverse tags [Gb]:	17.6284	4.4213	4.6641	5.3542	1.2451	0.2519	0.2931
>	% of total rev-tags:	79.08	39.12	41.6	24.32	35.46	19.07	15.89
>	uniquely mapped reverse tags [N]:	296136111	114290096	122811022	118884529	29327592	7108853	7740198
>	uniquely mapped reverse tags [Gb]:	14.8068	2.8573	3.0703	4.1610	1.0265	0.1777	0.1935
>	% of mapped rev-tags:	83.99	64.62	65.83	77.71	82.44	70.56	66.02
>	number of start points for uniquely mapped rev-tags:	254401952	105249172	114749501	106300687	18011844	2529329	2336671
>	total mapped tags [N]:	702691390	500487153	514134371	642785699	109316862	59741433	51449910
>	total mapped tags [Gb]:	35.1346	20.6031	21.0426	29.8446	4.9322	2.7352	2.2794
>	x coverage:	11.7115	6.8677	7.0142	9.9482	1.6441	0.9117	0.7598
<b>mapped pairs</b>	mapped pairs [N]:	403060374	352699284	363680372	499483939	75827547	51609455	43275253
>	mapped pairs [Gb]:	40.306	26.4524	27.276	42.4561	6.4453	3.8707	3.2456
>	x coverage [Gb]:	13.4353	8.8175	9.092	14.152	2.1484	1.2902	1.0819
>	non redundant pairs [N]:	376553573	344691280	357961748	472851886	65767936	51609455	43275253
>	non redundant pairs [Gb]:	37.6554	25.8518	26.8471	40.1924	5.5903	3.8707	3.2456
>	x coverage [Gb]:	12.5518	8.6173	8.949	13.3975	1.8634	1.2902	1.0819
>	% of mapped pairs:	93.42	97.73	98.43	94.67	86.73	100	100
>	unique AAA pairs [N]:	248571059	141781403	147201288	260881994	48378547	9178803	8434136
>	unique AAA pairs [Gb]:	24.8571	10.6336	11.0401	22.175	4.1122	0.6884	0.6326
>	x coverage [Gb]:	8.2857	3.5445	3.68	7.3917	1.3707	0.2295	0.2109
>	% of mapped pairs:	61.67	40.2	40.48	52.23	63.8	17.79	19.49
>	non redundant AAA pairs [N]:	223493788	135855494	143378507	237472494	40363516	9178803	8434136
>	non redundant AAA pairs [Gb]:	22.3494	10.1892	10.7534	20.1852	3.4309	0.6884	0.6326
>	x coverage [Gb]:	7.4498	3.3964	3.5845	6.7284	1.1436	0.2295	0.2109
>	% of mapped pairs:	55.45	38.52	39.42	47.54	53.23	17.79	19.49

MP = mate-pair, PE = paired-end, exome = agilent sureselect human all exon v1, trans = transcriptome (WTAK)

**Supplementary table 6 Summary of mapping statistics and raw data output for mother.**

		MP genome	PE1 genome	PE2 genome	PE3 genome	PE exome	PE trans
<b>raw data</b>	forward tags [N]:	458149091	473526191	472770316	523374482	102944572	66924522
>	forward tags [Gb]:	22.9075	23.6763	23.6385	26.1687	5.1472	3.3462
>	reverse tags [N]:	462199362	465875291	457813943	523374482	102944572	64533416
>	reverse tags [Gb]:	23.11	11.6469	11.4453	18.3181	3.6031	1.6133
>	total tags [N]:	920348453	939401482	930584259	1046748964	205889144	131457938
>	total tags [Gb]:	46.0174	35.3232	35.0839	44.4868	8.7503	4.9596
>	x coverage.:	15.3391	11.7744	11.6946	14.8289	2.9168	1.6532
<b>mapped data</b>	mapped forward tags [N]:	363842610	338137021	261876111	393970329	76588175	38924863
>	mapped forward tags [Gb]:	18.1921	16.9069	13.0938	19.6985	3.8294	1.9462
>	% of total fw-tags:	79.42	71.41	55.39	75.28	74.4	58.16
>	uniquely mapped forward tags [N]:	305403784	264501880	202670080	307624797	64306683	29854242
>	uniquely mapped forward tags [Gb]:	15.2702	13.2251	10.1335	15.3812	3.2153	1.4927
>	% of mapped fw-tags:	83.94	78.22	77.39	78.08	83.96	76.7
>	number of start points for uniquely fw-mapped tags:	259492160	240830477	185246492	264184978	38222279	4386661
>	mapped reverse tags [N]:	363384570	195026814	187937260	130802678	35213137	10012262
>	mapped reverse tags [Gb]:	18.1692	4.8757	4.6984	4.5781	1.2325	0.2503
>	% of total rev-tags:	78.62	41.86	41.05	24.99	34.21	15.51
>	uniquely mapped reverse tags [N]:	309028158	125945636	119893972	96523820	29038544	6656923
>	uniquely mapped reverse tags [Gb]:	15.4514	3.1486	2.9973	3.3783	1.0163	0.1664
>	% of mapped rev-tags:	85.04	64.58	63.79	73.79	82.47	66.49
>	number of start points for uniquely mapped rev-tags:	266971231	116910476	109576277	86032711	18418018	1910206
>	total mapped tags [N]:	727227180	533163835	449813371	524773007	111801312	48937125
>	total mapped tags [Gb]:	36.3614	21.7825	17.7922	24.2766	5.0619	2.1965
>	x coverage:	12.1205	7.2608	5.9307	8.0922	1.6873	0.7322
<b>mapped pairs</b>	mapped pairs [N]:	416112766	374416125	335080328	406170232	78481279	41203376
>	mapped pairs [Gb]:	41.6113	28.0812	25.131	34.5245	6.6709	3.0903
>	x coverage [Gb]:	13.8704	9.3604	8.377	11.5082	2.2236	1.0301
>	non redundant pairs [N]:	390720370	367976506	327688502	389301693	69938773	41203376
>	non redundant pairs [Gb]:	39.072	27.5982	24.5766	33.0906	5.9448	3.0903
>	x coverage [Gb]:	13.024	9.1994	8.1922	11.0302	1.9816	1.0301
>	% of mapped pairs:	93.9	98.28	97.79	95.85	89.12	100
>	unique AAA pairs [N]:	266799621	150484729	138515823	194874028	48595950	8792113
>	unique AAA pairs [Gb]:	26.68	11.2864	10.3887	16.5643	4.1307	0.6594
>	x coverage [Gb]:	8.8933	3.7621	3.4629	5.5214	1.3769	0.2198
>	% of mapped pairs:	64.12	40.19	41.34	47.98	61.92	21.34
>	non redundant AAA pairs [N]:	242663400	145707963	132432275	179946092	41779409	8792113
>	non redundant AAA pairs [Gb]:	24.2663	10.9281	9.9324	15.2954	3.5512	0.6594
>	x coverage [Gb]:	8.0888	3.6427	3.3108	5.0985	1.1837	0.2198
>	% of mapped pairs:	58.32	38.92	39.52	44.3	53.23	21.34

MP = mate-pair, PE = paired-end, exome = agilent suresselect human all exon v1, trans = transcriptome (WTAK)

**Supplementary table 7 Summary of mapping statistics and raw data output for father.**

		MP genome	PE1 genome	PE2 genome	PE3 genome	PE exome	PE trans
<b>raw data</b>	forward tags [N]:	483587949	455297734	545122442	441176353	105096625	77256598
>	forward tags [Gb]:	24.1794	22.7649	27.2561	22.0588	5.2548	3.8628
>	reverse tags [N]:	485942401	444781261	528971159	441244131	105096625	73769852
>	reverse tags [Gb]:	24.2971	11.1195	13.2243	15.4435	3.6784	1.8442
>	total tags [N]:	969530350	900078995	1074093601	882420484	210193250	151026450
>	total tags [Gb]:	48.4765	33.8844	40.4804	37.5024	8.9332	5.7071
>	x coverage.:	16.1588	11.2948	13.4935	12.5008	2.9777	1.9024
<b>mapped data</b>	mapped forward tags [N]:	344949116	337390407	415336167	334887372	79862683	49646837
>	mapped forward tags [Gb]:	17.2475	16.8695	20.7668	16.7444	3.9931	2.4823
>	% of total fw-tags:	71.33	74.1	76.19	75.91	75.99	64.26
>	uniquely mapped forward tags [N]:	287534155	260856808	333044788	267174954	67176631	38274337
>	uniquely mapped forward tags [Gb]:	14.3767	13.0428	16.6522	13.3587	3.3588	1.9137
>	% of mapped fw-tags:	83.36	77.32	80.19	79.78	84.12	77.09
>	number of start points for uniquely fw-mapped tags:	239941712	237512358	295760020	230668244	38049140	5916853
>	mapped reverse tags [N]:	364283204	210616370	253432829	200553150	39443555	15479389
>	mapped reverse tags [Gb]:	18.2142	5.2654	6.3358	7.0194	1.3805	0.387
>	% of total rev-tags:	74.96	47.35	47.91	45.45	37.53	20.98
>	uniquely mapped reverse tags [N]:	309041756	139570603	175273321	161309752	32708879	10878470
>	uniquely mapped reverse tags [Gb]:	15.4521	3.4893	4.3818	5.6458	1.1448	0.272
>	% of mapped rev-tags:	84.84	66.27	69.16	80.43	82.93	70.28
>	number of start points for uniquely mapped rev-tags:	259950464	129140741	159162927	142948390	19672063	2527013
>	total mapped tags [N]:	709232320	548006777	668768996	535440522	119306238	65126226
>	total mapped tags [Gb]:	35.4616	22.1349	27.1026	23.7637	5.3737	2.8693
>	x coverage:	11.8205	7.3783	9.0342	7.9212	1.7912	0.9564
<b>mapped pairs</b>	mapped pairs [N]:	414305583	370564765	446948994	355043327	82765958	53095185
>	mapped pairs [Gb]:	41.4306	27.7924	33.5212	30.1787	7.0351	3.9821
>	x coverage [Gb]:	13.8102	9.2641	11.1737	10.0596	2.345	1.3274
>	non redundant pairs [N]:	385776682	363128624	431857388	334450104	73154660	53095185
>	non redundant pairs [Gb]:	38.5777	27.2346	32.3893	28.4283	6.2181	3.9821
>	x coverage [Gb]:	12.8592	9.0782	10.7964	9.4761	2.0727	1.3274
>	% of mapped pairs:	93.11	97.99	96.62	94.2	88.39	100
>	unique AAA pairs [N]:	249184807	161643340	213783589	211810054	51714029	13031279
>	unique AAA pairs [Gb]:	24.9185	12.1233	16.0338	18.0039	4.3957	0.9773
>	x coverage [Gb]:	8.3062	4.0411	5.3446	6.0013	1.4652	0.3258
>	% of mapped pairs:	60.15	43.62	47.83	59.66	62.48	24.54
>	non redundant AAA pairs [N]:	221949584	156062383	201783825	193539561	44118023	13031279
>	non redundant AAA pairs [Gb]:	22.195	11.7047	15.1338	16.4509	3.75	0.9773
>	x coverage [Gb]:	7.3983	3.9016	5.0446	5.4836	1.25	0.3258
>	% of mapped pairs:	53.57	42.11	45.15	54.51	53.3	24.54

MP = mate-pair, PE = paired-end, exome = agilent sureselect human all exon v1, trans = transcriptome (WTAK)

**Supplementary table 8 Summary of achieved coverage.**

covered	genomic coverage (excluding Ns)						quantiles			average coverage per coverable base
	0x	1x	3x	5x	10x	20x	25%	50%	75%	
mother	0.46%	99.54%	98.96%	98.33%	95.85%	83.83%	39x	31x	23x	34.72
father	0.12%	99.88%	99.55%	99.10%	97.34%	89.58%	43x	35x	27x	40.28
son	0.13%	99.87%	99.48%	98.92%	96.51%	84.81%	42x	33x	24x	36.57

covered	coverage of refSeq annotated transcripts (excluding Ns)						quantiles			average coverage per coverable base
	0x	1x	3x	5x	10x	20x	25%	50%	75%	
mother	0.98%	99.02%	97.18%	95.80%	92.81%	84.31%	84x	40x	26x	66.23
father	0.72%	99.28%	97.54%	96.10%	92.78%	83.68%	89x	41x	26x	69.84
son	0.64%	99.36%	97.88%	96.64%	93.80%	85.83%	88x	43x	28x	68.58



**Supplementary table 9 SNV calling statistics.**

SNP-type/location	MOTHER		FATHER		SON	
	known	novel	known	novel	known	novel
<b>rsID SNVs:</b>	2995470	-	3055342	-	3027916	-
<b>novel SNVs:</b>	-	312818	-	319877	-	315156
<b>Venter+Watson SNVs:</b>	17	-	20	-	20	-
<b>Watson SNVs:</b>	1263	-	1377	-	1396	-
<b>Venter SNVs:</b>	1933	-	1883	-	2071	-
<b>gene associated SNVs:</b>	1173779	112517	1199115	113088	1194183	112535
<b>UTR-SNVs (intron):</b>	210478	21000	213777	21553	211519	21064
<b>UTR-SNVs (exon):</b>	26517	2344	26711	2331	27140	2355
<b>neighbourhood-SNVs:</b>	62072	6490	63147	6477	62915	6522
<b>coding SNV:</b>	13804	1367	13853	1403	14093	1452
<b>intronic SNV:</b>	860908	81316	881627	81324	878516	81142
<b>homozygous sense:</b>	7396	470	7422	508	7567	498
<b>heterozygous missense:</b>	2571	736	2583	734	2559	803
<b>homozygous missense:</b>	3212	80	3221	73	3321	71
<b>splice-junction (intron):</b>	157	34	155	28	159	41
<b>splice-junction (exon, heterozygous):</b>	20	2	27	3	19	5
<b>splice-junction (exon, homozygous):</b>	13	0	16	1	15	1
<b>read-through (heterozygous):</b>	10	3	7	2	9	4
<b>read-through (homozygous):</b>	12	0	14	1	17	0
<b>altered M (heterozygous):</b>	115	28	120	16	112	19
<b>altered M (homozygous):</b>	161	2	162	1	161	3
<b>nonsense (heterozygous):</b>	13	19	12	25	23	23
<b>nonsense (homozygous):</b>	15	4	15	0	10	2
<b>new M (heterozygous):</b>	144	22	123	37	152	22
<b>new M (homozygous):</b>	122	1	131	2	128	1

New M = base changes leading to a new methionine which might represent a new start-site for transcription; Altered M = changing a methionine into another amino acid, which could possibly remove an (alternative) start codon.

**Supplementary table 10 Bioscope SNV correlation with Illumina iScan 1M human Omniquad chip.**

Mother	SNV calls only				SNV calls + consensus calls*				
	2x	5x	10x	20x	1x	2x	5x	10x	20x
concordant	412156	406207	363984	149562	977270	975848	962268	888112	478435
unconcordant	3164	2394	1100	235	25679	25268	21904	14072	5831
concordant	99.24%	99.41%	99.70%	99.84%	97.44%	97.48%	97.77%	98.44%	98.80%
unconcordant	0.76%	0.59%	0.30%	0.16%	2.56%	2.52%	2.23%	1.56%	1.20%

Father	SNV calls only				SNV calls + consensus calls*				
	2x	5x	10x	20x	1x	2x	5x	10x	20x
concordant	418085	413801	385052	205650	984279	982935	971955	917402	587598
unconcordant	2658	2066	1130	290	20627	20312	17815	12648	6199
concordant	99.37%	99.50%	99.71%	99.86%	97.95%	97.98%	98.20%	98.64%	98.96%
unconcordant	0.63%	0.50%	0.29%	0.14%	2.05%	2.02%	1.80%	1.36%	1.04%

Son	SNV calls only				SNV calls + consensus calls*				
	2x	5x	10x	20x	1x	2x	5x	10x	20x
concordant	415218	409789	374426	180245	982764	981547	969291	905022	545816
unconcordant	2594	1955	1001	236	22243	21939	19336	13276	5895
concordant	99.38%	99.53%	99.73%	99.87%	97.79%	97.81%	98.04%	98.55%	98.93%
unconcordant	0.23%	0.47%	0.27%	0.13%	2.21%	2.19%	1.96%	1.45%	1.07%

\*includes consensus calls for all positions present on the Illumina chip where there was no SNV called by bioscope (reference call)

**Supplementary table 11 small structural variation (sSV) calling statistics.**

	mother		father		child	
	known	novel	known	novel	known	novel
<b>total sSVs detected:</b>	145,477	48,722	170,158	49,395	149,450	43,328
<b>no gene association:</b>	88,918	31,585	104,652	32,332	91,224	28,300
<b>5'-neighbourhood:</b>	1,244	361	1,378	352	1,253	296
<b>3'-neighbourhood:</b>	1,596	464	1,852	392	1,654	375
<b>5'-UTR (intronic):</b>	6,394	2,017	7,405	2,016	6,527	1,691
<b>3'-UTR (intronic):</b>	1,252	439	1,456	414	1,236	385
<b>5'-UTR (exonic):</b>	189	48	203	48	199	46
<b>3'-UTR (exonic):</b>	1,395	298	1,623	336	1,407	296
<b>intronic:</b>	44,269	13,417	51,346	13,417	45,708	11,860
<b>splice-junction (intronic):</b>	13	3	18	6	20	7
<b>splice-junction (exonic):</b>	11	7	12	1	10	4
<b>exonic:</b>	196	83	213	81	212	68

**Supplementary table 12 large structural variation (ISV) calling statistics.**

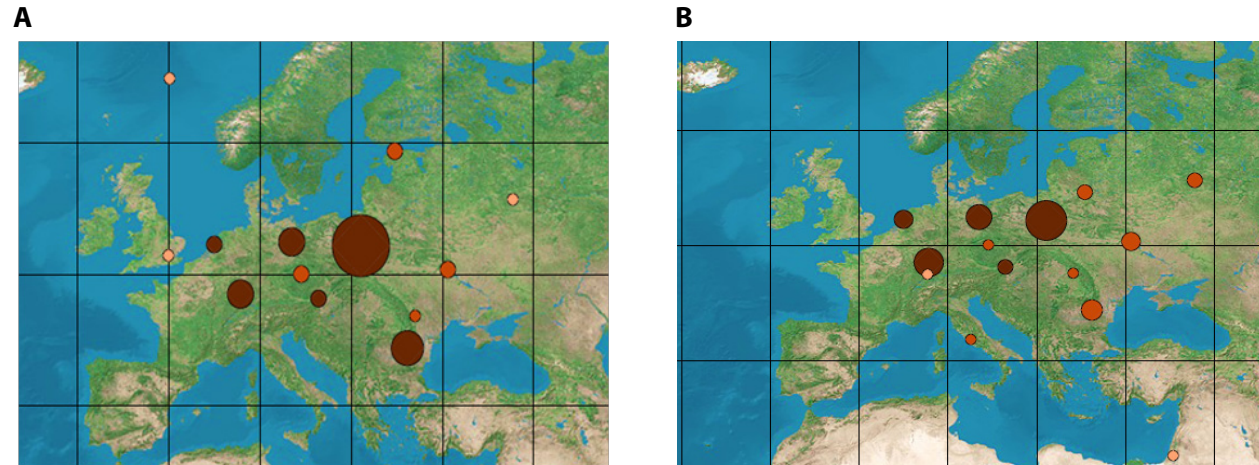
		mother	father	child
<b>Bioscope</b>	total ISVs detected [N]	5099	5059	4693
	heterozygous variants [N]	1827	1593	1062
	homozygous variants [N]	3272	3466	3631
	insertions [N]	1716	1832	814
	deletions [N]	3158	2979	3817
	double* [N]	225	248	62
	large indels >= 3000 bp	1349	1421	1351
	deletions >= 3000 bp	1271	1351	1319
<b>Korbel</b>	total deletions	15892	22763	113104
	deletions >= 3000 bp	3273	3491	6273
deletions >= 3000 bp, overlapping ANY Korbel ISV				1185

\* two ISVs at the same location.

Digital only tables:

**Supplementary table 13** Manually reviewed ISVs.

**Supplementary table 14** Sanger sequencing of potential *de novo* SNVs.



**Supplementary figure 4** Origin determination by mitochondrial variations based on Röhl *et al.* 2001.

Both parents have a „K“ mitotype, significant predominance of Ashkenazi matches, based on a comparison with > 40,000 mt DNA sequences. **A** father, **B** mother

**Supplementary table 15 Overview of noncoding SNVs detected in the child in genes associated with monogenic phenocopies of Crohn disease.**

trait	gene symbol	5'-neighbourhood	3'-neighbourhood	5'-UTR (intron)	5'-UTR (exon)	3'-UTR (exon)	intron
<b>Blau syndrome</b>	<i>NOD2</i> ■	1	2	0	1	0	24 (3)
<b>C1 esterase inhibitor deficiency</b>	<i>SERPING1</i>	1	1 (1)	0	0	0	6
<b>CGD</b>	<i>NCF2</i>	11 (1)	1	2	0	2	39 (1)
<b>CGD</b>	<i>NCF1</i>	0	0	0	0	0	0
<b>CGD</b>	<i>CYBA</i>	3	5 (1)	0	0	1	8
<b>CVID</b>	<i>ICOS</i>	3	3	0	0	3	24
<b>CVID</b>	<i>CD19</i> ■	3 (2)	2 (1)	0	0	0	2 (2)
<b>CVID</b>	<i>TNFRSF13B</i>	4	11	0	0	3 (2)	51 (2)
<b>CVID</b>	<i>TNFRSF13C</i>	2 (1)	0	0	0	1	0
<b>Hyper-IgM syndrome</b>	<i>CD40LG</i>	0	0	0	0	0	1 (1)
<b>IPEX</b>	<i>FOXP3</i>	0	1	0	0	0	1
<b>Wiskott-Aldrich syndrome</b>	<i>WAS</i>	0	0	0	0	0	2
<b>X-linked Agammaglobulinemia</b>	<i>BTK</i>	0	0	1	0	1	3

Numbers in brackets represent fraction of novel SNVs. ■ genes also associated with Crohn's Disease

**Supplementary table 16 Nonsynonymous SNVs detected in the child in genes associated with monogenic phenocopies of Crohn's disease.**

gene symbol	rsID	chrom	position	phyloP	type	zygosity	ref allele	allele A	allele B	reads ref	reads alt	freqA	freqC	freqG	freqT	SIFT	SIFT score	SIFT AA	PolyPhen	dScore
<i>NCF2</i>	rs2274064	chr1	181809010	-0.337	missense	hom	T	C	C	0	29	-	0.514	-	0.486	tolerated	1.00	K181R	benign	-0.089
<i>NCF1</i>	rs62475423	chr7	73831604	-1.034	missense	hom	G	A	A	0	3	na	na	na	na	tolerated	0.84	G99S	benign	-0.283
<i>CD19</i> ■*†	-	chr16	28851412	3.013	missense	het	G	T	G	2	2	na	na	na	na	<b>damaging</b> (low confidence)	0.00	W111C	<b>probably damaging</b>	3.083
<i>CD19</i> ■	rs2904880	chr16	28851897	0.938	missense	het	C	G	C	11	12	-	0.347	0.653	-	tolerated	0.17	L174V	benign	-0.054
<i>NOD2</i> ■	rs2066842	chr16	49302125	-1.986	missense	het	C	T	C	11	10	-	0.683	-	0.317	tolerated	0.26	P268S	benign	0.74
<i>NOD2</i> ■*†	rs2066845	chr16	49314041	3.304	missense	het	G	G	C	25	9	na	na	na	na	<b>damaging</b>	0	G908R	<b>probably damaging</b>	-3.854
<i>CYBA</i> ♦	rs4673	chr16	87240737	1.009	missense	hom	A	G	G	0	28	-	0.750	-	0.250	tolerated	0.56	Y72H	<b>probably damaging</b>	-3.516
<i>TNFRSF13B</i> *†	rs34562254	chr17	16783716	2.027	missense	het	G	G	A	7	6	-	0.903	-	0.097	tolerated	0.52	P251L	<b>probably damaging</b>	-3.265
<i>IL10RA</i>	rs2229113	chr11	117374880	0.451	missense	hom	A	G	G	0	41	0.392	-	0.608	-	tolerated	0.31	R351G	benign	-
<i>XIAP</i>	rs5956583	chrX	122862192	2.373	missense	hom	A	C	C	0	39	na	na	na	na	tolerated	0.15	Q423P	benign	-3.788

■ genes also associated with Crohn's Disease, \*† conserved region (phyloP >2.0), ♦ follows the recessive model. 1000G CEU frequencies taken from NCBI (these can be +-strand or -strand)

**Supplementary table 17 Noncoding SNVs detected in the child in conserved regions in genes associated monogenic phenocopies of Crohn disease.**

gene symbol	rsID	chrom	position	phyloP	location	zygosity	ref allele	allele A	allele B	reads ref	reads alt
<b>CD19</b> ■	-	chr16	28851705	2.352	intron	het	C	G	C	12	6
<b>SERPING1</b>	rs1005510	chr11	57123798	2.347	intron	hom	C	T	T	0	26
<b>NCF2</b>	rs3818364	chr1	181809370	2.191	intron	hom	G	T	T	0	13

Includes only SNVs with phyloP score  $\geq 2.0$ . ■ genes also associated with Crohn's Disease

**Supplementary table 18 small structural variations detected in the child in genes associated with monogenic phenocopies of Crohn disease.**

gene symbol	rsID	chrom	start	end	type	reference	alternative allele	zygosity	location
<b>NCF2</b>	rs10641967	chr1	181790126	181790126	insertion site	-	ATG	homozygous	3'-neighbourhood
<b>NCF2</b>	rs67311370	chr1	181806839	181806839	deletion	a	-	hemizygous	intronic
<b>ICOS</b>	rs10578138	chr2	204518019	204518020	deletion	TG	-	hemizygous	intronic
<b>ICOS</b>	rs56148633	chr2	204523890	204523890	deletion	t	-	hemizygous	intronic
<b>ICOS</b>	-	chr2	204527663	204527664	deletion	AGA	C	hemizygous	intronic
<b>ICOS</b>	rs67643841	chr2	204531749	204531749	insertion site	-	G	hemizygous	intronic

**Supplementary table 19 Nonsynonymous SNVs predicted to be damaging associated with genes in proximity of GWAS Crohn's disease SNVs.**

Two nonsense SNVs were detected in *FUT2* and *IKZF1*, which are located in conserved region (phyloP 2.791 and 3.308 respectively). The *IKZF1* SNV could not be replicated by Sanger and the *FUT2* SNV was confirmed even though the father is also heterozygous for this variant. SNVs in conserved regions were also detected in *CD19*, *MST1* (both not replicated by Sanger), *NOD2* and *GPX1*. One SNV following the recessive model of inheritance was detected in a site of acceleration in *SNAPC4* (phyloP -3.617).

rsID	gene symbol	chrom	position	phyloP	type	zygosity	ref allele	allele A	allele B	reads ref	reads alt	freqA	freqC	freqG	freqT	SIFT	SIFT score	SIFT AA	PolyPhen	dScore
<i>APOB48R</i>	rs40832	chr16	28416217	-1.271	missense	hom	T	C	C	0	12	-	1.000	-	-	not scored	NA	NA	possibly damaging	-3.166
<i>ATG16L1</i> *	rs2241880	chr2	233848107	-2.150	missense	het	A	G	A	22	20	-	0.528	-	0.472	tolerated	0.51	T204A	possibly damaging (uc002vtz.1)	-2.456
<i>BSN</i>	rs34762726	chr3	49664214	1.495	missense	hom	G	A	A	0	6	0.222	-	0.778	-	not scored	NA	A741T	possibly damaging	-2.8
<i>C1orf106</i>	rs296520	chr1	199147601	0.448	missense	hom	C	T	T	0	3	na	na	na	na	damaging (low confidence)	0.03	R538C	benign	-1.957
<i>CD19</i> *	-	chr16	28851412	3.013	missense	het	G	T	G	2	2	na	na	na	na	damaging (low confidence)	0	W111C	probably damaging	-4.544
<i>FUT2</i> *	rs601338	chr19	53898486	2.791	nonsense	het	G	G	A	13	18	na	na	na	na	-	-	W154*	-	-
<i>GALC</i>	rs421262	chr14	87470966	-0.894	missense	hom	T	C	C	0	71	-	0.992	-	0.008	tolerated	0.72	T625A	possibly damaging	-2.972
<i>GCKR</i>	rs1260326	chr2	27584444	0.550	missense	hom	T	C	C	0	68	-	0.556	-	0.444	tolerated	0.28	L446P	probably damaging	-3.325
<i>GPX1</i> *	rs11549656	chr3	49369838	2.870	missense	hom	G	A	A	0	12	na	na	na	na	not scored	NA	P200L	-	-
<i>GSDMB</i>	rs2305479	chr17	35315743	-1.142	missense	hom	C	T	T	0	56	0.472	-	0.528	-	damaging (low confidence)	0	G291R	probably damaging	-3.57
<i>IKZF1</i> *	-	chr7	50337822	3.308	nonsense	het	G	T	G	24	6	na	na	na	na	-	-	E29*	-	-
<i>IL27</i>	rs181206	chr16	28420904	1.956	missense	het	A	G	A	4	7	-	0.333	-	0.667	tolerated	0.09	L119P	possibly damaging	-3.708
<i>LRRK2</i>	-	chr12	38915756	1.354	missense	het	A	G	A	35	8	na	na	na	na	damaging	0.02	K137E	benign	-2.929
<i>MST1</i> *	rs71324987	chr3	49698754	2.623	missense	het	A	G	A	7	3	na	na	na	na	damaging	0	C298R	probably damaging	-3.975
<i>NOD2</i> *	rs2066845	chr16	49314041	3.304	missense	het	G	G	C	25	9	na	na	na	na	damaging	0	G908R	probably damaging	-3.854
<i>RASIP1</i>	rs2287922	chr19	53924038	1.266	missense	hom	G	A	A	1	2	0.592	-	0.408	-	damaging	0	R601C	probably damaging	-4.898
<i>RTEL1</i>	-	chr20	61791918	1.072	missense	het	T	T	C	15	6	na	na	na	na	tolerated	0.15	S726P	probably damaging	-3.989
<i>SNAPC4</i> ♦	known to 1000G	chr9	138391866	-3.617	missense	hom	C	A	A	0	11	na	na	na	na	damaging (low confidence)	0	A1412S	benign	-2.853

♦ follows the recessive model, \*conserved region (phyloP >2.0). \* known CD risk loci. 1000G CEU frequencies taken from NCBI (these can be +-strand or -strand).

**Supplementary table 20 Exonic and splice-junction small SVs in genes in proximity of GWAS Crohn's disease SNVs.**

One small deletion was detected in the end of the last exon in *TNFSF18*, which was not predicted to cause nonsense mediated decay. A 1-basepair insertion was detected very close to a splice-acceptor in *DENND1B* (3 bp from exon).

gene symbol	rsID	chrom	start	end	type	zygosity	ref allele	alternative allele	location	SIFT causes nonsense mediated decay?
<i>TNFSF18</i>	-	chr1	171277162	171277163	deletion	hemizygous	TG	-	exonic	no
<i>DENND1B</i>	rs11411505	chr1	195842930	195842930	insertion site	hemizygous	-	T	splice-junction (intronic)	-

Mother homozygous for TNFSF18 deletion, Father also hemizygous for DENND1B.

Digital only tables:

- Supplementary table 21** Detected CD associated SNVs (meta-analysis).
- Supplementary table 22** Known Crohn's Disease associated SNVs detected (GWAS catalog).
- Supplementary table 23** Overview of noncoding SNVs associated with Genes in proximity of GWAS CD SNVs in the child.
- Supplementary table 24** Noncoding SNVs in conserved regions of Crohn's Disease associated genes.
- Supplementary table 25** Overview of noncoding small structural variations in genes associated with Crohn's Disease.

**Supplementary table 26** Nonsynonymous SNVs in genes involved in other inflammatory diseases, which are predicted to be damaging.

Nine genes carry variants in conserved regions and five of them were predicted to be damaging by both applied prediction tools (*NOD2*, *MST1* (already excluded by Sanger sequencing), *ITGAX*, *IRF5* and *ERBB3*). Four of the variants pose novel variants, as they are not included in dbSNP130 (*FRMD4B*, *IRF5*, *KIAA1109* and *LRRK2*).

gene symbol	rsID	chrom	position	phyloP	type	zygosity	ref allele	allele A	allele B	reads ref	reads alt	freqA	freqC	freqG	freqT	SIFT	SIFT score	SIFT AA	PolyPhen	dScore
<i>BMP4</i>	rs17563	chr14	53487272	1.422	missense	hom	A	G	G	0	28	-	0.556	-	0.444	tolerated	0.06	V152A	<i>probably damaging</i>	-2.534
<i>C1QTNF6</i> ✠	rs229526	chr22	35911368	2.42	missense	het	G	G	C	26	26	-	0.778	0.222	-	tolerated	0.85	P42R	<i>probably damaging</i>	-3.786
<i>CCRL2</i>	rs3204849	chr3	46425074	-2.024	missense	het	T	T	A	23	15	0.233	-	-	0.767	<i>damaging</i>	0.00	F179Y	benign	-3.681
<i>CD6</i>	rs11230562	chr11	60532762	-0.349	new Met	het	C	T	C	2	4	-	0.833	-	0.167	tolerated	0.22	T217M	<i>probably damaging</i>	-4.44
<i>CD6</i>	rs11230563	chr11	60532785	0.577	missense	het	C	T	C	7	4	-	0.722	-	0.278	tolerated	0.10	R225W	<i>possibly damaging</i>	-2.186
<i>CIITA</i>	rs8046121	chr16	10903434	1.244	missense	hom	A	G	G	0	29	0.025	-	0.975	-	tolerated	1.00	R174G	<i>possibly damaging</i>	-3.505
<i>CIITA</i> ✠	rs7197779	chr16	10910428	2.299	missense	hom	A	G	G	0	17	-	-	1.000	-	tolerated	0.64	Q900R	<i>probably damaging</i>	-3.397
<i>ERBB3</i> ✠	rs773123	chr12	54781265	2.845	missense	het	A	T	A	20	22	0.117	-	-	0.883	<i>damaging (low confidence)</i>	0.01	S1119C	<i>probably damaging</i>	-4.703
<i>FRMD4B</i> ✠	rs9831516	chr3	69312751	3.208	missense	hom	G	A	A	0	25	1.000	-	-	-	-	-	-	<i>probably damaging</i>	-3.175
<i>FRMD4B</i> ✠	-	chr3	69313097	3.258	missense	het	G	G	C	12	12	na	na	na	na	-	-	-	<i>possibly damaging</i>	-2.837
<i>GSDMB</i>	rs2305479	chr17	35315743	-1.142	missense	hom	C	T	T	0	56	0.472	-	0.528	-	<i>damaging (low confidence)</i>	0.00	G291R	<i>probably damaging</i>	-3.57
<i>HLA-B</i>	rs41541519	chr6	31432043	-0.688	missense	het	T	T	A	3	2	0.967	-	-	0.033	<i>damaging (low confidence)</i>	0.02	T167S	benign	-1.86
<i>HLA-B</i>	rs1131215	chr6	31432495	-1.819	missense	hom	C	A	A	0	2	-	-	0.592	0.408	tolerated	1.00	D98Y	<i>probably damaging</i>	-3.522
<i>HLA-B</i>	rs1050529	chr6	31432594	0.357	missense	het	C	T	C	2	6	0.258	-	0.742	-	<i>damaging (low confidence)</i>	0.05	A65T	<i>possibly damaging</i>	-2.878
<i>HLA-B</i>	rs1050518	chr6	31432620	1.647	missense	het	T	T	A	15	7	na	na	na	na	<i>damaging (low confidence)</i>	0.04	Q56L	<i>possibly damaging</i>	-3.04
<i>HLA-C</i>	rs1050328	chr6	31346134	0.364	missense	hom	G	A	A	0	8	na	na	na	na	tolerated	0.07	R243W	<i>probably damaging</i>	-4.926
<i>HLA-C</i>	rs1050357	chr6	31346859	1.672	missense	het	C	T	C	13	3	na	na	na	na	<i>damaging (low confidence)</i>	0.03	E197K	<i>possibly damaging</i>	-2.689
<i>HLA-C</i>	rs2308574	chr6	31347039	0.654	missense	het	A	G	A	5	2	na	na	na	na	tolerated	0.06	Y137H	<i>probably damaging</i>	-3.462
<i>HLA-C</i>	rs1050409	chr6	31347480	1.9	missense	het	G	T	G	3	5	na	na	na	na	<i>damaging (low confidence)</i>	0.02	A73E	<i>probably damaging</i>	-3.122
<i>HLA-DQB1</i>	rs35418872	chr6	32737883	-0.916	missense	hom	T	A	A	0	3	na	na	na	na	-	-	D167V	<i>possibly damaging</i>	-3.685
<i>HLA-DRB1</i>	rs17879995	chr6	32660050	1.572	missense	het	T	T	G	2	10	0.900	0.100	-	-	-	-	N62H	<i>probably damaging</i>	-3.687
<i>HLA-DRB5</i>	known to 1000G	chr6	32597800	-0.595	missense	het	C	T	C	12	4	na	na	na	na	<i>damaging (low confidence)</i>	0.03	R77Q	<i>possibly damaging</i>	-2.43
<i>IL17REL</i>	rs5771069	chr22	48777607	-1.432	missense	hom	A	G	G	0	12	0.392	-	0.608	-	<i>damaging (low confidence)</i>	0.00	L333P	benign	-1.458
<i>IL1RL2</i>	rs2302612	chr2	102218140	0.347	missense	hom	T	C	C	0	20	0.792	-	0.208	-	<i>damaging (low confidence)</i>	0.00	L550P	benign	-2.407
<i>IL27</i>	rs181206	chr16	28420904	1.956	missense	het	A	G	A	4	7	-	0.333	-	0.667	tolerated	0.09	L119P	<i>possibly damaging</i>	-3.708



<b>IRF5</b> ♣	-	chr7	128375070	2.132	missense	het	T	T	C	6	2	na	na	na	na	<b>damaging</b>	0.03	L264P	<b>probably damaging</b>	-3.969
<b>ITGAX</b> ♣	rs2230429	chr16	31282036	2.052	missense	het	C	G	C	13	7	-	0.708	0.292	-	<b>damaging</b>	0.01	P517R	<b>possibly damaging</b>	-2.743
<b>KIAA1109</b> ♣	-	chr4	123445422	2.748	missense	het	C	T	C	33	10	na	na	na	na	-	-	-	<b>probably damaging</b>	-3.242
<b>LCE3D</b>	rs61745411	chr1	150818909	0.446	missense	het	C	A	C	13	22	na	na	na	na	-	-	G43V	unknown	?
<b>LRRK2</b>	-	chr12	38915756	1.354	missense	het	A	G	A	35	8	na	na	na	na	<b>damaging</b>	0.02	K137E	benign	-2.929
<b>MST1A</b> ♣	rs71324987	chr3	49698754	2.623	missense	het	A	G	A	7	3	na	na	na	na	<b>damaging</b>	0.00	C298R	<b>probably damaging</b>	-3.975
<b>NOD2</b> ♣	rs2066845	chr16	49314041	3.304	missense	het	G	G	C	25	9	na	na	na	na	<b>damaging</b>	0.00	G908R	<b>probably damaging</b>	-3.854
<b>RNF186</b>	rs1541185	chr1	20014115	-1.503	missense	het	C	T	C	6	5	-	0.764	-	0.236	tolerated	0.12	A23T	<b>possibly damaging</b>	-2.878
<b>SLC9A4</b>	rs1014286	chr2	102515532	-1.259	missense	hom	G	A	A	0	46	-	0.319	-	0.681	<b>damaging</b> (low confidence)	0.00	G784S	benign	-1.781
<b>SULT1A1</b>	rs9282861	chr16	28525015	1.603	missense	het	C	T	C	35	26	na	na	na	na	<b>damaging</b>	0.01	R213H	benign	-2.531
<b>SULT1A2</b> ♣	rs10797300	chr16	28514697	2.545	missense	het	G	G	A	8	9	na	na	na	na	tolerated	0.39	P19L	<b>probably damaging</b>	-3.36
<b>TLR8</b>	rs3764880	chrX	12834747	-1.233	altered Met	hom	A	G	G	0	13	na	na	na	na	<b>damaging</b> (low confidence)	0.00	M1V	benign	-2.941
<b>TNFRSF14</b>	rs4870	chr1	2486265	-0.599	missense	het	T	T	C	6	8	na	na	na	na	tolerated	0.17	K17R	<b>probably damaging</b>	-2.768
<b>WDFY4</b>	rs7072606	chr10	49603980	0.392	missense	het	T	T	C	13	7	-	0.075	-	0.925	<b>damaging</b>	0.05	S214P	<b>possibly damaging</b>	-3.6
<b>WDFY4</b>	rs2170132	chr10	49683408	0.407	missense	het	T	T	C	20	13	-	0.333	-	0.667	<b>damaging</b>	0.02	S239P	benign	-1.961
<b>WDFY4</b>	rs7097397	chr10	49695402	1.07	missense	hom	G	A	A	0	6	0.375	-	0.625	-	tolerated	0.44	R527Q	<b>possibly damaging</b>	-3.605
<b>WDFY4</b>	rs2292584	chr10	49856421	-0.452	missense	het	C	T	C	12	5	0.417	-	0.583	-	<b>damaging</b> (low confidence)	0.01	P256L	benign	-2.24

■ genes also associated with Crohn's Disease, ♣ follows the recessive model, ♣ conserved region (phyloP >2.0)

#### Supplementary table 27 Exonic and splice-junction sSVs in genes associated with other inflammatory diseases.

The variants in *DENND1B* and *TNFSF18* have been discussed above, as these genes are also possibly associated with CD. The remaining variant is a deletion in *CSMD1* (cub and sushi multiple domains 1) which was predicted to cause nonsense mediated decay and found to follow the recessive model of inheritance.

trait	rsID	chrom	start	end	type	reference allele	alternative allele	zygosity	associated gene	location	SIFT causes nonsense mediated decay?
<b>Asthma</b>	rs11411505	chr1	195842930	195842930	Insertion site <sup>2</sup>	-	T	hemizygous	<i>DENND1B</i> ■	splice-junction (intronic)	-
<b>Celiac disease</b>	-	chr1	171277162	171277163	deletion <sup>1</sup>	TG	-	hemizygous	<i>TNFSF18</i> ■	exonic	no
<b>Multiple sclerosis</b>	rs34128718	chr8	3254412	3254412	deletion <sup>1,2,3</sup>	C	-	homozygous	<i>CSMD1</i>	exonic	<b>yes</b>

<sup>1</sup> shared with mother; <sup>2</sup> shared with father; <sup>3</sup> following recessive model; ■ genes also associated with Crohn's Disease

#### Supplementary table 28 Large structural variations in proximity of genes involved in other inflammatory disease.

Six large deletions were detected in proximity of genes associated with inflammatory diseases. One of the deletions includes the third exon of the growth hormone receptor (*GHR*), which has been associated to an increased responsiveness to growth hormone. *GHR* has also been mentioned in the context with systemic lupus erythematosus. One deletion following the recessive model of inheritance was identified in *CSMD1*, a multiple sclerosis associated gene.

like DG Var [id]	gene_symb	chrom	loose start	loose end	length	teresa	val	dot	score	trait
------------------	-----------	-------	-------------	-----------	--------	--------	-----	-----	-------	-------

(20kb range)										
-	<i>CRB1</i>	chr1	195534187	195537759	3745	hemi	-	hemi	(+)	Asthma (GWAS catalog gene)
<b>59103</b>	<i>SCHIP1</i>	chr3	160738319	160741679	3538	-	hemi	hemi	+	Celiac disease (GWAS catalog gene)
<b>39518</b> ✧	<i>CSMD1</i>	chr8	4108127	4114595	6592	hemi	hemi	hom	+++	Multiple sclerosis (GWAS catalog gene)
<b>62242</b>	<i>BANK1</i>	chr4	103005737	103008451	3031	hemi	hom	hemi	+	Systemic lupus erythematosus (GWAS catalog gene)
<b>99119</b>	<i>GHR*</i> (exons, 65 bp)	chr5	42662319	42668587	6471	hemi	hemi	hemi	+	.0031 INCREASED RESPONSIVENESS TO GROWTH HORMONE [GHR, <b>EX3DEL</b> ] (OMIM) Systemic lupus erythematosus (GWAS catalog gene)
<b>102022</b>	<i>MEG3, RTL1</i>	chr14	100404330	100409664	5392	hemi	hemi	hemi	(+)	Type 1 diabetes (GWAS catalog gene)

\*causes nonsense mediated decay (SIFT), ✧follows recessive model. Fraction of deleted exons in [bp] given in brackets.

Digital only tables:

**Supplementary table 29** Overview of noncoding SNVs in genes involved in other inflammatory diseases.

**Supplementary table 30** Noncoding SNVs in genes associated with inflammatory diseases following the recessive model (also including one non-inflammatory disease: Parkinson).

**Supplementary table 31** Overview of noncoding SNVs in genes associated with other inflammatory diseases in the child in locations of high conservation.

**Supplementary table 32** Small structural variations in genes associated with inflammatory diseases in the GWAS catalog.

trait	3'-neighbourhood	5'-neighbourhood	5'-UTR (intronic)	5'-UTR (exonic)	3'-UTR (exonic)	intronic	splice-junction (intronic)	exonic
Ankylosing spondylitis	3 (1)	0	9 (4)	0	1	30 (4)	0	0
Asthma	2	0	7 (2)	0	2	114 (18)	1	0
Celiac disease	14 (2)	10	57 (12)	4	7 (3)	325 (73)	0	1 (1)
Crohn's disease	4 (1)	3 (1)	22 (1)	0	3	221 (32)	0	0
Crohn's disease and sarcoidosis (combined)	1 (1)	0	0	0	1 (1)	1	0	0
Inflammatory bowel disease	0	2	1	0	0	9 (1)	0	0
Inflammatory bowel disease (early onset)	3	0	16 (5)	0	0	20 (6)	0	0
Leprosy	0	0	0	0	1 (1)	1	0	0
Multiple sclerosis	8 (2)	7	11 (1)	0	4 (2)	465 (77)	0	1
Psoriasis	1	2	6	0	2	39 (3)	0	0
Rheumatoid arthritis	0	4	17 (1)	0	1	73 (10)	0	0
Systemic lupus erythematosus	5	4	27 (7)	0	6 (2)	188 (31)	0	0
Type 1 diabetes	4 (1)	4	29 (2)	0	4	160 (30)	0	0
Ulcerative colitis	2	10 (1)	9 (1)	1	0	120 (20)	0	0

**Supplementary table 33 Nonsynonymous SNVs detected in the child following the recessive model of inheritance which are predicted to be damaging.**

Among the genes with SNVs following the recessive model of inheritance are 15 located in conserved regions. Two of these are not included in dbSNP130 (*BEST2* and *MPHOSPH8*), though the *BEST2* can also be found in the 1000 Genome data set.

gene symbol	rsID	chrom	position	phyloP	type	ref	alt	reads_ref	reads_alt	freqA	freqC	freqG	freqT	SIFT	SIFT	SIFT	PolyPhen	dScore
<i>ANKRD35</i>	rs6670984	chr1	144272951	-0.572	missense	C	T	0	49	-	0.625	-	0.375	damaging (low confidence)	0.00	P428S	benign	0.379
<i>ANO7</i>	rs7590653	chr2	241812032	-0.33	missense	G	A	0	10	0.278	-	0.722	-	damaging (low confidence)	0.00	E912K	benign	-0.122
<i>BEST2</i> ✱	known to 1000G	chr19	12724429	2.236	missense	G	A	0	22	na	na	na	na	tolerated	0.30	R8Q	possibly damaging	0.576
<i>BRWD1</i> ✱	rs2183573	chr21	39496175	2.39	missense	A	G	0	29	0.403	-	0.597	-	tolerated	0.10	S1511P	probably damaging	2.08
<i>C12orf71</i>	rs61741737	chr12	27126181	1.729	missense	G	A	0	21	0.125	-	0.875	-	-	-	-	possibly damaging	1.728
<i>C14orf37</i>	rs2273442	chr14	57633447	0.996	missense	G	C	0	34	-	0.528	0.472	-	damaging (low confidence)	0.00	Q613E	benign	-1.705
<i>C17orf74</i>	rs72842820	chr17	7269858	-0.909	new Met	G	A	0	25	0.139	-	0.861	-	damaging (low confidence)	0.00	V43M	probably damaging	2.461
<i>C17orf90</i>	rs61742303	chr17	77245179	0.533	missense	A	T	0	46	0.819	-	-	0.181	tolerated	0.10	K50N	probably damaging	1.918
<i>C2orf73</i> ✱	rs55714450	chr2	54415516	2.365	missense	C	A	0	17	0.375	0.625	-	-	damaging (low confidence)	0.00	H30N	benign	0.068
<i>CCDC137</i>	rs11150805	chr17	77249210	1.095	missense	C	T	0	27	-	0.847	-	0.153	damaging (low confidence)	0.01	R177W	probably damaging	3.235
<i>CCDC60</i>	rs1064319	chr12	118350916	0.497	missense	A	G	0	31	0.625	-	0.375	-	tolerated	0.27	I46V	probably damaging	0.703
<i>CDC20B</i>	rs444527	chr5	54445856	1.299	missense	G	A	0	23	-	0.806	-	0.194	damaging (low confidence)	0.03	R499W	probably damaging	3.039
<i>COL4A3</i>	rs28381984	chr2	227843875	1.477	missense	C	T	0	50	-	0.5	-	0.5	-	-	-	probably damaging	1.751
<i>EPHA8</i>	rs606002	chr1	22788340	-2.083	missense	T	C	0	19	0.658	-	0.342	-	damaging (low confidence)	0.00	S457P	-	-
<i>FAM75D1</i>	-	chr9	83797224	0.337	new Met	T	G	1	11	na	na	na	na	tolerated	0.10	I673M	possibly damaging	1.785
<i>FCRLB</i>	rs34868416	chr1	159963943	0.844	missense	G	A	0	12	0.117	-	0.883	-	tolerated	0.13	G383D	possibly damaging	1.88
<i>FRG2B</i>	known to 1000G	chr10	135288967	-1.606	missense	G	A	0	15	na	na	na	na	-	-	-	possibly damaging	2.882
<i>GBE1</i> ✱	rs2229519	chr3	81780820	2.468	missense	T	C	0	21	0.681	-	0.319	-	damaging	0.04	R190G	benign	0.91
<i>GOLGA4</i> ✱	rs11718848	chr3	37341463	2.503	missense	C	A	0	25	0.389	0.611	-	-	tolerated	0.06	Q1028K	possibly damaging	1.253
<i>H1FNT</i>	rs1471997	chr12	47009862	0.307	missense	G	A	0	6	-	0.819	-	0.181	damaging (low confidence)	0.00	R174Q	benign	0.192
<i>HAP1</i>	rs4796693	chr17	37138109	0.478	splice-junction	G	A	0	15	0.792	-	0.208	-	tolerated	0.33	S357L	probably damaging	1.521
<i>HAP1</i>	rs4796603	chr17	37144241	1.512	missense	A	T	0	9	0.125	-	-	0.875	tolerated	0.26	S58T	possibly damaging	1.134
<i>HLA-DPB1</i>	rs1042131	chr6	33156580	1.949	missense	C	A	0	20	na	na	na	na	damaging (low confidence)	0.05	A85E	benign	0.645
<i>KATNAL2</i>	rs3744863	chr18	42814298	-0.378	missense	C	T	0	19	na	na	na	na	tolerated	0.31	A446T	probably damaging	1.757
<i>KCNJ12</i> ✱	rs73979896	chr17	21259801	2.462	missense	C	T	1	4	na	na	na	na	damaging	0.01	A185V	benign	0.899
<i>KRT76</i> ✱	rs11170271	chr12	51453662	2.408	missense	C	T	0	24	-	0.861	-	0.139	damaging	0.01	A283T	possibly damaging	1.31
<i>LLGL2</i>	rs1671036	chr17	71063780	1.362	missense	G	A	0	38	-	0.458	-	0.542	damaging	0.04	R45H	benign	0.573
<i>LOC100131320</i>	rs4662674	chr2	130454633	-0.549	missense	G	A	0	22	na	na	na	na	-	-	A159T	benign	0.083
<i>LOXL4</i> ✱	rs1983864	chr10	100007443	2.355	missense	T	G	0	34	0.708	0.292	-	-	damaging	0.00	D405A	probably damaging	2.431
<i>MCPH1</i>	rs2083914	chr8	6289562	-0.460	missense	G	T	0	25	-	-	0.847	0.153	damaging (low confidence)	0.02	R304I	probably damaging	2.11
<i>MCPH1</i>	rs12674488	chr8	6325714	-0.544	missense	C	A	0	47	0.153	0.847	-	-	tolerated	0.17	T682N	probably damaging	1.739
<i>MPHOSPH8</i> ✱	known to 1000G	chr13	19122202	3.165	missense	G	T	0	21	na	na	na	na	damaging (low confidence)	0.00	D460Y	probably damaging	2.655

<i>MUC20</i>	rs2688542	chr3	196938622	0.752	missense	G	C	0	15	na	na	na	na	-	-	-	<i>probably damaging</i>	2.683
<i>NOM1</i> ✘	rs2302443	chr7	156454985	3.014	missense	G	C	0	34	-	0.444	0.556	-	<i>damaging</i> (low confidence)	0.05	V804L	benign	0.217
<i>OBSL1</i> ✘	rs1983210	chr2	220129661	2.503	missense	C	G	0	9	-	0.292	0.708	-	tolerated	0.29	E264D	<i>possibly damaging</i>	0.194
<i>OR10A6</i>	rs7933807	chr11	7906367	0.708	missense	A	C	0	43	0.5	0.5	-	-	<i>damaging</i>	0.00	V140G	<i>probably damaging</i>	1.918
<i>OR10G4</i>	rs4936880	chr11	123392194	1.943	missense	A	G	0	18	0.592	-	0.408	-	<i>damaging</i>	0.00	R235G	<i>probably damaging</i>	1.674
<i>OR10G4</i>	rs4936881	chr11	123392374	1.961	missense	A	C	0	42	na	na	na	na	<i>damaging</i>	0.01	K295Q	benign	1.155
<i>OTOL1</i>	rs3921595	chr3	162704399	0.668	missense	A	C	0	14	0.5	0.5	-	-	<i>damaging</i> (low confidence)	0.03	E470A	<i>possibly damaging</i>	1.313
<i>OVCH2</i>	rs7927138	chr11	7684462	0.640	missense	C	T	0	22	-	0.556	-	0.444	<i>damaging</i> (low confidence)	0.00	R19Q	benign	-1.448
<i>PCDHA1</i> ✘	rs2240695	chr5	140148335	2.532	missense	G	T	0	13	0.569	0.431	-	-	<i>damaging</i>	0.00	C759F	-	-
<i>PCDHA1</i>	rs7701755	chr5	140162285	1.470	missense	G	T	0	37	-	-	0.4	0.6	<i>damaging</i>	0.02	S440I	-	-
<i>PCDHA1</i> ✘	rs4141841	chr5	140183616	2.044	missense	C	T	0	8	0.592	-	0.408	-	<i>damaging</i>	0.04	A691V	-	-
<i>PCDHB16</i>	rs61742755	chr5	140544447	0.342	missense	C	T	3	38	na	na	na	na	<i>damaging</i>	0.01	A710V	benign	0.812
<i>PCDHB7</i> ✘	rs62378900	chr5	140534665	2.089	missense	G	T	0	6	na	na	na	na	<i>damaging</i>	0.01	V689L	<i>probably damaging</i>	1.387
<i>PKD1L2</i>	rs7191351	chr16	79807455	0.523	missense	T	A	0	16	0.617	-	-	0.383	<i>damaging</i> (low confidence)	0.00	Q120L	<i>probably damaging</i>	1.719
<i>POLL</i>	rs3730477	chr10	103330046	0.400	missense	G	A	1	22	-	0.806	-	0.194	tolerated	0.06	R438W	<i>probably damaging</i>	3.207
<i>PON1</i>	rs854560	chr7	94784020	-0.305	new Met	A	T	0	43	0.569	-	-	0.431	<i>damaging</i>	0.04	L55M	benign	1.047
<i>PRPH2</i>	rs390659	chr6	42774142	1.302	missense	G	C	0	24	-	-	0.183	0.817	tolerated	1.00	Q304E	<i>probably damaging</i>	2.079
<i>PSMD9</i> ✘	rs14259	chr12	120838179	2.408	missense	A	G	0	25	0.431	-	0.569	-	<i>damaging</i>	0.05	E197G	benign	1.372
<i>PTRN2</i>	rs1130496	chr7	157652656	1.335	missense	C	T	0	28	0.825	-	0.175	-	tolerated	0.11	R213H	<i>possibly damaging</i>	2.14
<i>RALGAPA1</i>	rs2274068	chr14	35222928	-0.374	missense	T	C	0	35	0.444	0.556	-	-	<i>damaging</i> (low confidence)	0.00	T980A	benign	0.081
<i>RGL3</i>	rs167479	chr19	11387765	1.341	missense	G	T	0	4	na	na	na	na	<i>damaging</i>	0.00	P162H	<i>probably damaging</i>	1.684
<i>RGPD3</i>	rs72627454	chr2	106439933	1.283	missense	C	T	0	61	-	0.583	0.417	-	tolerated	0.46	D111N	-	-
<i>RIBC2</i>	rs1022478	chr22	44200620	-2.443	missense	C	G	0	10	0.875	-	0.125	-	<i>damaging</i>	0.04	F195L	benign	0.23
<i>SEC16B</i>	rs7522194	chr1	176172092	0.861	missense	C	A	0	34	0.167	0.833	-	-	tolerated	0.34	Q560H	<i>possibly damaging</i>	2.451
<i>SEL1L2</i>	rs41275404	chr20	13860309	0.936	missense	G	A	0	27	0.167	-	0.833	-	<i>damaging</i> (low confidence)	0.00	R75C	<i>possibly damaging</i>	2.855
<i>SEMA3F</i>	rs1046956	chr3	50197930	-0.574	new Met	T	A	0	11	0.667	-	-	0.333	tolerated	0.23	L503M	benign	0.232
<i>SERPINA3</i>	rs4934	chr14	94150556	0.343	missense	G	A	0	26	0.514	-	0.486	-	<i>damaging</i> (low confidence)	0.01	A9T	<i>possibly damaging</i>	1.571
<i>TBC1D1</i>	rs6811863	chr4	37638581	-1.976	missense	G	C	0	40	-	0.542	0.458	-	-	-	-	<i>possibly damaging</i>	2.164
<i>VWF</i>	rs1800378	chr12	6042463	0.413	missense	T	C	0	23	0.389	-	0.611	-	tolerated	1.00	H484R	<i>possibly damaging</i>	0.173
<i>ZNF142</i>	rs3770213	chr2	219216616	-0.646	missense	A	T	0	41	0.7	-	-	0.3	<i>damaging</i> (low confidence)	0.00	L956H	<i>possibly damaging</i>	2.595
<i>ZNF79</i>	rs13292096	chr9	129231007	0.486	missense	C	T	0	45	-	0.458	-	0.542	tolerated	0.13	T31I	<i>possibly damaging</i>	1.851

✘conserved region (phyloP >2.0)

**Supplementary table 34 Expression values calculated with Cufflinks (FPKM) for MPHOSPH8 and BEST2 in the transcriptomes from blood and colon biopsy (only for the child).**

Expression of the highest expressed isoform of MPHOSPH8 is very similar in all three individuals, though slightly higher in the mother (4.08) compared to the father and child (2.48, 2.74). The gene is also expressed in the tissue of the colon biopsy on a similar level. BEST2 is not expressed in the blood samples, but the colon biopsy shows expression.

trans_id	gene_symbol	FPKM_mother	FPKM_father	FPKM_child	FPKM_child_biopsy
uc001umf.1	MPHOSPH8	0.38	0.25	0.43	0.06
uc001umg.2	MPHOSPH8	0.27	0.17	0.21	0.16
uc001umh.1	MPHOSPH8	4.08	2.48	2.74	3.11
uc001umi.2	MPHOSPH8	0.01	0.01	0.01	0.09
uc002mux.1	BEST2	0.00	0.00	0.00	1.16

**Supplementary table 35 Exonic SNVs in other genes contributing to compound heterozygosity with at least one SNV predicted to be damaging.**

Among multiple genes with SNVs contributing to compound heterozygosity, only nine genes carry at least two SNVs which are distinctly inherited and predicted to be damaging (FBXW10, FLG, HLA-A, KCNJ12, MUC2, PNMT, SYNE2, TEK4, TLL12).

gene	rsID	#chr	pos	conservation	type	freqA	freqC	freqG	freqT	child alleleA	child alleleB	child reads ref	child reads alt	parent alleleA	parent alleleB	parent reads ref	parent reads alt	Mother	Father	SIFT	SIFT score	AA change	PolyPhen	dScore
C6orf15	rs1265054	chr6	31187622	-0.396	heterozygous missense	0.458	-	0.542	-	T	C	26	8	T	C	16	5	-	SNP	tolerated	1	K165E	benign	-0.672
C6orf15	rs2233976	chr6	31187973	0.485	heterozygous missense	0.083	-	0.917	-	T	C	22	19	T	C	13	14	SNP	-	not scored	N/A	0	probably damaging	2.541
CCHCR1	rs130072	chr6	31220463	2.235	heterozygous missense	na	na	na	na	T	C	23	15	T	C	14	11	SNP	-	not scored	N/A	0	probably damaging	1.945
CCHCR1	rs130067	chr6	31226490	-0.61	heterozygous missense	na	na	na	na	T	G	28	18	T	G	21	22	-	SNP	tolerated	0.35	E275D	benign	0.147
CCHCR1	rs11540822	chr6	31226877	1.915	heterozygous missense	na	na	na	na	T	A	14	13	T	A	8	6	SNP	-	not scored	N/A	0	probably damaging	2.709
CCHCR1	rs3130453	chr6	31232828	0.936	heterozygous nonsense	na	na	na	na	T	C	22	23	T	C	26	17	SNP	-	-	-	-	-	-
DNAH8	rs1738254	chr6	38855748	0.5	heterozygous missense	-	0.933	-	0.067	G	A	17	8	G	A	11	14	SNP	-	tolerated	0.53	G473R	benign	-0.062
DNAH8	rs1678674	chr6	38867357	1.814	heterozygous missense	0.05	-	0.95	-	G	A	18	20	G	A	20	20	SNP	-	damaging	0.04	A727T	benign	1.317
DNAH8	rs874808	chr6	38881271	1.448	heterozygous missense	-	0.583	-	0.417	G	A	35	16	G	A	39	16	-	SNP	tolerated	1	G807E	benign	-0.713
DSPP	?	chr4	88755386	-0.387	heterozygous missense	na	na	na	na	G	A	3	3	G	A	9	4	SNP	-	tolerated	0.21	G850S	unknown	?
DSPP	?	chr4	88756292	1.559	heterozygous missense	na	na	na	na	G	A	2	4	G	A	2	2	-	SNP	not scored	N/A	NA	possibly damaging	0.614
FBXW10	rs62073746	chr17	18588350	0.463	heterozygous missense	na	na	na	na	T	A	36	15	T	A	20	14	SNP	-	damaging (low confidence)	0	I23N	possibly damaging	1.98
FBXW10	?	chr17	18612686	0.607	heterozygous missense	na	na	na	na	T	C	9	3	T	C	7	6	-	SNP	not scored	N/A	0	probably damaging	3.586
FLG	?	chr1	150543068	1.231	heterozygous missense	na	na	na	na	A	C	26	6	A	C	20	23	SNP	-	damaging (low confidence)	0	S1046A	probably damaging	0.905
FLG	rs11582087	chr1	150545480	0.546	heterozygous missense	na	na	na	na	T	G	17	11	T	G	26	11	-	SNP	damaging (low confidence)	0.01	S181R	probably damaging	1.871
FRG2B	?	chr10	135289084	-1.759	heterozygous missense	na	na	na	na	T	C	9	2	T	C	12	3	SNP	-	not scored	N/A	0	benign	-1.372
FRG2B	?	chr10	135290193	0.485	heterozygous missense	na	na	na	na	T	A	41	8	T	A	34	13	-	SNP	not scored	N/A	0	benign	1.834
FRG2B	rs9630045	chr10	135290216	0.514	heterozygous missense	na	na	na	na	G	C	27	10	G	C	32	8	SNP	-	not scored	N/A	0	probably damaging	1.183

<b>FSIP2</b>	?	chr2	186363112	3.23	heterozygous missense	na	na	na	na	G	C	18	9	G	C	19	6	-	SNP	not scored	N/A	NA	<b>possibly damaging</b>	2.612
<b>FSIP2</b>	rs16827154	chr2	186379025	-0.704	heterozygous missense	0.5	-	-	0.5	T	A	31	16	T	A	16	14	SNP	-	-	-	-	benign	1.17
<b>FSIP2</b>	rs17826534	chr2	186379602	0.789	heterozygous missense	0.508	-	0.492	-	G	A	23	27	G	A	16	24	SNP	-	-	-	-	benign	1.146
<b>HLA-A</b>	rs707910	chr6	30018642	1.065	heterozygous missense	na	na	na	na	G	A	8	8	G	A	5	2	-	SNP	<b>damaging</b> (low confidence)	0.03	R68K	benign	0.313
<b>HLA-A</b>	rs1059526	chr6	30019197	0.999	heterozygous missense	na	na	na	na	G	A	3	5	G	A	5	3	SNP	-	<b>damaging</b> (low confidence)	0.03	A173T	benign	0.577
<b>HLA-A</b>	rs1137631	chr6	30021016	-0.573	heterozygous new Met	0.142	-	0.858	-	G	A	25	10	G	A	18	4	SNP	-	<b>damaging</b> (low confidence)	0.04	V364M	benign	0.931
<b>IGSF3</b>	rs41301291	chr1	116928940	0.773	heterozygous new Met	-	0.983	-	0.017	T	C	21	24	T	C	22	17	SNP	-	not scored	N/A	0	<b>probably damaging</b>	2.116
<b>IGSF3</b>	?	chr1	116948152	2.299	heterozygous missense	na	na	na	na	T	C	15	7	T	C	5	9	-	SNP	tolerated	0.3	E434G	benign	1.454
<b>KCNJ12</b>	rs8076599	chr17	21259545	-1.181	heterozygous missense	na	na	na	na	G	A	29	13	G	A	22	20	SNP	-	not scored	N/A	0	benign	-0.432
<b>KCNJ12</b>	?	chr17	21259662	2.841	heterozygous missense	na	na	na	na	G	A	24	7	G	A	26	6	-	SNP	<b>damaging</b>	0	E139K	<b>probably damaging</b>	1.91
<b>KCNJ12</b>	?	chr17	21259680	2.841	heterozygous missense	na	na	na	na	G	A	27	7	G	A	25	6	SNP	-	<b>damaging</b>	0	G145S	<b>probably damaging</b>	2.295
<b>KCNJ12</b>	rs1714864	chr17	21259714	2.462	heterozygous missense	na	na	na	na	T	C	34	18	T	C	34	11	SNP	-	not scored	N/A	0	<b>probably damaging</b>	2.579
<b>KCNJ12</b>	rs72846667	chr17	21259878	2.462	heterozygous missense	na	na	na	na	T	C	39	9	T	C	39	7	-	SNP	not scored	N/A	0	<b>possibly damaging</b>	1.284
<b>KRT27</b>	rs981684	chr17	36189338	1.951	heterozygous missense	0.433	-	0.567	-	G	A	27	11	G	A	25	13	SNP	-	not scored	N/A	0	benign	-1.392
<b>KRT27</b>	rs17558532	chr17	36189402	2.749	heterozygous missense	-	0.833	-	0.167	T	C	6	12	T	C	8	10	-	SNP	not scored	N/A	0	<b>possibly damaging</b>	0.862
<b>KRT27</b>	rs17558560	chr17	36190185	0.82	heterozygous missense	-	0.525	-	0.475	T	C	24	13	T	C	15	14	SNP	-	not scored	N/A	0	benign	1.306
<b>L1TD1</b>	?	chr1	62445477	-1.585	heterozygous missense	na	na	na	na	G	A	5	11	G	A	22	7	-	SNP	not scored	N/A	0	benign	-0.244
<b>L1TD1</b>	rs2457828	chr1	62445875	0.819	heterozygous missense	-	-	0.133	0.867	A	C	5	4	A	C	6	5	SNP	-	<b>damaging</b> (low confidence)	0	K329N	benign	1.421
<b>LRRC37A3</b>	rs17857225	chr17	60322621	-6.264	heterozygous missense	na	na	na	na	T	G	2	2	T	G	3	2	SNP	-	not scored	N/A	0	benign	0.01
<b>LRRC37A3</b>	rs62071406	chr17	60322733	-1.046	heterozygous missense	na	na	na	na	T	A	21	9	T	A	4	8	-	SNP	not scored	N/A	0	<b>possibly damaging</b>	1.28
<b>MUC16</b>	rs61742668	chr19	8854018	-0.67	heterozygous missense	na	na	na	na	G	A	9	11	G	A	22	13	SNP	-	<b>damaging</b> (low confidence)	0	P731L	<b>probably damaging</b>	1.728
<b>MUC16</b>	rs61732552	chr19	8929374	-1.299	heterozygous missense	0.975	-	0.025	-	G	A	17	12	G	A	23	11	-	SNP	not scored	N/A	0	benign	0.653
<b>MUC16</b>	?	chr19	8932562	-1.312	heterozygous missense	na	na	na	na	T	C	13	12	T	C	15	17	-	SNP	not scored	N/A	0	benign	1.242
<b>MUC16</b>	?	chr19	8936370	-0.855	heterozygous missense	na	na	na	na	T	C	20	20	T	C	23	17	-	SNP	not scored	N/A	T4024A	benign	0.494
<b>MUC16</b>	?	chr19	8950974	-0.965	heterozygous missense	na	na	na	na	T	C	9	11	T	C	15	8	SNP	-	not scored	N/A	0	benign	0.647
<b>MUC2</b>	rs2856111	chr11	1065747	1.688	heterozygous missense	-	0.092	-	0.908	T	C	6	3	T	C	3	2	-	SNP	<b>damaging</b>	0.03	L58P	<b>probably damaging</b>	2.309
<b>MUC2</b>	rs41411848	chr11	1071074	1.657	heterozygous missense	-	0.183	-	0.817	T	C	7	6	T	C	6	4	-	SNP	tolerated	0.33	V457A	benign	0.555
<b>MUC2</b>	rs57737240	chr11	1071757	-0.521	heterozygous missense	-	0.175	0.825	-	G	C	4	5	G	C	5	6	-	SNP	tolerated	0.5	S562T	benign	0.327
<b>MUC2</b>	?	chr11	1082760	-3.036	heterozygous missense	na	na	na	na	A	C	11	3	A	C	9	6	-	SNP	not scored	N/A	NA	benign	0.225
<b>MUC2</b>	rs56219745	chr11	1082872	-0.413	heterozygous missense	na	na	na	na	G	C	18	11	G	C	16	14	-	SNP	not scored	N/A	0	benign	1.397
<b>MUC2</b>	rs11245947	chr11	1083057	-1.01	heterozygous missense	na	na	na	na	G	A	12	6	G	A	16	5	SNP	-	not scored	N/A	0	<b>possibly damaging</b>	1.245
<b>MUC4</b>	rs62284986	chr3	196959830	1.596	heterozygous missense	0.25	-	0.75	-	G	A	26	16	G	A	18	9	SNP	-	tolerated	0.1	A2102V	<b>probably damaging</b>	1.656
<b>MUC4</b>	rs2550240	chr3	196981550	1.781	heterozygous missense	-	0.45	0.55	-	G	C	22	13	G	C	17	9	-	SNP	tolerated	0.16	N1228K	benign	0.548
<b>MYOM1</b>	rs2230165	chr18	3178873	0.368	heterozygous new Met	-	0.85	-	0.15	G	A	12	6	G	A	6	5	-	SNP	<b>damaging</b> (low confidence)	0.04	T215M	<b>possibly damaging</b>	2.631
<b>MYOM1</b>	rs1791085	chr18	3205158	1.884	heterozygous missense	-	0.1	0.9	-	G	C	9	3	G	C	9	5	SNP	-	not scored	N/A	0	benign	-1.381
<b>NBPF10</b>	?	chr1	144009678	-0.727	heterozygous missense	na	na	na	na	G	A	40	10	G	A	22	4	SNP	-	<b>damaging</b> (low confidence)	0.04	E209K	-	-
<b>NBPF10</b>	?	chr1	144038905	-1.544	heterozygous missense	na	na	na	na	G	A	14	5	G	A	6	6	SNP	-	not scored	N/A	0	-	-
<b>NBPF10</b>	?	chr1	144072072	-0.941	heterozygous missense	na	na	na	na	G	A	7	3	G	A	12	9	-	SNP	not scored	N/A	V628I	-	-

<b>NEB</b>	?	chr2	152114413	3.45	heterozygous missense	na	na	na	na	G	A	18	13	G	A	9	13	-	SNP	<b>damaging</b>	0	R4977C	<b>probably damaging</b>	2.714
<b>NEB</b>	rs6713162	chr2	152204772	1.779	heterozygous missense	0.825	-	0.175	-	G	A	18	8	G	A	20	18	SNP	-	not scored	N/A	0	benign	-0.855
<b>NLRP1</b>	rs11651270	chr17	5365801	0.646	heterozygous altered Met	-	0.483	-	0.517	T	C	11	3	T	C	9	5	-	SNP	tolerated	1	M1188V	benign	-2.797
<b>NLRP1</b>	?	chr17	5374527	-0.518	heterozygous missense	na	na	na	na	T	C	33	17	T	C	18	20	SNP	-	<b>damaging (low confidence)</b>	0.01	Q1177R	benign	0.487
<b>NLRP1</b>	?	chr17	5403637	0.322	heterozygous missense	na	na	na	na	G	C	32	13	G	C	44	8	SNP	-	not scored	N/A	0	benign	0.585
<b>PIK3C2G</b>	rs11044004	chr12	18326719	1.66	heterozygous missense	-	0.583	-	0.417	T	C	17	11	T	C	12	5	SNP	-	not scored	N/A	0	benign	0.56
<b>PIK3C2G</b>	rs7133666	chr12	18335076	2.501	heterozygous missense	0.075	0.925	-	-	A	C	15	13	A	C	18	12	-	SNP	not scored	N/A	0	<b>possibly damaging</b>	1.875
<b>PLA2G4F</b>	?	chr15	40226736	0.85	heterozygous missense	na	na	na	na	T	C	34	19	T	C	33	14	SNP	-	not scored	N/A	0	<b>possibly damaging</b>	1.815
<b>PLA2G4F</b>	rs73403546	chr15	40230115	-0.783	heterozygous missense	-	0.067	0.933	-	G	C	6	4	G	C	6	3	-	SNP	tolerated	1	L248V	benign	0.103
<b>PLB1</b>	rs6753929	chr2	28615485	3.131	heterozygous missense	-	0.225	0.775	-	G	C	26	14	G	C	15	18	-	SNP	<b>damaging</b>	0	V212L	<b>probably damaging</b>	1.503
<b>PLB1</b>	rs11681826	chr2	28667537	-0.548	heterozygous altered Met	0.758	-	0.242	-	G	A	13	10	G	A	8	10	SNP	-	not scored	N/A	0	benign	-0.891
<b>PNMT</b>	rs72554035	chr17	35079524	1.075	heterozygous missense	na	na	na	na	T	C	20	15	T	C	17	16	SNP	-	not scored	N/A	0	<b>probably damaging</b>	3.736
<b>PNMT</b>	?	chr17	35079776	-0.642	heterozygous missense	na	na	na	na	T	C	11	4	T	C	4	6	-	SNP	not scored	N/A	0	<b>probably damaging</b>	3.355
<b>PRIM2</b>	rs6913546	chr6	57354843	0.384	heterozygous missense	na	na	na	na	G	A	36	21	G	A	27	18	SNP	-	not scored	N/A	0	benign	-0.304
<b>PRIM2</b>	rs9476080	chr6	57506116	1.942	heterozygous missense	na	na	na	na	G	A	27	9	G	A	25	9	-	SNP	not scored	N/A	0	<b>probably damaging</b>	3.192
<b>SLC19A1</b>	?	chr21	45776396	2.486	heterozygous missense	na	na	na	na	G	A	14	5	G	A	11	3	SNP	-	not scored	N/A	0	<b>probably damaging</b>	1.617
<b>SLC19A1</b>	rs1051266	chr21	45782222	0.909	heterozygous missense	0.433	-	0.567	-	T	C	8	4	T	C	5	3	-	SNP	tolerated	0.35	H27R	benign	-1.909
<b>SYNE2</b>	rs9944035	chr14	63517529	2.357	heterozygous missense	-	0.042	-	0.958	T	C	19	23	T	C	6	14	SNP	-	<b>damaging</b>	0.01	I574T	<b>probably damaging</b>	1.913
<b>SYNE2</b>	rs12881815	chr14	63674348	1.947	heterozygous missense	0.058	-	0.942	-	G	A	21	9	G	A	25	11	-	SNP	tolerated	0.06	E1298K	<b>possibly damaging</b>	0.767
<b>SYNM</b>	rs3743242	chr15	97487151	0.357	heterozygous missense	0.917	-	-	0.083	T	A	25	13	T	A	20	19	SNP	-	not scored	N/A	0	<b>probably damaging</b>	3.682
<b>SYNM</b>	rs3743244	chr15	97487788	-2.088	heterozygous missense	0	0.867	-	0.133	T	C	12	17	T	C	21	14	-	SNP	not scored	N/A	0	benign	-0.283
<b>SYNM</b>	rs5030699	chr15	97489318	1.289	heterozygous missense	0.075	-	0.925	-	T	C	2	5	T	C	4	4	-	SNP	not scored	N/A	0	benign	1.563
<b>SYNM</b>	rs7167599	chr15	97490122	0.325	heterozygous missense	-	0.133	0.867	-	G	C	27	10	G	C	9	10	-	SNP	tolerated	1	G1344A	benign	0.881
<b>TEKT4</b>	rs11164112	chr2	94901349	-0.486	heterozygous missense	0.775	-	0.225	-	G	A	8	2	G	A	5	2	SNP	-	not scored	N/A	0	benign	-0.594
<b>TEKT4</b>	?	chr2	94904333	0.441	heterozygous missense	na	na	na	na	T	C	6	2	T	C	2	5	-	SNP	not scored	N/A	0	<b>probably damaging</b>	1.506
<b>TEKT4</b>	?	chr2	94904378	1.401	heterozygous missense	na	na	na	na	G	C	8	3	G	C	9	4	SNP	-	not scored	N/A	0	<b>possibly damaging</b>	0.888
<b>TPCN2</b>	rs72928978	chr11	68587940	2.498	heterozygous missense	0.142	-	0.858	-	G	A	5	5	G	A	2	4	SNP	-	tolerated	0.06	V219I	<b>probably damaging</b>	1.05
<b>TPCN2</b>	rs3750965	chr11	68596736	0.689	heterozygous missense	0.717	-	0.283	-	G	A	18	9	G	A	11	7	-	SNP	not scored	N/A	0	benign	0.56
<b>TLL12</b>	rs11704935	chr22	41902298	1.288	heterozygous new Met	-	0.942	0	0.058	T	C	13	9	T	C	9	5	-	SNP	not scored	N/A	0	<b>possibly damaging</b>	2.097
<b>TLL12</b>	rs138951	chr22	41909027	-0.61	heterozygous missense	0.208	-	0.792	-	G	A	24	11	G	A	16	9	SNP	-	<b>damaging (low confidence)</b>	0.02	R84W	<b>possibly damaging</b>	2.842
<b>TXNRD2</b>	rs5992495	chr22	18262984	-0.695	heterozygous missense	-	-	0.2	0.8	T	G	8	3	T	G	7	7	-	SNP	tolerated	0.65	S299R	benign	-0.908
<b>TXNRD2</b>	rs5748469	chr22	18287099	2.528	heterozygous missense	0.325	0.675	-	-	A	C	6	5	A	C	7	4	SNP	-	not scored	N/A	0	<b>probably damaging</b>	1.774
<b>ZAN</b>	?	chr7	100187883	-1.252	heterozygous missense	na	na	na	na	T	C	8	2	T	C	11	4	-	SNP	not scored	N/A	0	benign	0.013
<b>ZAN</b>	?	chr7	100188297	-0.353	heterozygous missense	na	na	na	na	T	C	18	6	T	C	23	7	SNP	-	tolerated	0.32	L878P	benign	0.173
<b>ZAN</b>	rs314298	chr7	100209050	-0.475	heterozygous missense	0.633	-	0.367	-	T	C	4	3	T	C	5	3	SNP	-	not scored	N/A	0	benign	0.806
<b>ZAN</b>	rs314300	chr7	100212023	2.573	heterozygous missense	-	0.542	-	0.458	G	A	14	12	G	A	17	14	SNP	-	not scored	N/A	0	<b>probably damaging</b>	1.836
<b>ZAN</b>	?	chr7	100226564	1.283	heterozygous missense	na	na	na	na	A	C	12	6	A	C	11	10	SNP	-	not scored	N/A	0	<b>possibly damaging</b>	1.582
<b>ZNF518A</b>	rs41291604	chr10	97909001	1.431	heterozygous altered Met	0.925	-	0.075	-	G	A	12	14	G	A	11	11	SNP	-	not scored	N/A	0	benign	0.57

Gene	rsID	chr	pos	beta	type	na	na	na	na	G	A	9	11	G	A	10	11	-	SNP	damaging (low confidence)	0	R1328Q	probably damaging	2.108
ZNF518A	rs3814226	chr10	97910052	3.455	heterozygous missense	na	na	na	na	G	A	9	11	G	A	10	11	-	SNP	-	-	-	benign	0.37
ZNF717	rs63417460	chr3	75868726	0.663	heterozygous missense	na	na	na	na	G	A	39	7	G	A	34	7	-	SNP	-	-	-	benign	0.753
ZNF717	rs3009024	chr3	75869071	1.56	heterozygous missense	na	na	na	na	T	C	23	24	T	C	39	13	-	SNP	-	-	-	benign	0.373
ZNF717	?	chr3	75869075	-0.825	heterozygous missense	na	na	na	na	G	C	25	20	G	C	20	26	-	SNP	-	-	-	benign	0.165
ZNF717	?	chr3	75869431	-1.356	heterozygous altered Met	na	na	na	na	G	A	27	24	G	A	16	8	-	SNP	-	-	-	benign	0.361
ZNF717	rs3009020	chr3	75869443	-2.345	heterozygous missense	na	na	na	na	T	C	34	20	T	C	15	22	-	SNP	-	-	-	benign	-0.448
ZNF717	?	chr3	75869449	-1.613	heterozygous missense	na	na	na	na	T	C	22	31	T	C	27	16	-	SNP	-	-	-	benign	0.283
ZNF717	?	chr3	75870713	-2.49	heterozygous missense	na	na	na	na	T	C	28	18	T	C	32	22	-	SNP	-	-	-	possibly damaging	1.307
ZNF717	?	chr3	75870820	-0.994	heterozygous missense	na	na	na	na	A	C	42	9	A	C	41	9	-	SNP	-	-	-	benign	-0.992
ZNF717	?	chr3	75870842	-0.529	heterozygous missense	na	na	na	na	T	G	21	13	T	G	19	13	-	SNP	-	-	-	benign	0.883
ZNF717	?	chr3	75873510	1.181	heterozygous missense	na	na	na	na	G	A	40	17	G	A	36	8	-	SNP	not scored	N/A	NA	benign	0.883

### Supplementary table 36 Functional associations of genes listed in Supplementary table 35.

The list of genes containing compound heterozygous SNVs contains several genes associated with CD relevant immune functions (mucin genes *MUC2*, *MUC4*, *MUC16*; inflammasome *NLRP1*) and related diseases (colitis-associated colon cancer: *SLC19A1*; psoriasis: *CCHCR1*).


gene	function / association (as hyperlinks)
<i>C6orf15</i>	na
<i>CCHCR1</i>	the aberrant function of CCHCR1 may lead to abnormal keratinocyte proliferation which is a key feature of psoriatic epidermis. <sup>112</sup>
<i>DNAH8</i>	na
<i>DSPP</i>	expressed by the ectomesenchymal derived odontoblast cells (OMIM)
<i>FBXW10</i>	protein-ubiquitin ligase (OMIM)
<i>FLG</i>	facilitates epidermal differentiation and maintains barrier function (OMIM)
<i>FRG2B</i>	na
<i>FSIP2</i>	na
<i>HLA-A</i>	class I molecules play a central role in the immune system by presenting peptides derived from the endoplasmic reticulum lumen.
<i>IGSF3</i>	na
<i>KCNJ12</i>	potassium channel (OMIM)
<i>KRT27</i>	na
<i>L1TD1</i>	na
<i>LRRC37A3</i>	na
<i>MUC16</i>	Identification of Siglec-9 as the receptor for MUC16 on human NK cells, B cells, and monocytes. <sup>141</sup>
<i>MUC2</i>	The MUC2 is a major carrier of the sialyl-Tn antigen in all IM cases and in most gastric carcinoma cases. <sup>142</sup>
<i>MUC4</i>	Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. <sup>143</sup>
<i>MYOM1</i>	na
<i>NBPF10</i>	na
<i>NEB</i>	involved in skeletal muscle morphology (OMIM)
<i>NLRP1</i>	significant KPNA1-, NLRP1- and NLRP3-gene expression phenotypes associated with human genotypes of Crohn's disease, Huntington's disease and rheumatoid arthritis <sup>110</sup>
<i>PIK3C2G</i>	regulates diverse cellular responses, such as cell proliferation, oncogenic transformation, cell migration, intracellular protein trafficking, and cell survival (OMIM)
<i>PLA2G4F</i>	na
<i>PLB1</i>	expressed in the entire human epidermis with an accumulation in the dermoepidermis junction (OMIM)



<b>PNMT</b>	phenylethanolamine N-methyltransferase (OMIM)
<b>PRIM2</b>	primase *warning* pseudogenes probably present (OMIM)
<b>SLC19A1</b>	Use of a novel genetic mouse model to investigate the role of folate in colitis-associated colon cancer. <sup>113</sup>
<b>SPTBN5</b>	linking membrane proteins, membrane lipids, and cytosolic factors with the major cytoskeletal filament systems of the cell (OMIM)
<b>SYNE2</b>	na
<b>SYNM</b>	na
<b>TEKT4</b>	na
<b>TPCN2</b>	putative cation-selective ion channel (OMIM)
<b>TLL12</b>	na
<b>TXNRD2</b>	thioredoxin reductase that could directly reduce proteins such as insulin
<b>ZAN</b>	species-specific binding of sperm to eggs (OMIM)
<b>ZNF518A</b>	na
<b>ZNF717</b>	na

**Supplementary table 37 Excerpt of potential damaging exonic SNVs in conserved regions of other functional interesting genes in the child.**

This list exhibits a very short list of possibly damaging missense SNVs which were selected by gene function. A complete list can be found in Supplementary table 38 (only digital available). Among the genes in this list are *IL22RA2*, *TLR10*, *SIGLEC1*, *ECD*, *MMP9*, *ITGA3* and *XRCC6* which have a very low minor allele frequency (*IL22RA* 0.017; *TLR10* 0.028; *SIGLEC1* 0.083, *ECD* 0.069) or are not included in dbSNP130 (*MMP9*, *ITGA3*, *XRCC6*).

gene symbol	rsID	phyloP	chrom	pos	type	freqA	freqC	freqG	freqT	ref	genotype	ref reads	alt reads	SIFT	SIFT Score	SIFT AA exchange	PolyPhen	dScore
<i>XRCC6</i>	-	2.599	chr22	40362166	heterozygous missense	na	na	na	na	A	W	4	2	damaging	0.00	K100I	possibly damaging	2.706
<i>UNC5C</i>	rs2289043	2.711	chr4	96325345	homozygous altered Met	-	0.597	-	0.403	A	G	0	116	damaging	0.03	M721T	benign	0.975
<i>LIMS1</i>	-	2.196	chr2	108642546	heterozygous missense	na	na	na	na	T	W	18	6	damaging	0.01	F17Y	probably damaging	2.024
<i>ITGAX</i>	rs2230429	2.052	chr16	31282036	heterozygous missense	-	0.708	0.292	-	C	S	13	7	damaging	0.01	P517R	possibly damaging	1.241
<i>ILF3</i>	-	2.746	chr19	10653866	heterozygous missense	na	na	na	na	A	R	10	3	damaging (low confidence)	0.00	K460E	probably damaging	1.724
<i>IL22RA2</i>	rs28385692	2.490	chr6	137524533	heterozygous missense	-	0.017	-	0.983	A	R	25	20	damaging (low confidence)	0.03	L16P	benign	0.642
<i>CD5</i>	rs2229177	2.317	chr11	60649811	heterozygous missense	-	0.403	-	0.597	C	Y	23	23	damaging (low confidence)	0.00	A471V	probably damaging	1.985
<i>BEST2</i> 	known to 1000G	2.236	chr19	12724429	homozygous missense	na	na	na	na	G	A	0	22	tolerated	0.30	R8Q	possibly damaging	0.576
<i>ANP32A</i>	-	2.024	chr15	66900116	heterozygous missense	na	na	na	na	T	K	13	4	damaging	0.00	E8A	possibly damaging	0.69
<i>BIRC7</i>	-	2.637	chr20	61338227	heterozygous missense	na	na	na	na	G	R	13	4	damaging	0.00	G112S	possibly damaging	1.114
<i>DOCK5</i>	rs61732769	2.622	chr8	25230527	heterozygous new Met	na	na	na	na	C	Y	13	18	damaging	0.01	T469M	probably damaging	1.452
<i>ECD</i>	rs2271904	2.399	chr10	74564381	heterozygous missense	-	0.069	-	0.931	T	Y	32	19	damaging	0.00	D634G	possibly damaging	0.92
<i>ECD</i>	-	2.307	chr10	74566661	heterozygous missense	na	na	na	na	T	K	17	17	damaging	0.03	D504A	probably damaging	2.234
<i>FURIN</i>	-	2.262	chr15	89220552	heterozygous missense	na	na	na	na	C	Y	9	2	damaging	0.01	R81C	probably damaging	2.927
<i>LRMP</i>	rs1908946	3.220	chr12	25134382	homozygous missense	-	0.542	0.458	-	G	C	0	47	tolerated	0.09	C253S	probably damaging	1.586
<i>MAP3K1</i>	rs702689	3.272	chr5	56213200	homozygous missense	0.681	-	0.319	-	G	A	0	89	-	-	-	probably damaging	1.641
<i>MMP9</i>	-	2.604	chr20	44076428	heterozygous missense	na	na	na	na	A	M	3	2	tolerated	0.11	K638T	probably damaging	1.892
<i>NLRP12</i>	rs34436714	2.275	chr19	59019125	heterozygous missense	0.236	0.764	-	-	C	M	32	20	tolerated	0.17	G39V	probably damaging	1.99
<i>SIGLEC1</i>	rs34924243	2.633	chr20	3630126	heterozygous missense	0.083	-	0.917	-	C	Y	23	14	damaging (low confidence)	0.00	R464H	probably damaging	2.568
<i>TLR10</i>	rs11466653	2.376	chr4	38452630	homozygous altered Met	-	0.028	-	0.972	A	G	0	70	damaging	0.00	M326T	possibly damaging	1.258

<b>TLR10</b>	rs11096957	2.095	chr4	38452886	homozygous missense	0.653	0.347	-	-	T	G	0	61	<b>damaging</b> (low confidence)	0.00	N241H	<b>probably damaging</b>	2.641
<b>HGF</b>	-	3.285	chr7	81230076	heterozygous missense	na	na	na	na	G	R	24	27	tolerated	0.15	A46V	<b>probably damaging</b>	1.608
<b>IGSF3</b>	rs61786577	3.238	chr1	116948027	heterozygous missense	na	na	na	na	G	R	28	19	tolerated	0.06	R476C	<b>probably damaging</b>	2.888
<b>IGSF3</b>	rs61786651	2.450	chr1	116957982	heterozygous missense	na	na	na	na	C	Y	43	11	<b>damaging</b>	0.02	D254N	<b>probably damaging</b>	2.055
<b>FREM1</b>	rs7023244	3.278	chr9	14809370	heterozygous missense	-	-	0.819	0.181	G	K	21	15	-	-	-	<b>possibly damaging</b>	2.491
<b>HBXIP</b>	-	3.196	chr1	110751878	heterozygous missense	na	na	na	na	G	R	25	10	<b>damaging</b> (low confidence)	0.00	P45L	-	-
<b>NUAK1</b>	rs3741883	3.086	chr12	104985068	heterozygous missense	-	0.139	0.861	-	G	S	18	18	tolerated	0.15	P543R	<b>possibly damaging</b>	1.833
<b>ITGA3</b>	rs61730081	3.064	chr17	45506848	heterozygous new Met	na	na	na	na	G	R	22	11	tolerated	0.06	V474M	<b>probably damaging</b>	2.243

❖ follows the recessive model, ✱ verified by Sanger

Digital only table:

**Supplementary table 38 Genetic variants in conserved coding regions of genes.**

**Supplementary table 39 Functional associations of genes listed in Supplementary table 37.**

Functional descriptions are mainly from OMIM. Several SNVs were identified in genes involved in apoptosis (*UNC5C*, *LIMS1*, *ANP32A*, *BIRC7*, *DOCK5*, *MAP3K1*, *HGF*, *HBXIP*, *NUAK1*), T-cell development and function (*ITGAX*, *CD5*, *FURIN*, *LRMP*, *IGSF3*), pattern recognition (*TLR10*) and inflammatory signaling (*ILF3*, *IL22RA2*, *TILRR*).

gene symbol	function / OMIM (as hyperlinks)
<b><i>XRCC6</i></b>	High levels of autoantibodies to p70 and p80 have been found in some patients with systemic lupus erythematosus.[provided by RefSeq 2008].
<b><i>UNC5C</i></b>	Thiebault <i>et al.</i> (2003) hypothesized that the UNC5 netrin-1 receptors may also be dependence receptors that induce apoptosis when unbound to their ligand. <sup>144</sup>
<b><i>LIMS1</i></b>	Fukuda <i>et al.</i> (2003) concluded that PINCH1 is an obligate partner of ILK and both are indispensable for proper control of cell shape change, motility, and survival. <sup>145</sup>
<b><i>ITGAX</i></b>	Leukocyte surface antigen p150,95, plays a part in leukocyte adhesion [OMIM]
<b><i>ILF3</i></b>	In nonstimulated cells, NF90 was mostly nuclear, but T-cell activation resulted in its accumulation in the cytoplasm. The authors concluded that nuclear export of NF90 is required for IL2 mRNA stabilization. <sup>146</sup>
<b><i>IL22RA2</i></b>	Xu <i>et al.</i> (2001) concluded that IL22RA2 may be important as an IL22 antagonist in the regulation of inflammatory responses. <sup>116</sup>
<b><i>CD5</i></b>	human T-cell surface glycoprotein [OMIM]
<b><i>BEST2</i></b>	Best2 <sup>-/-</sup> mice exhibited enhanced inflammation, slow recovery from DSS-induced colitis, and altered mucin biogenesis raise the possibility that Best2 may play a role in inflammatory bowel diseases. <sup>109</sup>
<b><i>ANP32A</i></b>	Jiang <i>et al.</i> (2003) identified a pathway that regulates mitochondria-initiated caspase activity. In this pathway, PHAP protein promoted caspase-9 activation after apoptosome formation. <sup>147</sup>
<b><i>BIRC7</i></b>	expression of BIRC7 blocked apoptosis induced by engagement of FAS, tumor necrosis factor receptor-1 (TNFR1), death receptor-4 (DR4), and DR5), as well as apoptosis induced by chemotherapeutic agents, with a potency comparable to that observed with BIRC2 and BIRC4. <sup>148</sup>
<b><i>DOCK5</i></b>	Failure of DOCK5 signaling, together with p53 and TNFRSF10A/D-related cell cycle and death pathways, may play a critical role in abrogating apoptosis. <sup>149</sup>
<b><i>ECD</i></b>	Novel regulator of p53 stability and function. May also be a transcriptional activator required for the expression of glycolytic genes (GeneCards, UniProtKB/Swiss-Prot).
<b><i>FURIN</i></b>	Furin-deficient regulatory T cells were less protective in a T-cell transfer colitis model and failed to induce FOXP3 in normal T cells. <sup>114</sup>
<b><i>LRMP</i></b>	is expressed in a developmentally regulated manner in lymphoid cell lines and tissues <sup>150</sup>
<b><i>MAP3K1</i></b>	full-length mouse Mekk1 generates antiapoptotic signals, while a 91-kD C-terminal Mekk1 fragment induces apoptosis. <sup>151</sup>
<b><i>MMP9</i></b>	Two known gelatinases, MMP-2 and MMP-9, are upregulated during IBD. Epithelial-derived MMP-9 is an important mediator of tissue injury in colitis, whereas MMP-2 protects against tissue damage and maintains gut barrier function <sup>119</sup>
<b><i>NLRP12</i></b>	NALPs are implicated in the activation of proinflammatory caspases (e.g., CASP1) via their involvement in multiprotein complexes called inflammasomes (Tschopp <i>et al.</i> , 2003) [OMIM]
<b><i>SIGLEC1</i></b>	Serum IgM, but not IgG, titers were significantly decreased in Sn-deficient mice. Oetke <i>et al.</i> (2006) suggested that SN has a role in regulating cells of the immune system rather than influencing steady-state hematopoiesis. <sup>152</sup>
<b><i>TJP2</i></b>	Tight junction proteins
<b><i>TLR10</i></b>	TLR10 belongs to the Toll-like receptor (TLR) family of proteins homologous to the Drosophila Toll protein, which regulates hematopoiesis and innate immune responses to microbial pathogens in the adult fly. [OMIM]
<b><i>HGF</i></b>	Hepatocyte growth factor protects hepatoblastoma cells from chemotherapy-induced apoptosis by AKT activation <sup>153</sup>
<b><i>IGSF3</i></b>	The Ig-like domains are of the V type, and a search of the sequence databases revealed that the IGSF3 protein is 32% identical to human V7, a leukocyte surface protein. <sup>154</sup>
<b><i>FREM1/TILRR</i></b>	Data suggest that TILRR is an IL-1RI co-receptor, which associates with the signaling receptor complex to enhance recruitment of MyD88 and control Ras-dependent amplification of NF-kappaB and inflammatory responses <sup>120</sup>
<b><i>HBXIP</i></b>	suppressor of var1, 3-like 1 protein interacts with HBXIP, previously identified as a cofactor of survivin in suppression of apoptosis <sup>155</sup>
<b><i>NUAK1/ARK5</i></b>	ARK5 suppresses the apoptosis induced by nutrient starvation and death receptors via inhibition of caspase 8 activation. <sup>156</sup>
<b><i>ITGA3</i></b>	The results suggest that the production of MMP-9 by MKN1 cells was potentiated by the alpha3beta1 integrin-laminin-5 interaction, which facilitated their invasion via degradation of the matrix. <sup>118</sup>

**Supplementary table 40 sSVs in coding regions of other genes following the recessive model.**

Only a small number of sSVs was detected in exonic regions. Among these are small insertions in *ZNF717*, *REC8* and *TCEAL6* and small deletions in *MRPL18* and *KRTAP19-6*.

rsID	chr	start	end	length	type	ref allele	allele A	child		mother		father		transcript id	transcript symbol	type	SIFT causes nonsense mediated decay
								zyg. score	cov. cutoff	zyg. score	cov. cutoff	zyg. score	cov. cutoff				
-	chr3	75869245	75869245	2	insertion site	-	TG	0.0828	0.9394	0.9637	2.8387	0.9682	1.6818	NM_001128223	<i>ZNF717</i>	exonic	na
rs66936880	chr6	160131636	160131638	3	deletion	GTT	-	0.3176	1.1111	0.9721	2.1429	1.000	6.5000	NM_014161	<i>MRPL18</i>	exonic	yes
rs10690822	chr14	23716246	23716246	3	insertion site	-	GAA	0.4256	1.1538	0.9510	1.5833	0.9818	1.9615	NM_005132	<i>REC8</i>	exonic	yes
rs72241443	chr21	30835853	30835853	1	deletion	G	-	0.0039	0.0000	0.9842	3.4000	1.000	4.0800	NM_181612	<i>KRTAP19-6</i>	exonic	na
-	chrX	101282436	101282436	1	insertion site	-	G	0.0277	0.8182	0.9075	1.4138	0.5407	1.2000	NM_001006938	<i>TCEAL6</i>	exonic	na

zyg. score = experimental zygosity score; cov. cutoff = ratio of clipped normal coverage / number of non redundant reads

**Supplementary table 41 Summary of genes 5-fold up- or downregulated in the child compared to parents.**

All data derived from blood samples. Most of the genes upregulated in the child can be traced back to inflammatory functions, also some have known associations with IBD (*S100A12*, *MMP9*, *VNN1*, *TLR5*, *SLC2A3*, *MAPK3*, *MAPK14*, *TNFAIP6*, *VNN1*), which is also true for the downregulated genes (i.e. *A2M*, *TBET*, *PTGDR*, *FGFBP2*). Several of the differential expressed genes are associated with development and function of immune cells (*CD177*, *C19orf59*, *ARG1*, *CST7*, *TNFSF13B*, *IL4*) and antigen recognition and related downstream cascades (*CLEC4D*, *HMGB2*, *TLR5*, *NLRC4*, *IRAK3*, *CR1*).

gene	up/down	mother	father	son		function (hyperlinks to references, mainly from OMIM)	
		FPKM			x-fold of mother FPKM	x-fold of father FPKM	Possible immune-system relevant function / link to IBD
<i>CD177</i>	↑	2.149	1.531	118.097	54.97	77.14	Proteinase 3 (PR3) is a major autoantigen in anti-neutrophil cytoplasmic antibodies (ANCA)-associated systemic vasculitis (AASV), and the proportion of neutrophils expressing PR3 on their membrane (mPR3+) is increased in AASV. PR3 gene transcription is dependent upon <b>CD177</b> expression. <sup>157</sup>
<i>ARG1</i>	↑	1.005	1.532	34.678	34.51	22.63	Arginase 1 ( <b>ARG1</b> ) inhibits T-cell proliferation by degrading extracellular arginine, which results in decreased responsiveness of T cells to CD3/TCR stimulation. In humans, <b>ARG1</b> is stored in inactive form within granules of polymorphonuclear neutrophils. <sup>158</sup>
<i>S100A12</i>	↑	45.242	62.289	1326.530	29.32	21.30	Increased levels of circulating <b>S100A12</b> are found in IBD, compared to IBS. When used to distinguish IBD from IBS adult patients, serum <b>S100A12</b> levels exhibit moderate performance. <sup>121</sup>
<i>C19orf59</i>	↑	5.922	7.935	167.412	28.27	21.10	RT-PCR analysis showed that this gene is differentially expressed in mast cells. <sup>159</sup>
<i>PLSCR1</i>	↑	1.326	3.353	39.614	29.88	11.81	Herein, we provide evidence that PR3 was externalized during apoptosis and was associated with human phospholipid scramblase-1 ( <b>hPLSCR1</b> ), a protein that has been implicated in the bidirectional movement of plasma-membrane phospholipids in response to high cytosolic calcium levels, injury, or apoptotic insult. <sup>160</sup>
<i>SERPINB1</i>	↑	5.061	8.584	124.694	24.64	14.53	One function of <b>SERPINB1</b> may be to protect neutrophils from elastase and other serine proteases that are highly expressed in these cells. <sup>161</sup>
<i>IL4R</i>	↑	1.897	9.295	59.097	31.16	6.36	It had previously been demonstrated that gain-of-function polymorphisms in the <b>interleukin-4</b> gene are associated with increased output of <b>interleukin-4</b> , which in turn is associated with asthma, skin-test positivity, and higher total concentrations of serum IgE (OMIM). <sup>162</sup>
<i>S100A8</i>	↑	462.212	1070.270	11856.900	25.65	11.08	<b>S100A8</b> and <b>S100A9</b> induce a specific inflammatory pattern in endothelial cells which is characterised by the induction of a prothrombotic and pro-inflammatory response and an increased vascular permeability due to a loss of endothelial cell-cell contacts. <sup>123</sup>
<i>MMP9</i>	↑	1.757	1.499	28.963	16.48	19.32	<b>MMP-9</b> concentrations, but not those of TIMP-1, were significantly greater in the sputum supernatant in SA and CRD patients compared to controls. <sup>124</sup>
<i>GYG1</i>	↑	4.179	3.685	68.402	16.37	18.56	glycogen storage disease. <sup>163</sup>
<i>VNN1</i>	↑	3.639	2.881	52.793	14.51	18.32	<b>Vanin-1</b> is expressed by enterocytes, and its absence limits intestinal epithelial cell production of proinflammatory signals. <sup>164</sup>
<i>ANXA3</i>	↑	7.262	5.668	104.147	14.34	18.38	<b>Vnn1</b> <i>-/-</i> mice better controlled inflammatory reaction and intestinal injury. <sup>165</sup>
<i>TNFSF13B</i>	↑	1.736	5.359	40.461	23.30	7.55	Over-expression studies suggested <b>AnxA3</b> might be involved in death promotion during lactacystin-mediated neuronal death, since caspase-3 activation was significantly stronger upon neuronal <b>AnxA3</b> over-expression. <sup>166</sup>
<i>NLRC4</i>	↑	1.217	1.343	16.923	13.90	12.60	This cytokine is expressed in B cell lineage cells, and acts as a potent B cell activator. It has been also shown to play an important role in the proliferation and differentiation of B cells. [provided by RefSeq, Mar 2011]
<i>CLEC4D</i>	↑	3.123	8.749	56.205	18.00	6.42	Activation of caspase-1 through an <b>NLRC4</b> dependent pathway is closely associated with the subsequent death of the cell. <sup>167</sup>
<i>SLC37A3</i>	↑	2.496	1.008	16.504	6.61	16.38	may be an intracellular receptor for lipopolysaccharide and/or other bacterial products. <sup>168</sup>
<i>C5orf32</i>	↑	5.016	4.170	50.551	10.08	12.12	Functions as an endocytic receptor. May be involved in antigen uptake at the site of infection, either for clearance of the antigen, or for processing and further presentation to T cells. <sup>169</sup>
<i>PROK2</i>	↑	2.690	3.963	33.797	12.56	8.53	transmembrane sugar transporter
<i>C14orf94</i>	↑	2.363	1.206	16.733	7.08	13.88	KALLMANN SYNDROME 4
<i>DJ442471</i>	↑	8.187	2.750	42.219	5.16	15.35	Like other augmin subunits, fluorescence-tagged HAUS4 localized to the metaphase spindle and accumulated at centrosomes during interphase.
<i>ALPL</i>	↑	3.954	3.819	37.475	9.48	9.81	deficiency leads to Hypophosphatasia
<i>METTL9</i>	↑	6.184	6.892	59.821	9.67	8.68	
<i>CST7</i>	↑	17.574	18.113	162.629	9.25	8.98	
<i>CIR</i>	↑	1.064	1.114	9.510	8.94	8.54	<b>Cystatin F</b> is a recently discovered type II cystatin expressed almost exclusively in immune cells. <sup>170</sup>

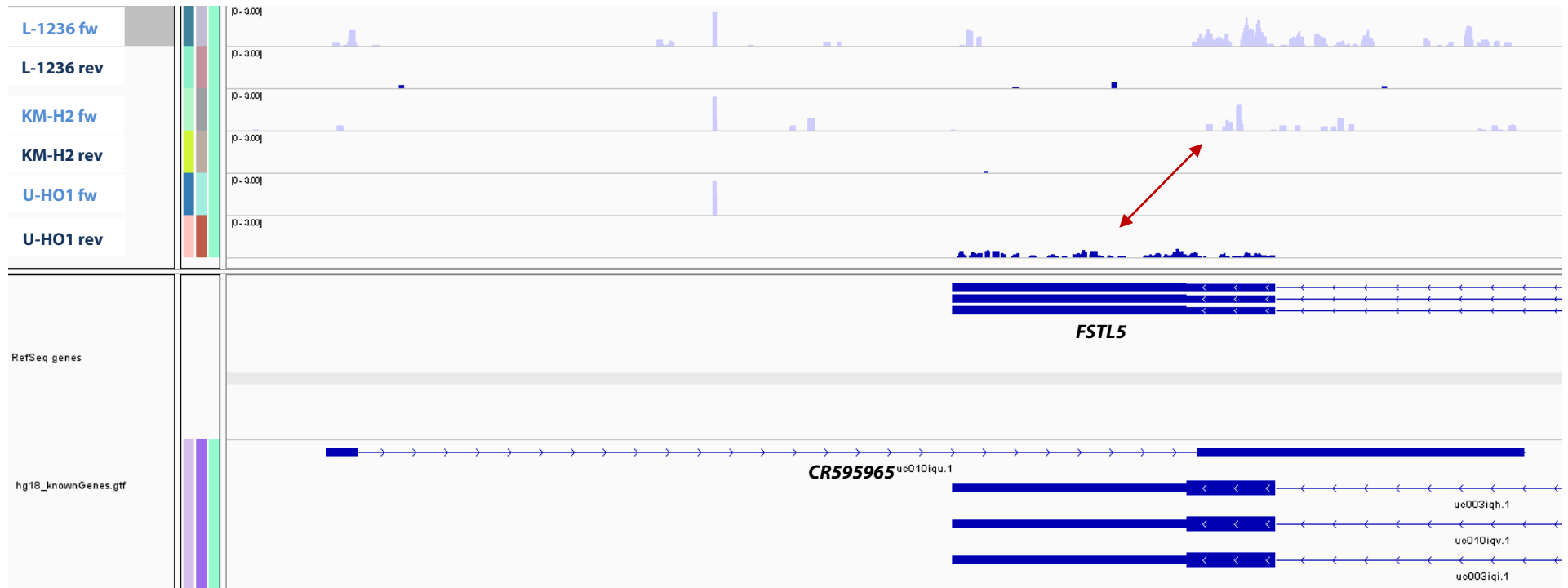
		mother	father	son		function (hyperlinks to references, mainly from OMIM)	
gene	up/down	FPKM			x-fold of mother FPKM	x-fold of father FPKM	Possible immune-system relevant function / link to IBD
<i>S100A9</i>	↑	659.149	1021.900	6943.040	10.53	6.79	<i>S100A8</i> and <b><i>S100A9</i></b> induce a specific inflammatory pattern in endothelial cells which is characterised by the induction of a prothrombotic and pro-inflammatory response and an increased vascular permeability due to a loss of endothelial cell-cell contacts <sup>123</sup>
<i>HK3</i>	↑	6.622	9.209	65.327	9.87	7.09	
<i>ACSL1</i>	↑	26.070	30.005	231.732	8.89	7.72	
<i>BCL2A1/BFL1</i>	↑	6.979	13.828	76.683	10.99	5.55	D'Sa-Eipper et al. (1996) showed that the <b><i>BFL1</i></b> protein suppresses apoptosis induced by the p53 tumor suppressor protein in a manner similar to other BCL2 family members. <sup>171</sup>
<i>ITGAM</i>	↑	4.384	2.692	27.381	6.25	10.17	Nath et al. (2008) identified and replicated an association between the <b><i>ITGAM</i></b> gene (120980) on chromosome 16p11.2 and risk of SLE in 3,818 individuals of European descent. <sup>172</sup>
<i>MAPK14/p38</i>	↑	4.591	7.930	47.577	10.36	6.00	<b><i>p38 mitogen-activated protein kinase</i></b> is activated and linked to TNF-alpha signaling in inflammatory bowel disease <sup>173</sup>
<i>CD36</i>	↑	2.087	1.914	16.197	7.76	8.46	It is the fourth major glycoprotein of the platelet surface and serves as a receptor for thrombospondin (188060) in platelets and various cell lines. [OMIM]
<i>RGS19</i>	↑	2.347	4.179	24.084	10.26	5.76	
<i>LEPROT</i>	↑	1.817	2.626	17.067	9.39	6.50	
<i>PGLYRP1</i>	↑	6.969	3.626	37.276	5.35	10.28	They concluded that PGRPs are innate immunity proteins that bind the cell wall or outer membrane and exploit the bacterial stress response to kill bacteria. <sup>174</sup>
<i>PGD</i>	↑	3.633	2.703	24.119	6.64	8.92	
<i>EMR1</i>	↑	1.718	2.252	15.165	8.83	6.73	
<i>HMGB2</i>	↑	3.415	6.009	33.701	9.87	5.61	The lack of HMGBs also resulted in poorer activation of Toll-like receptor-3 (Tlr3), Tlr7, and Tlr9 by their cognate nucleic acids (dsRNA, ssRNA, and hypomethylated DNA, respectively). <sup>175</sup>
<i>VCAN</i>	↑	2.916	3.055	22.646	7.77	7.41	
<i>UBE2H</i>	↑	1.237	1.411	9.838	7.95	6.97	
<i>FBXW7</i>	↑	1.094	1.063	7.894	7.22	7.43	
<i>SLC2A3</i>	↑	8.282	7.720	57.301	6.92	7.42	up-regulated in CD including <b><i>SLC2A3</i></b> <sup>176</sup>
<i>PPP4R2</i>	↑	1.422	2.350	12.647	8.90	5.38	
<i>ARNTL</i>	↑	1.518	1.702	11.117	7.32	6.53	
<i>CR1</i>	↑	2.696	2.696	18.215	6.76	6.76	Microbes as well as immune complexes and other continuously generated inflammatory particles are efficiently removed from the human circulation by red blood cells (RBCs) through a process called immune-adherence clearance. During this process, RBCs use <b>complement receptor 1 (CR1)</b> , CD35) to bind circulating complement-opsonized particles and transfer them to resident macrophages in the liver and spleen for removal. <sup>128</sup>
<i>YWHAB</i>	↑	3.038	3.583	22.125	7.28	6.18	
<i>BC073144</i>	↑	2.180	3.191	16.913	7.76	5.30	
<i>H2A/I</i>	↑	1.597	1.512	10.102	6.33	6.68	
<i>TNFAIP6</i>	↑	3.454	2.526	18.519	5.36	7.33	Class prediction analysis of CDc top 20 and top 5 significant genes allowed complete separation between CDc responders and CDc nonresponders. The CDc top 5 genes were <b><i>TNFAIP6</i></b> , <i>S100A8</i> , <i>IL11</i> , <i>G0S2</i> , and <i>S100A9</i> <sup>125</sup>
<i>CKAP4</i>	↑	4.961	6.191	34.281	6.91	5.54	
<i>PYGL</i>	↑	12.157	17.722	89.512	7.36	5.05	GLYCOGEN STORAGE DISEASE VI
<i>MAPK3</i>	↑	1.051	1.169	6.851	6.52	5.86	Pages et al. (1999) concluded that p44 Mapk apparently has a specific role in thymocyte development. <sup>177</sup>
<i>STK4</i>	↑	3.329	3.973	21.952	6.59	5.53	The particular phosphorylation catalyzed by this protein [STK4] has been correlated with apoptosis. [provided by RefSeq, Jul 2008]
<i>PIK3AP1/BCAP</i>	↑	4.126	4.166	24.274	5.88	5.83	Yamazaki and Kurosaki (2003) found that <b><i>Bcap</i></b> <i>-/-</i> mice had altered Nfkb (see 164011) activity with poor expression of Rel (164910) and, to a lesser extent, of RelA (164014), leading to reductions in target gene induction, cell survival, and cell division. <sup>178</sup>
<i>IRAK3/IRAKM</i>	↑	6.217	7.353	38.173	6.14	5.19	This contrasts <sup>130</sup> with endotoxin tolerance, in which IRAK-M up-regulation follows proinflammatory cytokine expression caused by LPS exposure
<i>CASP3/PPP32</i>	↑	1.397	1.495	7.878	5.64	5.27	Nicholson et al. (1995) developed a potent peptide aldehyde inhibitor and showed that it prevents apoptotic events in vitro, suggesting that apopain/ <b><i>PPP32</i></b> is important for the initiation of apoptotic cell death. <sup>179</sup>
<i>BGR/DECTIN1</i>	↑	2.050	2.011	10.948	5.34	5.44	Brown and Gordon (2001) identified dectin-1 as a beta-glucan receptor present on macrophages. <sup>129</sup>
<i>TLR5</i>	↑	1.451	1.437	7.654	5.27	5.33	indicating that TLR5-stop can protect persons of Jewish ethnicity against CD <sup>127</sup>
<i>HBA1</i>	↓	5083.940	220.249	43.593	116.622	5.052	
<i>PIGR</i>	↓	11.569	8.004	1.557	7.429	5.140	mediates transcellular transport of polymeric immunoglobulin molecules. [OMIM]
<i>ISCU</i>	↓	14.357	8.848	1.463	9.812	6.048	Tong and Rouault (2006) concluded that <b><i>ISCU</i></b> is involved in a coordinated response to iron deficiency that includes activation of iron uptake, redistribution of intracellular iron, and decreased utilization of iron in Fe-S proteins <sup>180</sup>

		mother	father	son			function (hyperlinks to references, mainly from OMIM)
gene	up/down	FPKM			x-fold of mother FPKM	x-fold of father FPKM	Possible immune-system relevant function / link to IBD
<i>SEPT9</i>	↓	15.087	9.597	1.373	10.988	6.990	<b>Septin</b> depletion resulted in chromosome loss from the metaphase plate, lack of chromosome segregation and spindle elongation, and incomplete cytokinesis upon delayed mitotic exit. <sup>181</sup>
<i>FIBP</i>	↓	7.755	7.294	1.359	5.706	5.367	
<i>BTF3</i>	↓	13.194	14.744	1.014	13.017	14.547	suppresses CED-3-independent apoptosis in <i>c. elegans</i> <sup>182</sup>

**Supplementary table 42 Exonic missense SNVs detected in genes differentially expressed in the child compared to the parents.**

Only five differentially expressed genes show potential damaging SNVs (*CR1*, *VCAN*, *IL4R*, *EMR1* and *MMP9*). Only the *MMP9* SNV is located in a site of high conservation (phyloP 2.604), but is supported only by a very limited number of reads.

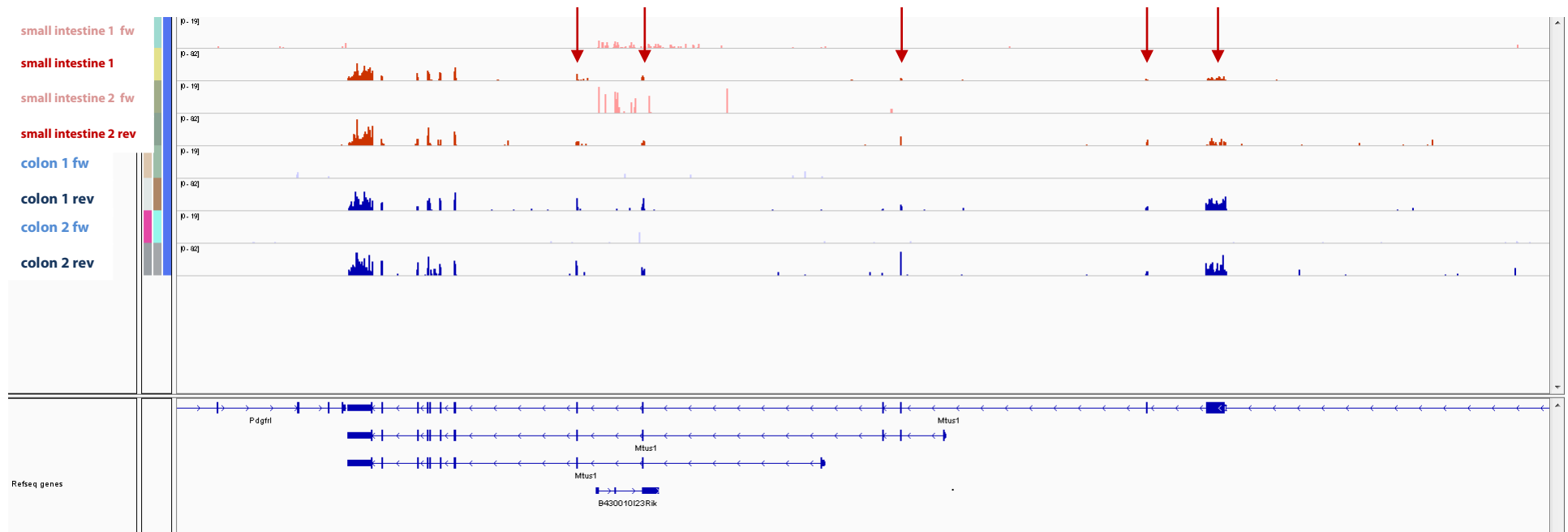
rsID	gene symbol	chrom	position	phyloP	subtype	zygosity	ref	allele A	allele B	reads ref	reads alt	SIFT	SIFT score	SIFT AA change	PolyPhen	dScore
rs2274567	<i>CR1</i>	chr1	205820244	0.394	missense	hom	A	G	G	0	31	tolerated	0.32	H1663R	<b>probably damaging</b>	1.538
rs309559	<i>VCAN</i>	chr5	82869125	0.903	missense	het	A	G	A	21	25	<b>damaging</b> (low confidence)	0.00	K1516R	benign	0.251
rs1805011	<i>IL4R</i>	chr16	27281373	-0.489	missense	het	A	A	C	38	18	tolerated	0.36	E400A	<b>possibly damaging</b>	0.764
rs1805015	<i>IL4R</i>	chr16	27281681	1.019	missense	het	T	T	C	34	22	tolerated	0.28	S503P	<b>possibly damaging</b>	1.839
rs897738	<i>EMR1</i>	chr19	6852891	1.113	missense	hom	G	A	A	0	62	tolerated	0.43	D174N	<b>possibly damaging</b>	0.754
-	<i>MMP9</i>	chr20	44076428	2.604	missense	het	A	A	C	3	2	tolerated	0.11	K638T	<b>probably damaging</b>	1.892



**Supplementary figure 5** Differentially expressed S/AS pair in cHL cell lines.

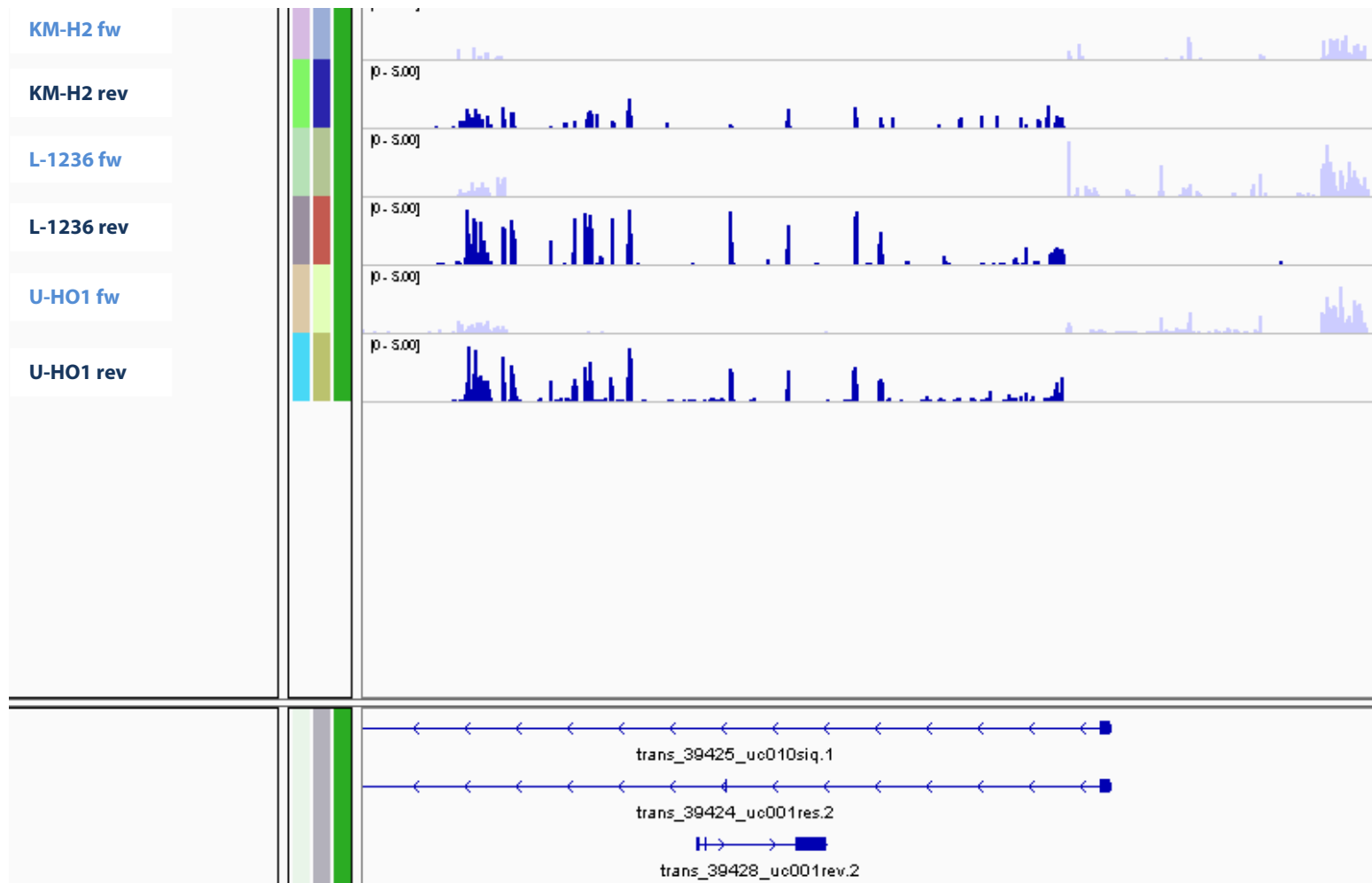
Expression of *uc010iqu.1* (*CR595965*, light blue) is mutually exclusive in L-1236 and KM-H2, while *uc003iqi.1* (*FSTL5*, dark blue) is only expressed in U-HO1. Scale: 0-3 for all tracks.





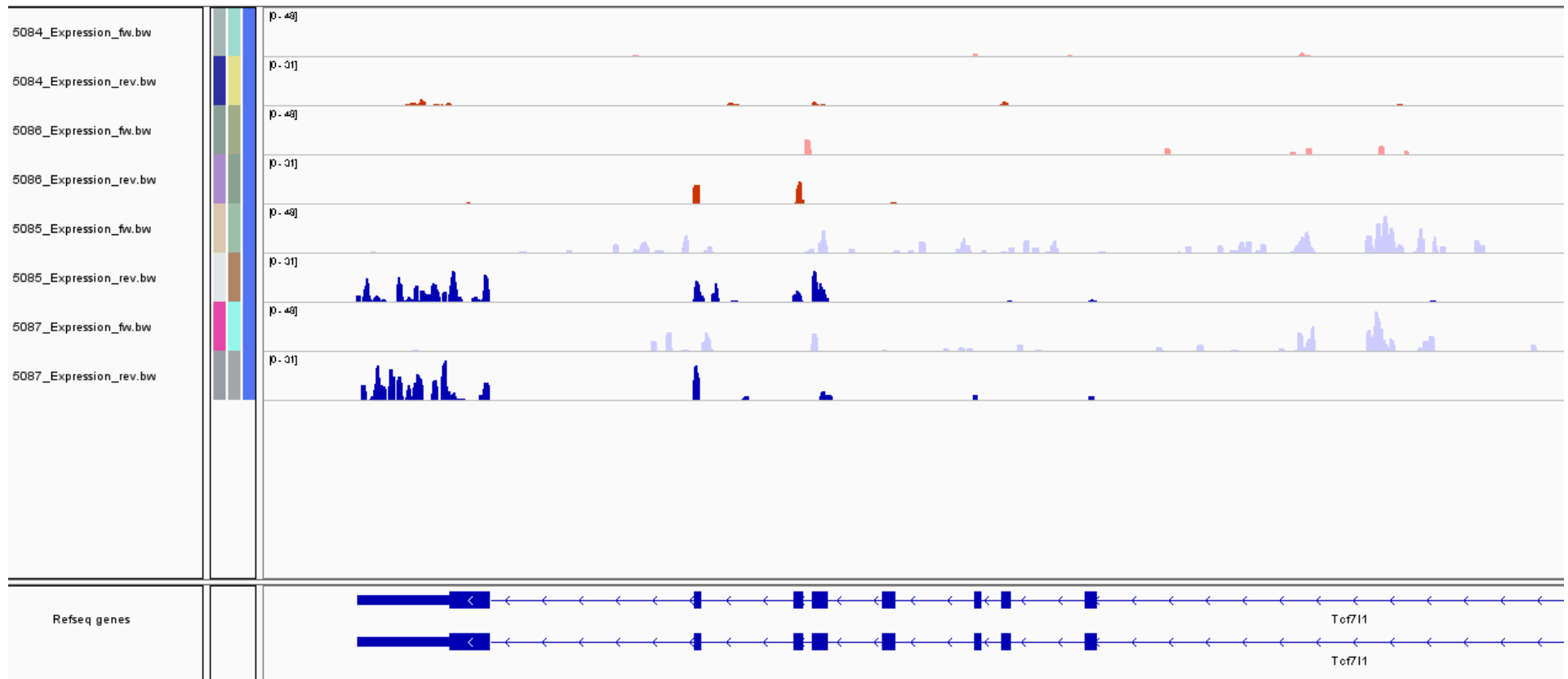
**Supplementary figure 6 Differentially expressed antisense event in *Mtus1* in murine tissues.**

*B430010123Rik* is expressed in the two small intestine samples (red) but not in colon (blue). Expression of (the long) *Mtus1* isoform is lower in the small intestine samples (exons marked by arrows). Scale for forward strand: 0-19, scale for reverse strand: 0-82.



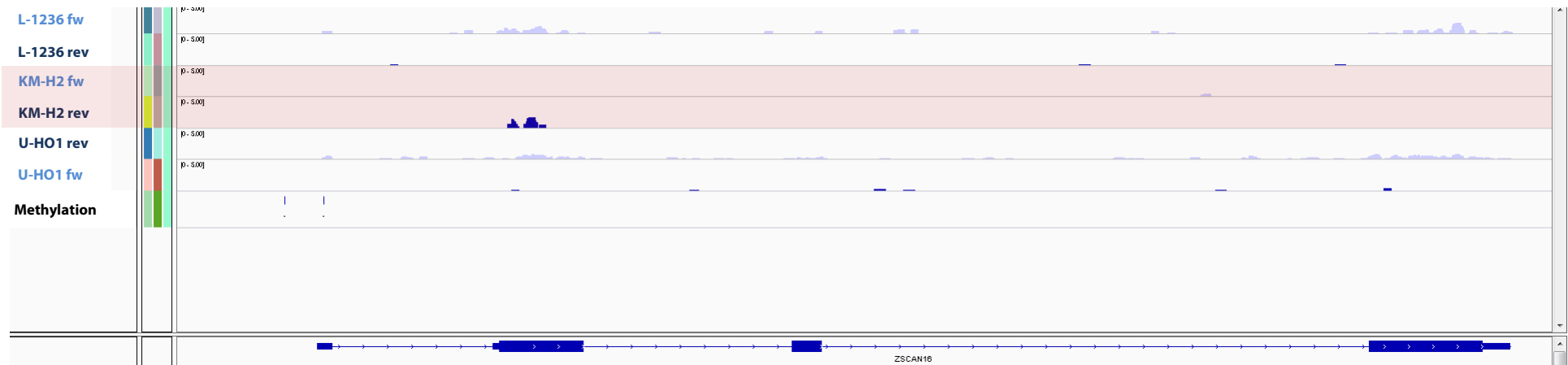
**Supplementary figure 7** Coexpressed S/AS pair in CHL cell lines.

A transcript in antisense orientation to *uc010siq.1* (light blue) correlates with the expression of the sense transcript (dark blue). Scale 0-5 for all tracks.



**Supplementary figure 8** Coexpressed S/AS example in *Tcf7l1* in murine tissues.

A novel transcriptionally active region is visible in both colon samples (blue), which also show expression of *Tcf7l1*, while neither is expressed in the small intestine samples (red). Scale 0-49 for forward and 0-31 for reverse strand.



**Supplementary figure 9** Differentially expressed antisense event in ZSCAN16 in cHL cell lines.

ZSCAN16 expression (light blue) was detected in samples L-1236 and U-HO1, but not in KM-H2. Instead KM-H2 shows antisense expression (dark blue) of a small unannotated region. Methylation data is available for the two marked positions (last track), where KM-H2 shows a higher degree of methylation than the other two samples (see Supplementary table 43). Scale 0-5 for all tracks.

**Supplementary table 43** Average beta methylation values for ZSCAN16 (ZNF435) promoter region.

Methylation was measured in duplicates for L-1236 and KM-H2. The methylation rate in KM-H2 is higher than in the other two samples. KM-H2 is also the only sample that shows expression on opposite strand of ZSCAN16, exon 2.

TargetID	SYMBOL	CHR	MAPINFO	L-1236 AVG_Beta	KM-H2 AVG_Beta	U-HO1 AVG_Beta
cg12706983	ZNF435	6	28200218	0,1541	0,5401	0.0433
cg13665593	ZNF435	6	28200399	0,0680	0,4470	0.0198