Evaluation of susceptibility genes for inflammatory bowel disease by association study and candidate gene analyses

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität

zu Kiel

vorgelegt von

Weiyue Zheng M.sc

Referent: Prof. Dr. h.c. Thomas C. G. Bosch

Koreferent/in: ……

Tag der mündlichen Prüfung: .....

Zum Druck genehmigt: Kiel, ...

Der Dekan

Abbreviations and symbols

AA	amino acid
AbD	assay(s) by design (Applied Biosystems, available at https://www.store.appliedbiosystems.com)
AGR2	human homologues anterior gradient proteins 2
AGR3	human homologues anterior gradient proteins 3
AoD	assay(s) on demand (Applied Biosystems, available at https://www.store.appliedbiosystems.com)
ASP	affected sibling pair
bp	base pairs
CD	Crohn's disease
cDNA	complementary DNA
cM	centiMorgan
CU	Colitis ulcerosa
°C	degree Celsius
χ^2	chi-square, measure of association or independence
D'	D prime (measure of LD)
dATP	2'-deoxyadenosine-5'-triphosphate
dCTP	2'-deoxycytidine-5'-triphosphate
DDW	double-distilled water
dGTP	2'-deoxyguanosine-5'-triphosphate
DNA	deoxyribonucleic acid
dNTP	2'-deoxynucleotide-5'-triphosphate
dsDNA	double-stranded deoxyribonucleic acid
dTTP	2'-deoxythymidine 5'-triphosphate
EDTA	ethylenediaminetetraacetic acid
FAM	6-carbofluorescein
Fig.	figure
Figs	figures
g	gram(s)
x*g*	relative centrifugal force (RCF)
h	hour(s)
HWE	Hardy-Weinberg-Equilibrium
IBD	Inflammatory bowel disease
ibd	identity by descent
ibs	identity by state
IFN	interferon (e.g. IFN-g)
kb	kilobase
kD	kilodalton
l	liter(s)
LD	linkage disequilibrium
LDU	linkage disequilibrium units
LOD	logarithm of odds
M	molar (mol/l)
Mb	Mega base

mg	milligram(s)
min	minute(s)
ml	millilitre(s)
MLE	maximum likelihood estimate
mM	millimolar (mmol/l)
MRC	Molecular Research Center
mRNA	messenger RNA
µg	microgram(s)
µl	microlitre(s)
µM	micromolar (µmol/l)
ng	nanogram(s)
nM	nanoMolar (nmol/l)
NPL	nonparametric linkage
OR	odds ratio
p	probability
p. a.	pro analysis
PCR	polymerase chain reaction
pH	potentia hydrogenii (hydrogen ion concentration)
pmol	picomol
r.t.	room temperature (ca. 20°C)
rmax	maximum radius (centrifuge parameter)
Rn	fluorescent emission of the normalised reporter dye
RNA	ribonucleic acid
rpm	rotations per minute (centrifuge parameter)
s	second(s)
SNP	single nucleotide polymorphism
Taq-polymerase	*Thermophilus aquaticus* DNA polymerase
TaqMan®	commercial name for sequence variation detection assay, using 5'→3' exonuclease activity
TaqMan®-MGB-	commercial name for sequence variation detection assay, using 5'→3' exonuclease activity and 3' minor groove binding probes
TBE	Tris borate EDTA
TCR	T-cell receptor
TDT	Transmission disequilibrium test
TE	Tris EDTA
Tm	melting temperature
Tris	tris-(hydroxymethyl)-aminomethane
UC	ulcerative colitis
UV	ultraviolet (light)
VIC	trade name for fluorescent dye

Amino acid symbols

A	Alanine
C	Cysteine
D	Aspartic acid
E	Glutamic acid
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
K	Lysine
L	Leucine
M	Methionine
N	Asparagine
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threonine
V	Valine
W	Tryptophan
Y	Tyrosine

DNA base nomenclature

A	Adenine
G	Guanine
C	Cytosine
T	Thymine

1. Introduction

1.1 Inflammatory bowel disease

1.1.1 Basic concept on inflammatory disease

1.1.1.1 Barrier organs

Barrier organs (skin, mucosa) are of pivotal importance for the defense of the organism against facultative pathogens of a hostile environment. Under normal conditions, such barriers are maintained without any inflammation. However, for most of the barrier organs chronic inflammatory disorders have been described (skin: atopic dermatitis, psoriasis; gastrointestinal mucosa-Crohn's disease, ulcerative colitis (inflammatory bowel disease); lung mucosa: asthma, sarcoidosis; oral mucosa: periodontitis. A large overlap in clinical features and in molecular pathophysiology between some of the diseases is seen. A polygenic susceptibility is documented for all inflammatory barrier diseases through genetic epidemiology (twin studies, large families) but also molecular findings (disease genes, linkage findings). A remarkable overlap in linkage findings suggests that some of the diseases may involve the same disease genes. This hypothesis could be substantiated by the analysis of CARD15, the first disease gene in Crohn's disease, that plays also a role in asthma, psoriatric arthritis and familiar sarcoidosis (Blau syndrome), although different sequence variants are relevant in the latter condition.

1.1.1.2 Inflammation as a common response to environmental triggers

Inflammation is one of the core reaction mechanisms of the human organism to environmental stimuli. While phylogenetically designed to combat infection, the same pathomechanisms are important in a number of autoimmune inflammatory disorders. Following the cloning algorithm from phenotype to genotype, these disorders share some key features that make a joint genomic analysis highly attractive:

1. All diseases are epidemiologically connected to a western, industrialized lifestyle and show an increasing incidence in these countries. They share some of the environmental factors (in particular disturbances in the bacterial flora on body surfaces) that are also amenable to genomic approaches.
2. The histological appearance (mucosa associated inflammation) and cellular activation patterns are similar. For instance, an increased expression of inflammatory cytokines (TNF-α, IL-6, IL-10, IFN-γ) and activation of common signal transduction pathways (NFκB, JAK-STAT systems, MAP kinases) has been demonstrated for asthma, IBD, psoriasis, sarcoidosis and atopic dermatitis. This makes them particularly suited to a joint genomic analysis of disease pathophysiology.
3. Most importantly, several genome-wide linkage studies have defined common chromosomal susceptibility regions that make a joint analysis (including the joint fine mapping, development of common makers sets, identification of candidate genes) scientifically highly attractive.
4. These entities also have many clinical commonalities, including overlap of the phenotypes and similar treatment strategies (i.e. immunosuppression using corticosteroids

and other agents). A true integration of phenotypic and genotype data, therefore calls for an integrated, interdisciplinary phenotyping using complex clustering algorithms.

1.1.1.3 Bacterial flora on body surfaces as a disease factor

Several arguments can be made to emphasize the critical role of bacterial populations on body surfaces in the manifestation of the autoimmune disease investigated in this proposal. These arguments include:

1. Epidemiological associations with childhood hygiene and antigen exposure exist for many of these disorders. The bacterial flora is clearly modified by behavioral changes influencing immune responses not only locally, but also systemically. Thus, many target organs (lung, skin and intestine) are affected by the exposure to bacteria.
2. The intestinal immune system (over 70% of the total immune cells of the body) exists in close interaction with the intestinal bacteria. A well-known example is the induction and maintenance of tolerance to the ABO blood group antigens.
3. Genetic animal models of intestinal inflammation demonstrate that germ-free conditions can completely abrogate the development of inflammatory disease. Early exposure to bacterial antigens attenuates later inflammatory responses to non-pathogenic bacterial challenges. Furthermore, it seems that sensitization to inhaled allergens is abrogated in animal models of asthma, if endotoxin exposure is completely absent. On the other hand, very high loads of microbial produces also seem to be protective against the development of asthma and atopic diseases.
4. A close association between exposure to microbial matter and the development of asthma and atopic dermatitis has been shown in children of farmers. The more early exposure to

microbes the higher is expression levels of innate immunity pathogen receptors and the lower is the risk to develop asthma.

Probiotic treatments have resulted in positive studies under evidence-based conditions (i.e. placebo controlled trials) in IBD and atopic dermatitis.

1.1.2 Inflammatory bowel disease

1.1.2.1 Pathogenesis and pathophysiology of IBD

Inflammatory bowel disease (IBD) refers to a complex chronically relapsing autoimmune disorder of the gastrointestinal tract of unknown etiology. It is a chronic relapsing destructive disease, which affects the intestine and other organs (including joints, liver, pancreas, skin, eyes) and which results in a severe impact on the health and economic capacities of the adolescents affected. The disease was first described in the 1930ies and has risen in incidence in Europe and the US since World War II. New data from Iceland indicate that increases in incidence have not stopped. Current lifetime prevalence is estimated at 0.5-1%[1] [2]. More importantly, it is thought that other unclear functional bowel disorders (IBS-irritable bowel syndrome, which affects up to 30% of the population > 40 years) can to a certain extent be attributed to abortive forms of IBD (i.e. resulting in disturbed bowel function but not the characteristic ulcerations). IBD is also considered as a prototype, which will help in the understanding of other chronic inflammatory disorders (e.g. rheumatoid arthritis) where the phenotype is less clear and the disease organ is less accessible. It has been classified into Crohn's disease (CD) and ulcerative colitis (UC) based on clinical, radiologic, endoscopic and pathological criteria [3] (Table 1.1.2.1). IBD is characterized by a chronic relapsing activity

with problem-free phases ("remission") intervening with phases of inflammatory activity ("flair"). The clinical features include abdominal pain, (bloody) diarrhea and complications such as groth retardation in children, anemia, toxic megacolon and stenosis and fistulae (in Crohn disease). While some patients develop a chronically active disease in which inflammatory activity of different degrees is always present, others come to a complete clinical remission between active episodes. The immunological causes responsible for the different types of disease behavior are unclear and the triggers for the development of relapses are unknown. Extra-intestinal manifestations can also occur, including skin ulcers, arthritis, and bile-duct inflammation, the last especially in UC. Both UC and CD are characterized by mucosal ulceration. In CD, ulcers penetrate into the gut wall, and fistulous tracts may develop between loops of bowel or to the skin. IBD is a multifactorial disease caused by the interplay of genetic, environmental and immunological factors [4]. Multiple studies have suggested that first-degree relatives of an affected patient have a risk of inflammatory bowel disease that is 4 to 20 times as high as that among the background population; the absolute risk of inflammatory bowel disease is approximately 7 percent among first-degree family members [5] [6]. Epidemiological and genetic linkage studies in IBD provide a thorough proof for a genetic background [7] [8], and a first disease gene (NOD2 or CARD15) has been identified for CD [9] [10] [11]. The lifestyle of an industrialized society is important for the penetrance of the genetic factors [12] (Figure 1.1.2.1).

Table 1.1.2.1 Differences between Crohn's disease and Ulcerative Colitis

Disease	Ulcerative Colitis	Crohn's disease
Symptoms	Diarrhea Bloody Diarrhea Pus or mucus in the stool Severe abdominal cramps Nausea Frequent fever	Diarrhea Rectal bleeding Wight loss Pain & tenderness in abdomen, especially the lower right side Low-grade fever Sometimes constipation because of a blockage Slowed growth and delayed sexual development in some childhood cases
Parts of digestive system affected	Only the top layers of the walls of the colon or rectum (most often in the lower part of the colon and rectum)	Deep in the lining of the walls of the colon and/or small intestine. Any part of the digestive tract from mouth to anus

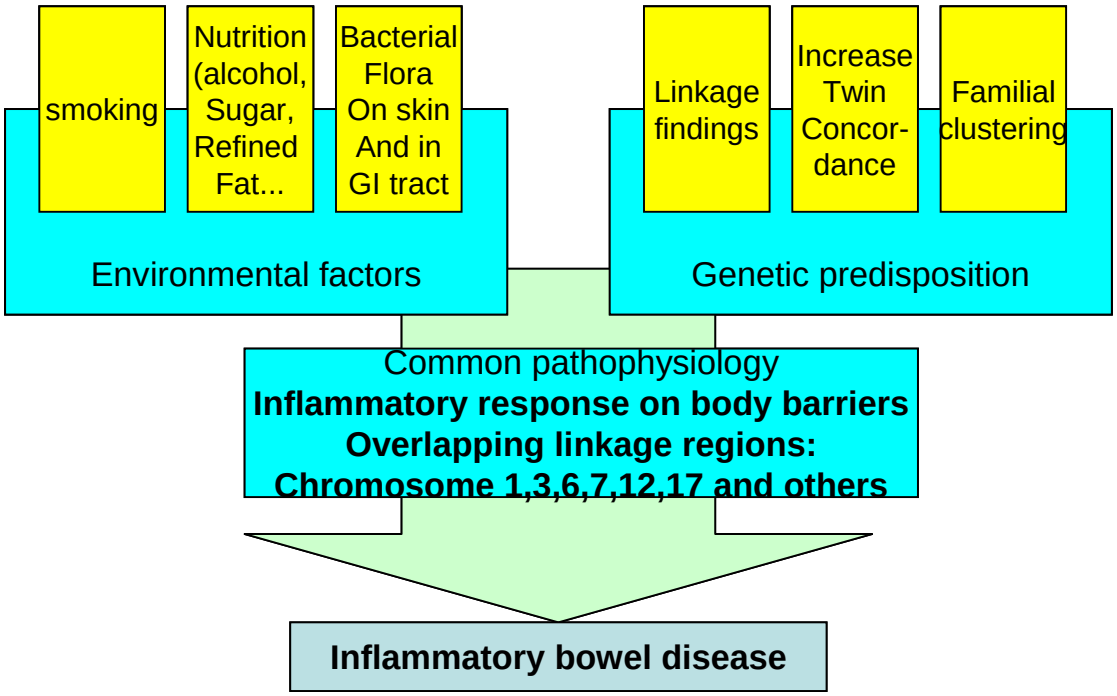


Fig 1.1.2.1 Possible causes of inflammatory bowel disease

The disorder of Crohn's disease is characterized by transmural inflammation that could affect any part of the gastrointestinal tract, and the disease relapses and remits throughout its course. In CD inflammation can arise anywhere in the digestive tract from the mouth to the anus. Aggregates of immune cells consisting of T cells, monocytes and macrophages in the bowel wall can be observed in half of the cases. The structure of the mucosa is distorted and shows progressive atrophy. A simple cause and effect relation probably does not account for most cases of Crohn's disease. The pathogenesis of the disease is complex and consists of three interacting elements: genetic susceptibility factors, priming by the enteric microflora, and immune-mediated tissue injury [13 14 15].

Ulcerative colitis is a worldwide, chronic, idiopathic inflammatory disease of the rectal and colonic mucosa. Inflammation in UC starts in the rectum and extends proximally up the bowel. The damage of the mucosa is continuous from the rectum to the proximal colon without evidence of skip lesions. The inflammation is located predominantly within the lamina propria of the mucosa. Scattered crypt abscesses with ulceration and islands of regenerating mucosa forming pseudopolyps are observed. A severe course of UC is the formation of a toxic megacolon, a sudden cessation of bowel function leading to toxic dilatation and eventual perforation of the bowel. As only the inner lining of the intestine is affected in UC, nearby organs are not affected by the formation of fistulae. Both CD and UC can lead to numerous extra-gastrointestinal inflammatory manifestations in the liver, eyes, skin and joints [3].The gut lamina propria is constantly exposed to antigens and pathogens coming along with the nutrition. After a primary inflammatory process microbes and microbial products may infiltrate the submucosa or lamina propria, resulting in a reinforced inflammation leading to an increased mucosal damage [16].

When the state of inflammation shifts from the physiologic to the pathogenic state, the pro- and anti-inflammatory cytokines become imbalanced [17 18]. In CD, a predominant TH1 response with elevated levels of IL-2 and IFN was observed in studies [19 20]. In UC, humoral immunity with an increased level of IL-5 and IL-10 seems to predominate, but evidence for a classical TH2 situation is scarce [4 21]. In both CD and UC, activated macrophages participate in the mucosal immune response, *e.g.* by producing pro-inflammatory cytokines such as TNF-a, IL-1, and the chemokine IL-8 [22 23 24]. It has been suggested that TNF-α plays a central role in the pathogenesis of CD and is likely to be at the apex of the inflammatory cascade [18 20 22 25]. In almost half of UC affected individuals and in a small group of CD cases, a perinuclear

antineutrophil cytoplasmic antibody (p-ANCA) can be detected. It seems to be more prevalent in more aggressive UC. P-ANCA reactivity is suggested to derive from the recognition of heterogeneous neutrophil-associated antigens [26] . A sub group of the antibody, called atypical p-ANCA, recognizes a 50 kD myeloid-specific nuclear envelope protein that can be detected in UC affected carriers of the antibodies [27].

1.1.2.2 Genetic background of inflammatory bowel disease

Epidemiological and family studies have provided overwhelming evidence that genetic factors have an important role in determining susceptibility to IBD [28]. Systematic investigations of affected families show relative sibling risks of 15-35 for Crohn's disease [29] [30] and 6-9 for UC [6] [31] [32]. This and an increased concordance of the IBD phenotype in monozygotic twins [5] [33] clearly show a genetic cause of the disease. In 25% of multiply affected families cases of CD and UC are present [34]. Several susceptibility loci were identified: IBD1 on chromosome 16 [35] [36]; IBD2 on chromosome 12[35]; IBD3 on chromosome 6 [37] [38]; IBD4 on chromosome 14[39] [40]; IBD5 on chromosome 5[41] [38]; chromosome 1 [42]; chromosome 19[38] and as well on chromosome X [43] (table 1.1.2.2). Most subsequent studies supported a polygenic mode of inheritance, and it has since become clear that environmental factors represent significant risk modifiers [44].

Table 1.1.2.2 Gene map locus for inflammatory bowel disease

Locus name	Region
IBD1	16q12
IBD2	12q13
IBD3	6p
IBD4	14q11-q12
IBD5	5q31
IBD6	19p13
IBD7	1p36
IBD8	16p

Within the linkage region on chromosome 16 (IBD1) a first disease gene CARD15 (or NOD2) has been identified. Variants of the NOD2 gene (C14772T (R702W) in exon 4, G25386C (G908R) in exon 8 and 32629insC (1007insC)) are highly associated with CD [9 10 11], but cannot account for all cases. Among other functions, NOD2 may be involved in NF-kB signalling [16]. Besides this first gene, the multiple linkage regions and segregation models indicate that more than one risk allele is involved in the pathogenesis of IBD [45 46]. A stronger association exists between genes of the human leucocytes antigen region and UC [47]. A positive association with DR2 and the rare alleles DRB1*0103 and DRB1*12 and negative association with DR4 and DRw6 have been reported [4 47].

1.1.2.3 Linkage region on chromosome 12

A candidate region on chromosome 12 (the IBD2 locus [MIM 601458]) has been reported for both CD and UC [8,35,48-51], but some studies claimed to exclude this region [52-54]. Miles Parkes et al [55] demonstrated that IBD2 appears to make a major contribution to UC susceptibility but to have only a relatively minor effect with regard to CD. A number of plausible candidate genes have been investigated with negative results, including the beta 7 integrin gene [56] and the STAT-6 gene [57].

1.1.2.4 Linkage region on Chromosome 7

Suggestive evidence for linkage at chromosome 7q near D7S669 was initially reported in a European cohort consisting of 186 affected sibling pairs from 160 families [35]. Several candidate genes on this region had been evaluated [58-60].

1.2 Globlet cell function

The epithelial mucosal layer is a very important barrier that can protect the animal body from dryness, harmful exogenous substances and pathogens. Mucus is secreted by the epithelial surfaces throughout the gastrointestinal tract from the stomach to the colon. It forms a gel adherent to the surface that provides a protective barrier between the epithelium and the exterior environment. Mucus serves many functions, including protection against shear stress and chemical damage. The mucus layer on top of the intestinal epithelium is the barrier between the host's internal milieu and gut bacteria [61]. Mucins are the primary component of the mucus. They are glycoproteins that are deemed to mediate many interactions between these cells and their milieu [62]. Mucins and other components of mucus are secreted from the apical surface of specialized columnar epithelial cells referred to as goblet cells [63].

Goblet cells are distributed among other cells in the epithelium of many organs, especially in the intestinal and respiratory tracts. Goblet cells reside throughout the length of the small and large intestine and are responsible for the production and maintenance of the protective mucus blanket by synthesizing and secreting mucins. Goblet cells have a characteristic

morphology, based on membrane-bound secretory granules, which contain mucus [64].

The goblet cells' function is the secretion of mucins and other products, including protease resistant peptides--like the trefoil peptide family, which protects epithelium from injury and promotes repair through restitution of epithelial cells [65]. Secretion of mucus occurs by exocytosis of secretory granules [66]. After the secretion, mucins have the ability to form a viscous gel, producing a protective scaffold overlaying epithelial surfaces.

Epithelial cell differentiation in mucosal tissues has been studied to some detail in the gastrointestinal tract endoderm and the bronchial airways [67 68]. The small intestinal epithelium consists of four principal cell types: enterocytes, goblet, enteroendocrine and Paneth cells [69].

When the mucus production is altered, it can cause various diseases. As in asthma, chronic obstructive pulmonary disease (COPD), and cystic fibrosis, the mucus productions are increased, whereas in dry eye syndrome, gastric disease, peptic ulcer, and inflammatory bowel disease are decreased. Altered mucus production is also described in malignancies like colorectal cancer [70 71 72].

1.3 Mouse model

A mouse phenotype with diarrhea and goblet cell dysfunction caused by anterior gradient protein 2 dysfunction was reported (European patent WO2004056858). In the patent, the authors performed a genome wide screen for mutations influencing epithelial functions in mice, e.g. nutrient absorption by intestinal mucosa. The responsible mutation was identified

by positional cloning and shown to result in an amino acid exchange within "anterior gradient 2" (*AGR2*) by the Mouse Genome Informatics database. This gene is expressed in murine intestinal tissues, specifically in intestinal goblet cells [73]. The human homologous gene is "Anterior gradient 2 homolog" (*AGR2*). Comparing to the mouse *AGR2* gene, it encodes a protein with 91% amino acid identity.

They analyzed the RNA expression profile of the mouse *Agr2* gene and the human *AGR2* gene. The phenotype observed in the mouse model demonstrates that *AGR2* function is required for normal goblet cell function in a mammalian model organism. In the patent, they provide mutated *AGR2* nucleic acids and polypeptides having modified sequences compared to the wild type sequences. In a specific embodiment, an *AGR2* mutein carries an amino acid substitution at residue 137. The amino acid substitution is the substitution of a codon encoding valine at position 137 to a codon encoding a non-valine (glutamic acid) substitution. They use gene knockout method to produce a mouse model that only contains mutated *AGR2* gene.

The goblet cells referred to in this study are specialized with respect to mucus secretion via granules, in particular in the gastrointestinal tract (GI) (examples in this regard are goblet cells of the esophagus, of the stomach surface, of the pyloric glands, and of the intestinal epithelium), or in the respiratory tract (examples in this regard are goblet cells of the nose epithelium, of the trachea, of the bronchius, and of the submucosal glands of the trachea).

The patent demonstrates that *AGR2* is required for normal goblet cell function and that mutating this gene and its gene product may result in goblet cell dysfunction and

corresponding physiological and medical disorders of the affected animal. The results of this experiment indicate that human *AGR2* mRNA is strongly expressed in stomach, duodenum, ileocecum, ileum, descending colon, transverse colon, caecum, and rectum. Weaker expression is detected in lung, cervix, and prostate. *AGR2* is a secreted protein released from cultured colon cancer cells. Western blot analysis predicted that any mutation in the *AGR2* gene resulting in abnormal *AGR2* peptide expression levels in an individual will interfere with the peptide's normal biological function, including in a manner analogous to that observed in the present invention. Mutations leading to abnormal *AGR2* peptide expression levels might affect any aspect of gene expression, e.g. DNA transcription, mRNA transport and processing, mRNA translation or *AGR2* peptide half-life itself.

The patent demonstrates for the first time that *AGR2* is required for normal goblet cell function, in particular mucin secretion.

1.4 *AGR2* function

The cement gland is an ectodermal organ in the head of frog embryos, lying anterior to any neural tissue. The cement gland was induced by the dorsal mesoderm. Mesoderm with the highest cement gland-inducing potential lay posterior to the ectoderm fated to form this organ, indicating that its induction occurred at a distance from the inducer source. Cement gland induction first occurred during early gastrulation.

The functional analysis of a protein homologue to human *AGR2* has been performed in Xenopus laevis. The overexpression of XAG-2 induces both cement gland differentiation and expression of anterior neural marker genes in the absence of mesoderm formation [74]. The Xenopus protein exhibits 59% amino acid identity to mouse *AGR2* protein, and exhibit 60% amino acid identity to human *AGR2* (Fig 1.4a and Fig 1.4b).

```
AGR2_Homo          AAAGGACACAAAGGACTCTCGACCCAAACTGCCCCAGACCCTCTCCAGAGGTTGGGGTGA
AGR2_Bos           GAAGGACACGAAGGACTCTCAACTCAAACTGCCCCAGACCCTCTCCAGAGGTTGGGGGGA
AGR2_Mus           AAAGGACCCAAAGGACTCTCGGCCCAAACTACCTCAGACACTCTCCAGAGGTTGGGGCGA
AGR2_Rattus        AAAGGACCCAAAGGACTCTCGACCCAAACTACCCCAGACCCTGTCCAGAGGTTGGGGAGA
AGR2_Danio         GAAGGA---GAAGGA------AAAGAGAGTTCCACAGACTCTCTCCAGAGGATGGGGTGA
AGR2_Xenopus       GTCGGTCAGAAAAGA------AATCCGGGCCCCGCAGACACTATCAAGAGGTTGGGGAGA
                      **     ** **                ** ***** ** ** ***** ***** **

AGR2_Homo          CCAACTCATCTGGACTCAGACATATGAAGAAGCTCTATATAAATCCAAGACAAGCAACAA
AGR2_Bos           CCAACTCATCTGGACCCAGACATATGAAGAAGCTTTATATAAATCCAAGACAAGCAATAA
AGR2_Mus           TCAGCTCATCTGGACTCAGACATACGAAGAAGCTTTATACAGATCCAAGACAAGCAACAG
AGR2_Rattus        TCAGCTCATCTGGACTCAGACTTACGAAGAAGCCTTATACAAATCCAAGACAAGCAACAG
AGR2_Danio         TCAGCTGATTTGGGCACAGACATACGAGGAAGCTCTGTTTTGGTCACGATCCAAGAACAA
AGR2_Xenopus       TGATATCTCATGGGTGCAAACATATGAAGAAGGACTTTACAATGCAAAGAAAAGAAATAA
                     *  *    ***   ** ** ** ** ****   * *      *       *  ** *

AGR2_Homo          ACCCTTGATGATTATTCATCACTTGGATGAGTGCCCACACAGTCAAGCTTTAAAGAAAGT
AGR2_Bos           ACCCTTGATGATTATTCACCACTTGGATGAATGCCCACACAGTCAAGCTTTAAAGAAGGC
AGR2_Mus           ACCCTTGATGGTCATTCATCACTTGGACGAATGCCCACACAGTCAAGCCTTAAAGAAAGT
AGR2_Rattus        ACCCTTGATGGTCATTCATCACTTGGACGAATGCCCGCACAGTCAAGCTTTAAAGAAAGT
AGR2_Danio         GCCCCTCATGGTCATCTTTCACCTGGAAGACTGTCCACACAGCCAGGCTCTGAAGAAAGC
AGR2_Xenopus       GCCCTTAATGGTAATTCATCATTTAGAAGATTGTCAGTATTGCCAAGCATTGAAGAAAGT
                    *** * *** * **    **  * ** ** ** * *  * * ** **  * ***** *

AGR2_Homo          GTTTGCTGAAAATAAAGAAATCCAGAAATTGGCAGAGCAG---TTTGTCCTCCTCAATCT
AGR2_Bos           ATTTGCTGAAAATAAAGAAATCCAGAGATTGGCAGAGCAG---TTTGTCCTCCTCAATCT
AGR2_Mus           GTTTGCTGAACATAAAGAAATCCAGAAATTGGCAGAGCAG---TTTGTTCTCCTCAACCT
AGR2_Rattus        GTTTGCTGAAAATAAGGAGATCCAGAAATTGGCAGAGCAG---TTTGTTCTCCTCAACTT
AGR2_Danio         ATTTGCTGAGGATAAAGAAATCCAGAAGTTGGCTGATGAAGACTTTGTGATCTTGAACTT
AGR2_Xenopus       ATTTGCAGAAAGTGATGAAGCTCAGACATTAGCCCAAGAGCAATTCATAATGCTCAACCT
                    ***** **   * * **    ****  ** **  *  *    **  *  *  * **  *

AGR2_Homo          GGTTTATGAAACAACTGACAAACACCTTTCTCCTGATGGCCAGTATGTCCCCAGGATTAT
AGR2_Bos           AGTTTATGAAACAACTGACAAACACCTCTCTCCTGATGGCCAGTATGTGCCCAGGATTTT
AGR2_Mus           GGTCTATGAAACAACCGACAAGCACCTTTCTCCTGATGGCCAGTACGTCCCCAGAATTGT
AGR2_Rattus        GATCTATGAAACAACTGACAAGCACCTTTCTCCTGATGGCCAGTACGTCCCCAGAATTGT
AGR2_Danio         GGTGTACGAAACCACAGATAAGCACTTGTCTCCTGATGGCCAGTACGTTCCCAGAATCAT
AGR2_Xenopus       TATGCATGAAACAACTGACAAAAACCTTTCCCCTGATGGACAGTATGTGCCTCGGATAAT
                     *  * ***** ** ** **  ** * ** ******** ***** ** **  * **  *
```

Fig 1.4a Nucleotide sequences of *AGR2* cDNA and alignment with other species using clustal X [75].

```
AGR2_Homo       MEK-IPVSAFLLLVALSYTLAR--------DTTVKPGAKKDTKDSRPKLPQTLSRGWGDQ
AGR2_Mus        MEK-FSVSAILLLVAISGTLAK--------DTTVKSGAKKDPKDSRPKLPQTLSRGWGDQ
AGR2_Xenopus    MQTGLSLACLVLLCSVLGEAALRKPKRQAGATDTNGAAKSEPAPVKTKGLKTLDRGWGED
                *:. :.::.::** ::    *          * .: .**.:.   :.*  :**.****::

AGR2_Homo       LIWTQTYEEALYKSKTSNKPLMIIHHLDECPHSQALKKVFAENKEIQKLA-EQFVLLNLV
AGR2_Mus        LIWTQTYEEALYRSKTSNRPLMVIHHLDECPHSQALKKVFAEHKEIQKLA-EQFVLLNLV
AGR2_Xenopus    IEWAQTYEEGLAKARENNKPLMVIHHLEDCPYSIALKKAFVADKMAQKLAQEDFIMLNLV
                : *:*****.* ::: .*:***:****::**:* ****.*. .*  **** *:*::****

AGR2_Homo       YETTDKHLSPDGQYVPRIMFVDPSLTVRADITGRYSNRLYAYEPADTALLLDNMKKALKL
AGR2_Mus        YETTDKHLSPDGQYVPRIVFVDPSLTVRADITGRYSNRLYAYEPSDTALLYDNMKKALKL
AGR2_Xenopus    HPVADENQSPDGHYVPRVIFIDQSLTVRSDLKGRYGNKLYAYDADDIPELITNMKKAKSF
                : .:*:: ****:****::*:* *****:*:.***.*:****:. * . *  ***** .:

AGR2_Homo       LKTEL
AGR2_Mus        LKTEL
AGR2_Xenopus    LKTEL
                *****
```

Fig 1.4b Amino acid sequences of *AGR2* cDNA and alignment with other species using clustal X [75].

Human *AGR2* (also termed BCMP7 and XAG-1) is a known protein. There are several patents which study the elements (such as growth factor, PRO 1030, EST sequence) whose response to *AGR2* can cause colon cancer (WO 98/07749, WO 99/55858, WO 00/53755). Human *AGR2* mRNA was shown to be expressed in trachea, lung, stomach, colon, prostate and small intestine [76].

1.5 Disease gene detection in complex disease

1.5.1 Linkage analysis

Epidemiological evidence shows that many diseases have genetic background and it is important to identify the genes that carry the disease causing mutation. There is more than 20,000 genes are known and the amount of genes that is estimated for the human genome is about 30,000 to 50,000 genes [77] [78] [79]. The purpose of genome-wide linkage analysis is to reduce the area of the genome in which such a gene could be situated. In complex disease, linkage analysis is performed to detect a co-segregation between a marker and a putative disease locus. In linkage analysis, co-segregation of two or more genes is examined in a family unit to determine if they segregate independently according to Mendel's laws or if they do not segregate independently because of their close physical proximity. If the two homologous chromosomes segregate independently, alleles at loci on the same chromosome should co-segregate at a rate (θ) which is somehow related to the distance between them on the chromosome. This rate is the probability of a recombination event occurring between the two loci. Multiple recombination events can occur on the same chromosome. Two loci are said to be genetically linked when $\theta \leq 0.5$, and the phenomenon describing this occurrence is genetic linkage. The object of linkage analysis is to estimate θ and to test if $\theta \leq 0.5$. Recombination between each of the genetic markers and the disease "marker" is determined based on the frequency each allele was observed. The maximum likelihood that the observed data are caused by a determined recombinant fraction is tested. The likelihood for each value of the recombinant fraction between pairs of markers is compared and thereby the odds ratio and the logarithm of the odds ratio (LOD score) are determined. The LOD score indicates the likelihood of linkage [80]. The most probable position of the disease "marker" is between those 2 genetic markers where the smallest recombination frequency is measured, indicating the strongest linkage and the shortest genetic distance. The maximum likelihood estimate (MLE) is the value that gives rise to the largest value of the likelihood (or LOD score) [81]. The MLE

is highly efficient, in the sense of the precision obtained using a given set of data. With increasing sample size it will converge closer to the true value.

The usual likelihood-based approaches for genetic linkage analysis require that a correct model be specified for the relationship between an individual's genotype and the corresponding chance of displaying a disease or other trait phenotype. The non-parametric linkage (NPL) analysis methods, like the affected-sib-pair (ASP) method, do not require that the genetic model explaining disease inheritance be explicitly specified. The NPL is usually measured in the analysis of a qualitative trait.

Multipoint linkage analysis is the usual method where several markers at the same time are used to map the disease locus. The analysis of the polymorphic markers is performed in affected sibling pairs (ASP). They share the disease allele through identity by descent (ibd). With the determination of the allele status at the polymorphic marker, recombination can be detected between the marker and the disease allele. The closer the disease allele is linked to the marker, the less recombination can be expected and the higher the LOD score is. Taking the possibility of false positive results in account, a critical value of > 3.6 for the LOD score were suggested [82]. However, LOD scores depend on the density and information content of the marker, the size of the ASP population and preciseness of the disease phenotype [83].

1.5.2 Association mapping analysis

Linkage analysis locates the approximate position of genes. The association design takes a known gene and tests whether individual differences in the gene are statistically associated with a phenotype. A linkage region can be several Mb long and contain several hundred

genes. For association analysis, the density of markers is increased (3 to 50 kb) and usually diallelic single nucleotide polymorphisms (SNPs) are employed as markers. Association studies can be family or population based, resulting in two different analysis methods: the transmission disequilibrium test (TDT) for family based studies [84] and the case control analysis for population-based studies.

In a population based study with the case control set-up, the Pearson's χ_2 statistic (or any other appropriate χ_2 statistic) is calculated. P value calculated from association studies must be corrected for the number of loci analyzed and the number of alleles at each loci. Number of samples required depends on the number of markers tested, the population frequency of the susceptibility allele and the relative risk of the susceptibility allele [85].

The transmission disequilibrium test (TDT) was developed by Spielman [86], as a method to test for linkage in the presence of association. The original version of the TDT used a McNemar chi-square to assess preferential transmission of allele A to affected offspring from AB heterozygote parents. If allele A is transmitted significantly more than half of the time, it can be concluded that the AB marker locus is linked to a disease predisposition locus, and that allele A is positively associated with an allele that increases disease susceptibility. Different methods to evaluate the transmission to the affected offspring have been developed. Either transmitted alleles are compared direct to untransmitted alleles or the difference to the number of alleles expected to be transmitted according to HWE is analyzed [87] [88] [89]. TDT association studies are considered to be more sensitive (having greater power) than linkage methods to detect genes with small to moderate effects on disease [89]. The TDT is only a valid

test of association if one affected offspring per family is used. It is always valid as a test of linkage, but its power is low when association is weak or absent.

1.5.3 Candidate gene analysis

The systematic positional cloning approach outlined above is extremely time and cost consuming. The direct evaluation of functional candidate genes is one possibility to directly identify the causative molecular variants. The chances for success of this process depend critically on the pathophysiology-based choice of the candidate gene. In addition, expression studies together with the chromosomal localization of a transcript can be used to define "positional candidate genes". Differentially regulated transcripts identified through expression profiling are another important contributor to the selection of candidate genes.

In fact, once disease related disequilibrium is identified in a genomic region, the identified transcripts have to be also tested in a candidate gene approach. Study design is a critical factor as case-control studies are susceptible to false positive findings due to population stratification. Performing case-control studies, special care needs to be taken in order to ensure a genetic matching of study cohorts. A robust alternative is the use of family based association designs like the haplotype relative risk (HRR) or transmission disequilibrium test (TDT). Furthermore, the replication of results in different, independent and large cohorts is thought to be critical to assess the true value of association results. The choice of the candidates to be studied is driven by the prevailing immunological theories and restricted to known genes. However, in recent years the knowledge about disease pathways has increased dramatically and genetic studies of disease pathways, e.g. in the innate and adaptive immune system, start to replace single gene association studies. Thus, gene-gene interactions are

increasingly taken into account in a more systematic approach of extended haplotype studies. Furthermore, gene expression profiling and comparative studies between murine and human disease models add to the potentials of candidate gene approaches. Carefully conducted candidate gene studies remain a powerful tool, also in view of the methodological problems outlined for the positional cloning method. In many situations, like for instance the search for disease modifying genes or in pharmacogenetics, case-control candidate gene studies will remain the only viable instrument. In general variants in the candidate genes will either be taken from the public SNP databases or will be generated in-house by re-sequencing of affected individuals.

The analysis of candidate genes is a key step in strategies for disease gene identification. In simple diseases, linkage and positional cloning can be used. But for complex diseases, these methods have not been successful. The failure is the result of three main features of complex disease: (1) the diseases vary in severity of symptoms and age of onset; (2) they can vary in their aetiological mechanisms; (3) they are more likely to be caused by several genes, each with a small overall contribution and relative risk. Candidate-gene studies focus on genes that are selected because of a priori hypotheses about their aetiological role in disease. Furthermore, a candidate-gene study is usually conducted in a population-based sample of affected and unaffected individuals. Therefore, candidate-gene study takes advantage of both the increased statistical efficiency of association analysis of complex diseases and the biological understanding of the phenotype, tissues, genes and proteins that are likely to be involved in the disease.

The selection of candidate genes has many parallels with identifying and ranking risk factors in an epidemiological study. Candidate genes may be identified based on functional information about the disease and/or on genetic map information obtained by linkage or linkage disequilibrium. Candidate genes are genes of known biological action involved with the development or physiology of the trait biological candidates. They may be structural genes or genes in a regulatory or biochemical pathway affecting trait expression. Some genes may be excellent candidate genes based on similar phenotypes seen in other species. Expression studies might provide important information about the tissues and cells that are involved in the disease [90 91 92]. The chosen number of candidate genes and variants is influenced by many considerations [93].

The criticisms of candidate-gene approaches are rooted in a fundamental challenge to the study of the genetics of complex diseases. Some of the candidate-genes cannot be replicated in different populations. Lack of replication may be caused by different reasons. First, the finding of association studies often cannot be replicated [94]. Second, discrepant findings are often due to the variations in study design. Third, the selections of polymorphisms are not likely to be causal. Most candidate genes considered only a small number of candidate genes and variants.

A polymorphism is a variation in DNA sequence that has an allele frequency of at least 1% in a population. Most of the DNA sequence variation in the human genome is in the form of single nucleotide polymorphism (SNP) [95]. The statistical power to detect a significant association depends on the size of the association and the frequency of the allele of interest [85,93,96]. The SNP with very low allele frequencies would need to have very large relative risks

associated with them to be detected in a candidate gene. SNPs with frequencies of at least 5% are generally more likely to be useful in a candidate-gene study [97]. Another important consideration in selecting SNPs is whether there is significant linkage disequilibrium in the candidate gene in the study population. The use of rigorous epidemiological principles for the choice and analysis of candidate genes and SNPs in disease studies is one tool that might improve the chance of a successful outcome.

1.5.4 Pathway-mapping to establish the link between genetic variants and pathophysiology

Positional cloning results and positive candidate gene studies lead to the identification of sequence variants that are strongly associated with disease and therefore suggested as possible causative factors. However, the proof of a causative relationship requires functional studies detailing disease mechanisms. Likewise, only functional studies can explore the pathways originating from disease genes and their uplink into disease pathophysiology. Most of the genes identified lack a full gene structure or a functional annotation, respectively. Detailed functional analysis and evaluation of biological relevance in the context of barrier function and chronic inflammation requires a systematic approach using advanced molecular and bioinformatics tools.

The molecular tools for pathway mapping have to address a wide range of mechanisms observed in disease genes described or under study in this network (from clear structural changes due to amino-acid exchanges or truncating mutations (e.g. an insertion mutation causing a frame shift as demonstrated for NOD/CARD15-SNP13) to more complex effects including alterations in mRNA stability, promoter activity or regulation of alternative splicing

(e.g. IL-13, STAT6). The mappping of pathways originating from disease genes will ultimately reveal new therapeutic targets in the uplink to more generic mechanisms in disease pathophysiology. For this purpose explorative techniques to systematically analyze expression regulation (i.e. microarrays, real time PCR) and to detect protein-protein interactions are used.

Pathway mapping integrates functional genomical approaches generated to assign a cellular function to a molecular variant ultimately leading to a disease phenotype. The connections have been established to gain access to key technologies including systematic splicing and allelic expression analysis, RNAi, high-throughput yeast two hybrid, proteomics (2DGE, MALDI-TOF), HTS-reporter gene assays, ChIP on Chip and standardized generation of stably transfected cell lines will be used to address key questions that evolve from the identified disease genes. As an important element, pathway-mapping needs to combine a series of highly standardized technological platforms and resources that provide complex material like tissue from animal models and patients available in a systematic fashion.

1.6 Aims of this study

1. Evaluation of *AGR2* and *AGR3* as candidate genes for inflammatory bowel disease.
2. Identification of susceptibility marker for inflammatory bowel disease on chromosome 12q linkage region.

2. Materials and methods

2.1 Materials

Table 2.1 Materials part 1

Material	Manufacturer / Supplier
100 bp DNA ladder	Invitrogen; Karlsruhe, Germany
2.2 ml storage plate (96 well)	ABgene, Epsom, UK
384 deep well storage plate	ABgene, Epsom, UK
Agarose	Eurogentec; Köln, Germany
AmpliTaq® DNA Polymerase	Applied Biosystems; Weiterstadt, Germany
AmpliTaq® Gold DNA Polymerase	Applied Biosystems; Weiterstadt, Germany
Bacillol®	Bode Chemie; Hamburg, Germany
BigDye Terminator Ready reaction kit	Applied Biosystems; Weiterstadt, Germany
Bromophenol blue	Sigma; München, Germany
Cell culture flasks (250 ml; canted neck)	BD Biosciences; Heidelberg, Germany
Cryotubes (2ml)	Greiner Bio-One GmbH; Frickenhausen, Germany
DNAzol®	Molecular Research Center, inc. Cincinatti, Ohio, USA
dNTP set (100mM solutions @ 100µM each)	Amersham Biosciences; Freiburg, Germany
Easy peel heat seal foil	ABgene, Epsom, UK
EDTA	Sigma; München, Germany
EDTA blood vial 9 ml	Sarstedt; Nümbrecht, Germany
Ethanol p. a.	Merck; Darmstadt, Germany
Ethanol technical	Bundesmonopol für Branntwein (BfB); Offenbach, Germany
Ethidium bromide solution (10mg/ml)	Invitrogen; Karlsruhe, Germany
ExoI (Exonuclease I)	Amersham Biosciences; Freiburg, Germany
GeneAmp PCR buffer system (10x buffer w/o $MgCl_2$; 25mM $MgCl_2$ solution)	Applied Biosystems; Weiterstadt, Germany
Glycerol	Sigma; München, Germany
Invisorb Blood Giga Kit	Invitek, Berlin, Germany
isopropanol	Merck; Darmstadt, Germany
$MgCl_2$	Merck; Darmstadt, Germany
MicroAmp® optical 96 well reaction plate	Applied Biosystems; Weiterstadt, Germany
MicroAmp® single strips	Applied Biosystems; Weiterstadt, Germany
MicroAmp® single tubes	Applied Biosystems; Weiterstadt, Germany
Microtiter 384 well plates	Sarstedt; Nürnberg, Germany

Table 2.1 Materials part 2

Material	Manufacturer/Supplier, Country
Microtiter 384 well plates	Greiner Bio-One GmbH; Frickenhausen, Germany
Microtiter 96 well plates	Sarstedt; Nürnberg, Germany
Microtiter 96 well plates	Costar Corning Incorporated; Cambridge, MA, USA
Microtiter plates, 96-well, round bottom, with lid	Sarstedt; Nümbrecht, Germany
Microtiter strips (Nunc-Immuno Maxisorp™ , 8-well, flat bottom) with 96-well frame	Nunc; Wiesbaden, Germany
Multiscreen column loader	Amersham Biosciences; Freiburg, Germany
PEQLAB DNA isolation system	PEQLAB Biotechnology GmbH; Erlangen, Germany
PicoGreen Molecular Probes	Europe BV, Leiden, The Netherlands
Pipette (serological, sterile with filter 5 / 10 / 25 ml)	Sarstedt; Nümbrecht, Germany
Pipette tips with filter (10 / 200 / 1,000 µl)	Sarstedt; Nürnberg, Germany
protein-kinase K	Invitek, Berlin, Germany
proteinase K	Molecular Research Center, inc. Cincinatti, Ohio, USA
saccharose	Merck; Darmstadt, Germany
SAP shrimp alkaline phosphatase	Amersham Biosciences; Freiburg, Germany
Sephadex powder (G50 superfine)	Amersham Biosciences; Freiburg, Germany
Sephadex spin column plates MAHVN 4550	Amersham Biosciences; Freiburg, Germany
SmartLadder DNA marker	Eurogentec; Köln, Germany
TAE Buffer 25x ready pack	Amresco; Solon, OH, USA
TaqMan ® Universal PCR Master Mix	Applied Biosystems; Weiterstadt, Germany
TBE buffer 10x ready pack	Amresco; Solon, OH, USA
TEMED	Sigma; München, Germany
Tris	Merck; Darmstadt, Germany
Triton-X	Sigma; München, Germany
Trypsin / EDTA (0.25% / 1 mM)	Invitrogen/Gibco; Karlsruhe, Germany
Tubes (0.5 / 1.5 / 2.0 mL)	Eppendorf; Köln, Germany
Tubes (0.5, 1.5, 2 mL)	Eppendorf; Köln, Germany
Tubes, flat bottom (60 mL)	Sarstedt; Nümbrecht, Germany
Tubes, sterile (15 mL)	Sarstedt; Nümbrecht, Germany
Tubes, sterile (50 mL)	BD Biosciences; Heidelberg, Germany
Xylene Cyanol FF	Sigma; München, Germany

2.2 Electronic database

Applied Biosystems	http://www.appliedbiosystems.com/
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
Celera Discovery Systems	http://www.celera.com/
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/index.html
Gene Cards	http://www.genecards.org/
Locus Link	http://www.ncbi.nlm.nih.gov/LocusLink/
MAP viewer	http://www.ncbi.nlm.nih.gov/mapview/
NCBI	http://www.ncbi.nlm.nih.gov/
PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
UCSC Genome Browser	http://genome.ucsc.edu/
Repeat Masker	http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker

2.3 Participants and study design

2.3.1 Investigated patient sample of AGR gene

In this study, two groups of participants were investigated. The first group included a sample of 383 cases from families and 565 trios with IBD (631 with CD, 317 with UC) and 537 unrelated healthy control individuals of German extraction. The second group consisted of 604 cases from families and 91 trios (384 with CD and 311 with UC) and 360 unrelated healthy control individuals from the United Kingdom. For each affected individual, the diagnosis of either CD or UC was confirmed by standard diagnostic criteria [3] [98]. Ascertainment criteria were determined prior to the initiation of patient collection. German patients and their family members were recruited at the 1st Department of Medicine at the Hospital Schleswig-Holstein, Campus Kiel (Kiel, Germany), the Charité University Hospital (Berlin, Germany) and other collection centers in Germany. The control individuals were collected through the Department of Transfusion Medicine at the University Hospital Schleswig-Holstein and through the POPGEN population project (www.popgen.de). UK families were sampled through the King's College School of Medicine, Guy's Hospital, and the St. Mark's Hospital London (UK). The cohorts have been used in a number of previous studies and the clinical and ethical issues of the recruitment process were reviewed as part of these publications [99] [50] [11] [46]. An overview of the investigated sample is given in table 2.3.1.

Table 2.3.1: Overview of the investigated cohort: Non-overlapping categories are given. Single cases were randomly selected from IBD families.

Population	Cases						Controls
	CD			UC			
	Trios	Cases from families	Total independent cases	Trios	Cases from families	Total independent cases	
Germany	377	254	631	188	129	317	537
UK	56	328	384	35	276	311	360

2.3.2 Association study population on chromosome 12

In this study, one group of participants was investigated. This group included a sample of 776 trios with IBD (484 with CD, 292 with UC) and 360 unrelated healthy control individuals of German extraction. For each affected individual, the diagnosis of either CD or UC was confirmed by standard diagnostic criteria [3 98]. Ascertainment criteria were determined prior to the initiation of patient collection. German patients and their family members were recruited at the 1st Department of Medicine at the Hospital Schleswig-Holstein, Campus Kiel (Kiel, Germany), the Charité University Hospital (Berlin, Germany) and other collection centers in Germany. The control individuals were collected through the Department of Transfusion Medicine at the University Hospital Schleswig-Holstein and through the POPGEN population project (www.popgen.de). The cohorts have been used in a number of previous studies and the clinical and ethical issues of the recruitment process were reviewed as part of these publications [99 50 11 46].

2.3.3 Sequencing samples

For the sequencing of genomic DNA, the DNA was extracted from peripheral blood lymphocytes as described in the DNA isolation section and arrayed into a 96 well plate format. Each well contained 5 ng of liquid DNA. The samples consisted of unrelated German individuals affected with IBD from the population sample for association studies. For the mutation detection of the AGR gene, one IBD population sample, which was composed of 47 unrelated German individuals (24 with CD, 23 with UC) and one empty control, was used for sequencing. The samples were applied to a 96 well plate in duplicate. Another sequencing population, which included 47 UC patients and one control, was used for mutation detection of candidate genes on chromosome 12.

2.4 Handling of samples

The local ethics committees at participating institutions recruited patients. IBD patients were chosen according to standard diagnostic criteria [3 98]. Patient contact was initiated through the general practitioner or a medical center and informed written consent was obtained from all participants. Two tubes of 9 ml EDTA blood were collected from the patients. Additionally, a questionnaire inquiring about disease circumstances, other diseases, and environmental factors was received from every patient and their family members. All material was sent back by regular post [37 50]. The blood was immediately frozen at –80°C and stored at this temperature until the preparation of the DNA. Samples that were collected by collaborators were either sent as frozen EDTA blood or as extracted DNA. Informed written consent was obtained as well from the non-IBD sample populations.

2.4.1 DNA isolation

DNA isolation from EDTA-full blood by guanidine-detergent lysis with DNAzol® was performed either from fresh or from frozen samples, stored at -80°C [100]. About 9 ml blood was used. Frozen samples were thawed at room temperature immediately before preparation by gently inverting the blood vial occasionally. All of the following procedures were conducted on ice. Eighteen ml of MRC buffer (Molecular Research Center; 320 mM saccharose, 5 mM $MgCl_2$, 1 % Triton-X, 1 mM Tris at pH 7.5) was transferred with the blood sample to sterile 50 ml tubes and admixed and incubated for 10 minutes. During this step the red blood cells were destroyed and protein particles were separated from intact leukocytes. The samples were centrifuged for 10 min at 10000 *xg* at 4°C. The supernatant was discarded and the pellet rinsed with 5 ml MRC buffer, which was discarded directly afterwards. The pellet was resuspended in 9 ml MRC buffer and then MRG buffer was added to a total volume of 18 ml. Incubation and centrifugation were repeated as described. The pellet was resuspended in 5 ml of DNAzol® and incubated at room temperature for 15 min until the solution was transparent. In this step the leukocytes were destroyed and the DNA was set free from the nuclei. Proteinase K (stable for 4 weeks at -20°C, 100 µg/ml final concentration) was added and the solution was incubated at room temperature for up to 12 h for enzymatic digestion of proteins. 2.5 ml ice-cold absolute ethanol (0.5 vol. per 1 vol. DNAzol®) was added and gently mixed by inverting the tube. A small thread of DNA precipitate floating in the liquid was transferred into a labelled DNAse free 1.5 ml tube. The precipitate was then washed with 96% and then briefly with 70% ethanol. Supernatants from each washing step were

discarded. The precipitate containing the DNA was then dried at room temperature for several hours (overnight) and then reconstituted in 500 – 1000 µl Tris EDTA buffer (TE), depending on the approximate yield of DNA. To maximize solubilization of the DNA pellet, the solution was incubated 24 hours at room temperature and stored at 4°C until further use.

Invisorb Blood Giga Kit (Invitek, Berlin, Germany) was also used for preparing DNA. Frozen blood samples were thawed in cold water, while gently inverting the blood vial occasionally. Thawed blood samples were transferred to labeled sterile 50 ml tubes previously filled with 30 ml of "Buffer 1" (4°C), shaken and incubated for 10 min at room temperature. During this step the red blood cells were destroyed while the leukocytes stayed intact. The samples were centrifuged for 3 min at 6000 *xg* and the supernatant discarded. Another 20 ml of "Buffer 1" was added to the vial, mixed until the pellet was dissolved, and then centrifuged as above. This step was repeated until the pellet was white, then the pellet was re-suspended in 3 ml "Buffer 2" and 50 µl proteinkinase K followed by 2 h incubation at 60°C while rocking at 95 rpm. If a clear, transparent DNA solution was not obtained after this step, the incubation was extended for another 30 minutes. The solubilzed DNA was then transferred to labelled 15 ml tubes, and 1.8 ml of "Buffer 3" were added. The solution mixed by agitation and incubated for 5 min on ice, followed by centrifugation for 15 min at 10000 x*g*. The supernatant was transferred to a new, labeled 15 ml tube and the two volumes (9.6 ml) of 96% ethanol added. Inverting the tube resulted in precipitation of the DNA. After centrifugation for 3 min at 10000 *xg* the supernatant was discarded, the pellet transferred to a labeled 2 ml tube previously filled with 1 ml 70% ethanol, stirred and centrifuged for 2 min at 10000xg. The

supernatant was discarded, the pellet dried in the tube with open lid until the ethanol was completely evaporated. Finally, 500 µl TE (1x) buffer was added to dissolve the DNA.

For small volumes of blood, DNA was extracted with the PEQLAB DNA isolation system (PEQLAB Biotechnology GmbH, Erlangen, Germany). All procedures were conducted according to the protocol provided with the kit.

DNA samples were stored either at 4°C or at -20°C. The quality and the concentration of the DNA were measured when all DNA was dissolved. The genomic DNA was quantified using PicoGreen (Molecular Probes Europe BV, Leiden, The Netherlands). The automated concentration measurement was carried out on a TECAN SPECTRO FLUOR fluorescence microplate reader (Tecan Deutschland GmbH, Crailsheim, Germany).

2.4.2 Plate design

Individual DNA samples were arranged in 96 well microtiter plates. Each 96 well plate contained a maximum of 93 DNA samples. Two wells contained the same CEPH (Centre d'Etude du Polymorphisme Humain, Paris, France) cell line (as a control in the diallelic discrimination assays), located for each plate layout in the same position. Negative controls (wells containing no DNA) were increased from one to four wells over the course of plate production and were present at the same position for each plate layout. Four 96 well plates were merged into one 384 well plate. Whenever family samples were involved, the plates were designed to keep the families on one plate. The plate-layout with individual identification through barcodes was entered to the database system before the plate was produced. Each plate received an identification number. Application to 96 and

384 deep well plates (ABgene, Epsom, UK) and adjustment of the concentration was performed with the aid of a TECAN Genesis RSP 150 multipipetting robot (Tecan Deutschland GmbH, Crailsheim, Germany). An aliquot of DNA -TE solution containing 2 ng (for diallelic discrimination assays) or 5 ng (for sequencing) DNA was then distributed via 96- and 384-channel Robbins Scientific Hydra microdispensers (Dunn Labortechnik GmbH; Asbach, Germany) to the 96 or 384 well microplates (Costar Corning Incorporated; Cambridge MA, USA, Greiner Bio-One GmbH, Frickenhausen; Germany). The DNA plates were dried at 60°C, and the plates then were heat sealed with a ABGENE ALPS 300 (ABgene, Epsom, UK) for storage at –20°C.

2.4.3 Whole Genome Amplification (WGA) plate preparation

The purpose of making the WGA plate is to perform unlimited genetic tests from limited source material. Genomiphi DNA Amplification Kit (Amersham Biosciences; Freiburg, Germany) was used for preparing WGA plates. The Genomiphi method utilizes bacteriophage Phi29 DNA polymerase to exponentially amplify single- or double-stranded linear DNA templates during an isothermal strand displacement reaction. Microgram quantities of DNA are generated from nanogram amounts of starting material after an overnight incubation. One microlitre of 10-400ng of DNA samples was mixed with the sample buffer, which contained random hexamers and heated at 95°C for 3 minutes. The mixture was quickly cooled on ice. Enzyme mix (1µl) was diluted in 9µl of reaction buffer and added to cooled sample. After incubating the sample at 30°C for 16 hours, the sample was heated at 65°C for 10 minutes and cooled down to 4°C. This sample was now ready for genotyping.

2.5 Diallelic genotyping

2.5.1 Taqman assays

For the analysis of allele status at a known SNP position in a large group of DNA samples, the method of diallelic genotyping (5' nuclease assays) was applied. The 5' nuclease PC assay detects the accumulation of specific PCR product by hybridisation and cleavage of a double-labeled fluorogenic oligonucleotide during the amplification reaction. Besides the forward and reverse primer, two oligonucleotides (called probes), one for each allele, are involved in the reaction. The probe has a reporter fluorescent dye attached at the 5' end and a quencher dye at the 3' end. The two probes have reporter dyes with fluorescent emission at different wavelengths. During the PCR reaction, several events occur. First, each TaqMan MGB probe anneals specifically to a complementary sequence between the forward and reverse primer sites. When the probe is intact, the proximity of the reporter dye to the quencher dye results in suppression of the reporter fluorescence. Second, AmpliTaq Gold DNA polymerase cleaves only probes that are hybridized to the target. Third, cleavage separates the reporter dye from quencher dye, which results in increased fluorescence by the reporter. The increase in fluorescence signal occurs only if the amplified target sequence is complementary to the probe [101]. Mismatches between a probe and target reduce the efficiency of probe hybridization. Furthermore, AmpliTaq Gold DNA polymerase is more likely to displace a mismatched probe without cleaving it, which does not produce a fluorescent signal. After separation from the quencher, the dye becomes fluorescent (Fig. 2.5.1). Visualization takes place through laser scanning technology (ABI Prism® 7700 Sequence Detection System or ABI

Prism 7900HT Sequence Detection System). In each cycle, fluorescence increases, proportionally to the rate of probe cleavage. To induce fluorescence, laser light is distributed to the 96 or 384 sample wells via a multiplexed array of optical fibers. The resulting fluorescent emission returns via the fibers and is directed to a spectrograph with a charge-coupled device (CCD) camera. The fluorescent dyes employed for the analysis are TET (tetrachloro-6-carboxyfluorescein) and FAM™ (6-carboxyfluorescein) and VIC® (trade name by Applied Biosystems) and the quencher TAMRA (6-carboxytetramethylrhodamine). The systems employed (Applied Biosystems TaqMan® technology und TaqMan® -MGB technology) allow 96 well and 384 well platforms. The integrated software depicts the fluorescent status for each well in a diagram and allows assigning an allele calling to each well. A general setting of the software is that the TET (VIC® for TaqMan® -MGB probes) allele is always called 1, FAM™ -allele 2. Each well of the 96 well or 384 well plate results in a dot on the diagram according to the fluorescent intensity of the reporter dyes. The distribution is a cloud of the dots for each homozygous allele 1/1 or 2/2 calling group, and the heterozygous 1/2 calling group.

In contrast to the TaqMan® probes, the TaqMan® -MGB probes contain a minor groove binder (MGB), which attaches to the minor groove of duplex DNA and by this stabilizes hybridisation products and allows shorter probes. Furthermore, a non-fluorescent quencher replaces the TAMRA™ quencher, reducing the background fluorescence and improving the spectral discrimination. Different levels of automation were employed in the diallelic genotyping. Many allelic genotyping assays were ordered in a ready to use

status (Assay-on-Demand, Applied Biosystems Inc., Foster City, CA, USA), or were designed from a sequence by a company (Assay-by-Design Applied Biosystems Inc., Foster City, CA, USA) (both TaqMan®-MGB technology). Both need no further optimisation. Several of the assays employed were self-designed using the TaqMan ® technology. The oligonucleotides were manufactured and labelled with fluorescent dyes by a company (Eurogentec S.A., Seraing, Belgium) according to the defined sequence.

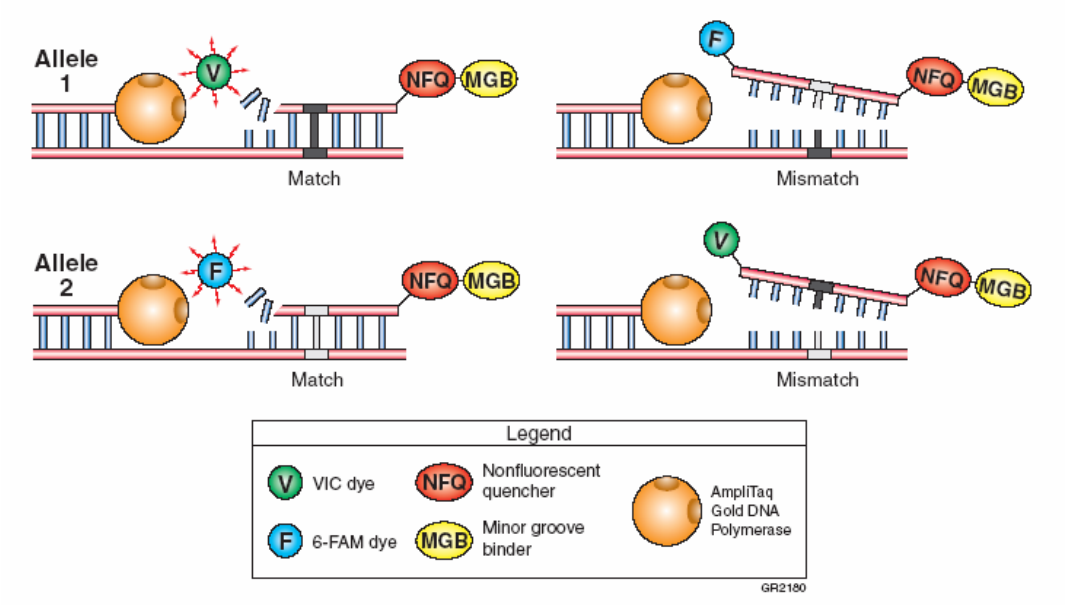


Fig. 2.5.1 The principle of Taqman diallelic genotyping

For the design of oligonucleotides employed in the diallelic assay, the program Primer Express 2.0 (Applied Biosystems Inc., Foster City, CA, USA) was used. The guidelines proposed in the program description were applied [102 103]. The polymorphism was in the middle or the last third of the probe (minimum 5 bases before the end of the probe). The probe was not allowed to start with a guanine (G) as the first base. The required melting

temperature (Tm) for the probes was about 70°C, with a flexible range of ±2°C. The difference in Tm between the probes was not larger than 0.5°C. The sequence of the probes belonging to one assay was allowed to vary in length, with a maximum length of 40 bases. The minimum length was 17 bases. The primer Tm was about 10°C below the Tm of the probes; therefore, the optimal Tm for primers was between 58 to 60°C. Primers were not selected from repeat masked regions or regions with other SNPs. The average amplicon length varied between 50 to 150 bp. Where the SNP needed verification, the amplicon was chosen to be close to 150 bp or even longer to allow a clean sequencing product around the SNP. The reverse primer was allowed to overlap with the probe, but not to reach the SNP. Primers with high variation in the bases at the 3'end were allowed. A thymidine was not allowed in the last 3' position and three same bases in a row within the last third of the oligonucleotide were not allowed. TaqMan® probes were labelled with the fluorescent dyes FAM™ or TET and with the quencher TAMRA™.

2.5.2 SNPlex assays

The SNPlex genotyping system uses Applied Biosystems oligonucleotide ligation assay to achieve allelic discrimination and target amplification. Each assay includes three SNP-specific ligation probes. Two of the probes are allele-specific oligos (ASOs). These are designed specifically for the polymorphism by having the discriminating nucleotide on the 3' end. Each ASO probe contains one of the 96 unique ZipCode sequences for ZipChute probe binding. The third probe is a locus-specific oligo (LSO). Its sequence is common to both alleles of a given locus and anneals adjacent to the SNP site on its target DNA. A set of Universal ASO/LSO linkers is needed. Each ASO is ligated to a universal

ASO-specific linker, which contains a PCR primer sequence corresponding to the universal forward primer and a partial ZipCode sequence. The ASO linkers anneal to the universal ZipCode sequence of the ASO probes. Another linker is ligated to the LSO and has a universal sequence that is compatible with all LSOs, which includes a partial binding site for a universal reverse primer. (Fig 2.5.2a)

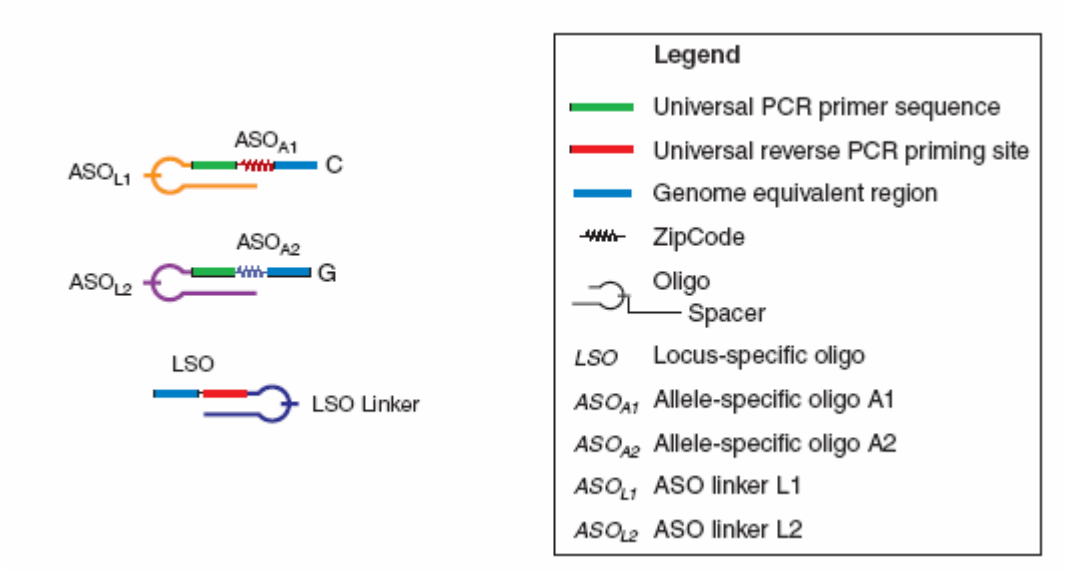


Fig 2.5.2a interaction between SNP-specific probes and universal linkers

ZipChute probes are one of the key components of the SNPlex assays. Each Zipchute probe has a ZipCode-binding sequence, mobility modifiers, which enable size separation during electrophoresis, and a fluorescuent label. (Fig 2.5.2b)

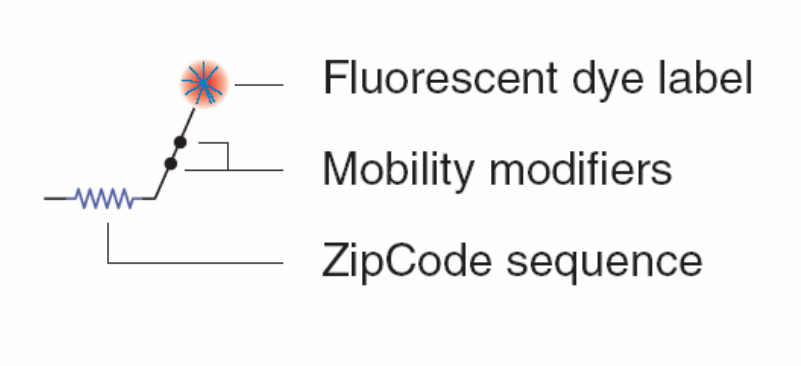


Fig 2.5.2b the parts of a ZipChute probe

The processes required to perform the SNPlex assay are shown in Fig 2.5.2c.

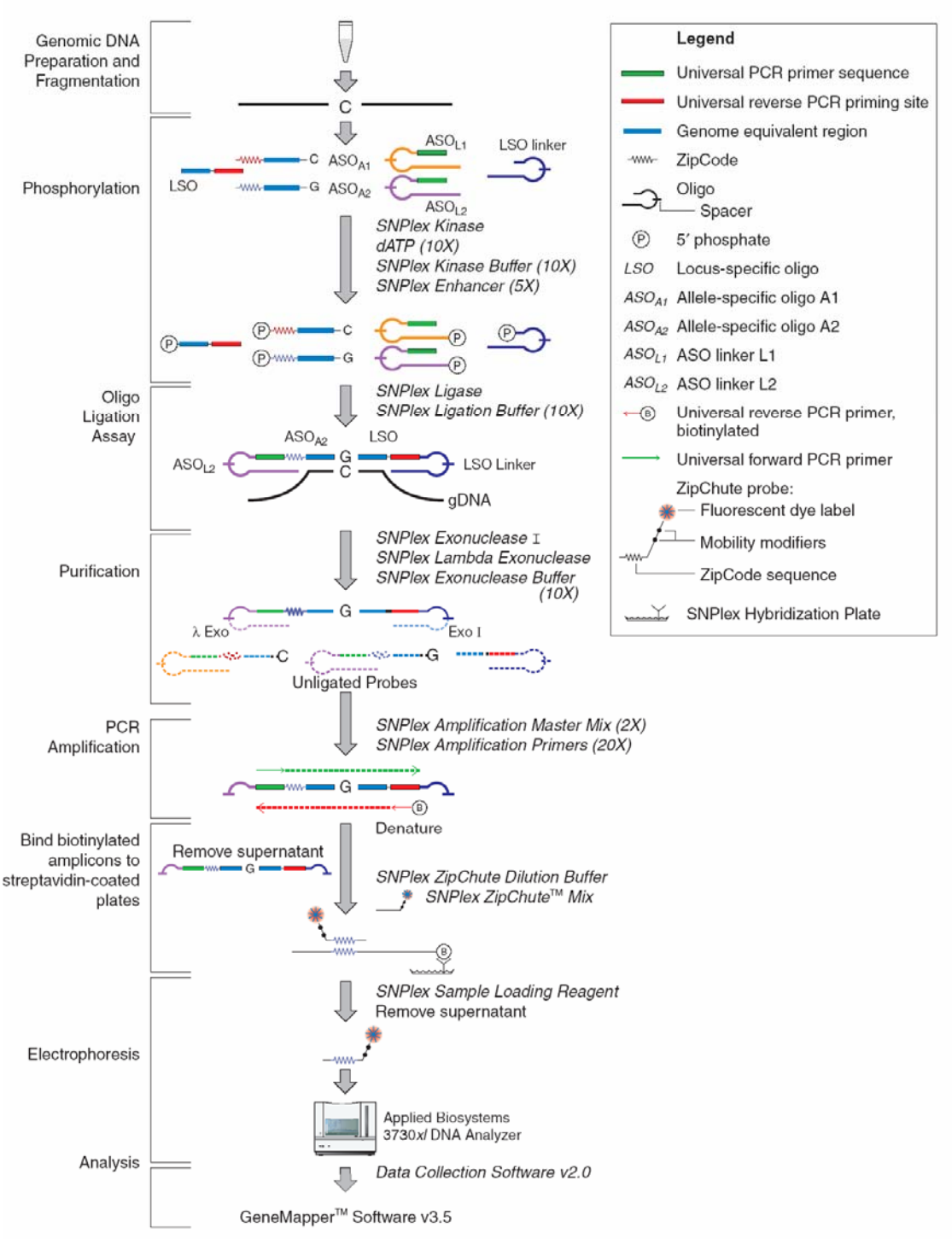


Fig 2.5.2c SNPlex Assay Flowchart

After the reaction, plates are loaded onto 3730xl instrument. All the data were processed using GeneMapper Software 4.0.

2.6 Mutation detection in candidate genes

2.6.1 PCR optimisation

The strategy for the optimization of PCR conditions for different primer-pairs was developed on the basis of standard PCR protocols [104] [105]. First we diluted the oligonucleotide to a storage concentration of 100 pmol/µl with DDW, and then made a working solution of 20 pmol/µl. The reaction solution contained the following reagents: 2.5 µl Qiagen Buffer (10 x concentrated) 0.5 µl dNTP (10 mM), 0.2 µl primer forward; 20 pmol/µl, 0.2 µl primer reverse; 20 pmol/µl 1.0 µl $MgCl_2$ (25 mM) 0.15 µl Taq polymerase (QIAGEN GmbH; Hilden, Germany) and DDW, to a final amount of 25.0 µl per reaction. Five ng of liquid DNA were used. Also the amount of $MgCl_2$ could be changed according to the conditions needed. The PCR reaction was set up in a T1 Gradient PCR machine (Whatman Biometra GmbH, Göttingen, Germany) with the following cycle program: 96°C - 10 min, (96°C - 1 min; 64°C - 1 min [– 0.5 °C per step]; 72°C - 1 min) 16x, (96°C - 1 min; 56°C - 1 min; 72°C - 1 min) 15x, 72°C - 10 min. The gradient covered the temperatures up to 5°C higher and lower than the annealing temperature given in the protocol. A range of temperatures could be tested at one time in this way (e.g. a range from 59°C to 69°C in the example above). Five µl of PCR products together with 2 µl of 2-times concentrate loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF, 30% glycerol in water) were applied to a 1.5% agarose gel. A 100 bp ladder (Invitrogen GmbH, Karlsruhe, Germany) was applied next to the PCR products; electrophoresis

conditions were 150 V for 50 min. A picture was taken under UV light and observed bands of PCR products were compared to expected product length. Quality criteria were one clear band at the correct length without smear, no double bands, and a low amount of primer-dimer.

2.6.2 Sequence analysis

The solutions for the sequencing reaction were prepared according to previously optimized conditions. The solution was applied to a 96 well with 5ng liquid DNA in each well. Five micro liters of PCR product were mixed with 2 µl 2-times loading buffer and tested on a 1.5% agarose gel (in 300 ml TBE, 1% ethidiumbromide, 150 V for 50 min). After finishing the PCR reaction, digestion was required to be remove the unwanted primer-dimers and free dNTPs in the PCR product before sequencing. For a highly concentrated PCR product, a 1:5 dilution with DDW was necessary before the digestion. The enzymatic digestion was performed in a new plate with 8 µl PCR product at the following conditions: 0.30 µl SAP (Shrimp Alkaline Phosphatase; 1 U/µl) 0.15 µl ExoI (Exonuclease I; 10 U/µl) 1.55 µl DDW adding up to 2.0 µl complete reaction mix, with following incubation conditions: 37°C - 15 min, 72°C - 15 min. The chemicals used for sequencing were from the BigDye Terminator Ready Reaction kit (Applied Biosystems Inc., Foster City, CA, USA) based on fluorescent terminator dNTPs. Two microliters of digested PCR product were used for the sequencing reaction, 1.0 µl primer (forward or reverse); 3.2 pmol/µl, 1.0 µl BigDye version1.0 Ready Reaction Mix from the kit, and 6.0 µl DDW for a reaction volume of 8.0 µl. The same reaction was performed with the reverse primer. The following cycle protocol for the sequencing reaction was used: 95°C - 5 min; (95°C - 1 min; 50°C - 15 sec; 60°C - 4min) 25x; 60°C -.5 min. Sample cleanup

was performed with Sephadex spin columns. The Sephadex spin column plates (MAHVN 4550) were prepared by adding Sephadex powder (G50 Superfine) with the aid of a Multiscreen Column loader (gel exclution products obtained from Amersham Biosciences, Freiburg, Germany), 300 µl DDW were added to each column. The plate was incubated at room temperature for 2 hours. Centrifuge the plate at 2100 r/m for 5 minutes and wash the plate with 150 µl DDW and centrifuge once again at 2100 r/m for 5 minutes. Then add 10 µl of DDW to the sequencing produces, after short centrifuge, pipetted it to the centre of the spin columns. The plate was fitted to a MicroAmp ® Optical 96-Well Reaction Plate (Applied Biosystems Inc., Foster City, CA, USA). After centrifugation at 2100 r/m for 5 min, the flow through contained the purified sequencing product. The sequence detection was performed with an automated 96-capillary fluorescence detection system ABI PRISM ® 3700 DNA Analyzer (Applied Biosystems Inc., Foster City, CA, USA), for fluorescent labelled DNA fragments. For the sequence analysis, chromatograms were aligned and compared to the consensus sequences using Sequencher Version 4.0.5 (Gene Codes Corporation, Ann Arbor, MI, USA) and InSNP program [106].

2.6.3 Mutation detection in candidate gene

2.6.3.1 AGR2 and AGR3 gene on chromosome 7

A total of 21 pairs of primers were design to cover all exons and the promoter region of the *AGR2* and *AGR3* genes. Amplification was performed with an ABI GeneAmp ® PCR System 9700 (Applied Biosystems Inc., Foster City, CA, USA) using the following

thermoprofile: 96°C - 10 min, (96°C - 1 min; 59-68.7°C - 1 min [–0.5 °C per step]; 72°C - 1 min) 16x, (96°C - 1 min; 51.2-60.7°C - 1 min; 72°C - 1 min) 25x, 72°C -10 min (Table 2.6.3.1a, Table 2.6.3.1b)

Table 2.6.3.1a Primers and PCR protocol of *AGR2* gene

Exon	Primer	Mixture	Cycle	amplicon
Celera promoter	F:AGTTTAGGTGGAAATTGCTAATGGC R:CTCATGGCTTAATGACTTTGGGTT	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-59.7° 1min-72° 1min (16)/96° 1min-51.9° 1min-72° 1min (20)/72° 10min	746bp
Celera exon1	F: TCGTCTGGCTCCACTTACTCAGA R: ACAGGCAATAACAAATGCTGGC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-59.7° 1min-72° 1min (16)/96° 1min-51.9° 1min-72° 1min (20)/72° 10min	759bp
Celera exon2	F: GACTAAGTGATCCTTTCATTCGGC R: CTAGTGAATGGGCTTGTGCTTGT	Buffer 2.5/Q buffer 5/dNTP 0.5/Primer F(0.3) R(0.3)/Taq Polymerase 0.15/DNA 5/Water 11.25	96° 10min/96° 1min-59° 1min-72° 1min (16)/96° 1min-51.2° 1min-72° 1min (20)/72° 10min	646bp
Ncib promoter	F: GAACTCTGTGGAGGGAAGTTGCT R: CCACCCACTAGTGCTGCATTTAC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-64.5° 1min-72° 1min (16)/96° 1min-56.6° 1min-72° 1min (20)/72° 10min	607bp
Ncbi exon1	F: AATCTGCCCACGGAGCAGA R: GAGGCCCGAATTCACTCCAG	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	480bp
Ncbi exon2-4	F: CAACCTGCAGACCCTGAAGACT R: GAACCCTGGCCCTAAGGCTA	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-64.5° 1min-72° 1min (16)/96° 1min-56.6° 1min-72° 1min (20)/72° 10min	780bp
Ncbi exon5	F: TTGTTCACTGCACCATCCCTAGT R: GGAAGCAATCCCAGTTAATGCA	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	482bp
Ncbi exon6	F: GCCGGAAATGGACAGATTCTT R: ACGTGGCAAGGGACAGTGG	Buffer 2.5/Q buffer 5/dNTP 0.5/Primer F(0.75) R(0.75)/Taq Polymerase 0.15/DNA 5/Water 10.35	96° 10min/96° 1min-60.8° 1min-72° 1min (16)/96° 1min-53° 1min-72° 1min (20)/72° 10min	239bp
Ncbi exon7	F: ACAGCTGTTGATGTTTCCAGCC R: GAGGCGTCCCTAAGCCTAGC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	405bp
Ncbi exon8-2	F: ATGCCAGCCTGAGTGGGAGT R: CTGAAGGAGAAAGCTACTTGCCAG	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-64.5° 1min-72° 1min (16)/96° 1min-56.6° 1min-72° 1min (20)/72° 10min	835bp
Ncbi exon8-3	F: GAGTCAACTCTGGCCAGGAACTC R: ACCTTATTTCACGTCCGCCC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-64.5° 1min-72° 1min (16)/96° 1min-56.6° 1min-72° 1min (20)/72° 10min	924bp

Table 2.6.3.1b Primers and PCR protocol of *AGR3* gene

Exon	Primer	Mix	Cycle	amplicon
Celera promoter	F: AACCGTTCCTGTTTCCTGGG R:TAATGTTTGGCTGAAGTTCAGCAC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	427bp
Celera exon1	F:TCTTGTTTCTATGCAGGTGAGGTTG R: TGGTTACAGGCCTACAGCAGCT	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	581bp
Ncbi promoter	F:GGAACTGAAGGACAAGAATCCTCC R: CACGAAGCCTGCTTCTGAACC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.1) R(0.1)/Taq Polymerase 0.15/DNA 5/Water 15.65	96° 10min/96° 1min-62° 1min-72° 1min (16)/96° 1min-54.2° 1min-72° 1min (20)/72° 10min	260bp
Ncbi exon1	F: TGACTTCTCACCAACTGAGCATGT R: TGCTAGTTGATTTGATTCACTGCC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	541bp
Ncbi exon2	F: AGCCAACAGCTTGATGGCTTAG R: GAGGGTTCAGAGCTGGAAGGAT	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	522bp
Ncbi exon3	F: CTCCTGGGTTCAAGCGATTCT R: TTGCTGAGCGCCTTCTTAGC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.1) R(0.1)/Taq Polymerase 0.15/DNA 5/Water 15.65	96° 10min/96° 1min-62° 1min-72° 1min (16)/96° 1min-54.2° 1min-72° 1min (20)/72° 10min	489bp
Ncbi exon4	F: GACAAAGTGGCAATAGGCCAAT R: GTGCCTGTAGTCCCAGCTACTTG	Buffer 2.5/Q buffer 5/dNTP 0.5/Primer F(0.75) R(0.75)/Taq Polymerase 0.15/DNA 5/Water 10.35	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	438bp
Ncbi exon5	F: CACCTTGTTGGTTAGGCTGGTC R: GATAATGGCCTCTGGCTACATCC	Buffer 2.5/Q buffer 5/dNTP 0.5/Primer F(0.75) R(0.75)/Taq Polymerase 0.15/DNA 5/Water 10.35	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	771bp
Ncbi exon6	F: GGATGTAGCCAGAGGCCATTATC R: GGTGGGAGCTAGAAGTTGGCA	Buffer 2.5/Q buffer 5/dNTP 0.5/MgCl 1.0/Primer F(0.3) R(0.3)/Taq Polymerase 0.15/DNA 5/Water 11.35	96° 10min/96° 1min-60.8° 1min-72° 1min (16)/96° 1min-53° 1min-72° 1min (20)/72° 10min	605bp
Ncbi exon7	F: TCATAGACGAATTCCCATGTTCAA R: CTCCACTATAACTCTTGGCAGGCT	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.1) R(0.1)/Taq Polymerase 0.15/DNA 5/Water 15.65	96° 10min/96° 1min-63.2° 1min-72° 1min (16)/96° 1min-55.4° 1min-72° 1min (20)/72° 10min	416bp

2.6.3.2 Candidate genes on chromosome 12

A total of 29 pairs of primers were design to cover all exons and the promoter region of the *LOC115749*, *BC042855* and *FLJ32549* genes. Amplification was performed with an ABI GeneAmp ® PCR System 9700 (Applied Biosystems Inc., Foster City, CA, USA) using the following thermoprofile: 96°C - 10 min, (96°C - 1 min; 59-68.7°C - 1 min [–0.5 °C per step]; 72°C - 1 min) 16x, (96°C - 1 min; 51.2-60.7°C - 1 min; 72°C - 1 min) 25x, 72°C -10 min (Table 2.6.3.2a, 2.6.3.2b, 2.6.3.2c)

Table 2.6.3.2a Primers and PCR protocol of *LOC115749* gene

Exon	Primer	Mixture	Cycle	amplicon
BC03 Exon1_1	F: GATCCGCAACAGAGTACGCAGT R:CTTGAGGATGTGGTTCTCAGAGTTG	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	542bp
BC03 Exon1_2	F: AGCCGACCTAGCGTCGATTC R: AGGGAGCTCGGTTCTCAAACC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	694bp
BC03 Exon2	F: GCTGTTAGCTTTGGGGGATTTATT R: ACTGTCCGCATCTTTCAATGTGT	Buffer 2.5/Q buffer 5/dNTP 0.5/Primer F(0.3) R(0.3)/Taq Polymerase 0.15/DNA 5/Water 11.25	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	417bp
BC03 Exon3	F: GTGTGAGCCACCGTGCTTG R:CACAGAACCAGAGCTCAACAGATG	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	574bp
BC03 Exon3b_1	F: CCATGAAGGACAGGCACAATACTT R: TCATTTCCATTCCCATTGCACT	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-65.7° 1min-72° 1min (16)/96° 1min-57.8° 1min-72° 1min (20)/72° 10min	550bp
BC03 Exon3b_2	F: GAGCTCTCAGTCTGCACCAACAA R: CAAAGTGCTGGGATTACAGGAATG	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	539bp
BC03 Exon3c	F: AAGGTAGGGGGAATTGATCTTCAG R:CCCATGCAACTCTATCATACATCATC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	174bp
BC03 Exon4	F: CCAGGGACAGTGAGAGTATTCCTG R: ACCCAATGCATGGCAGCTACT	Buffer 2.5/Q buffer 5/dNTP 0.5/Primer F(0.75) R(0.75)/Taq Polymerase 0.15/DNA 5/Water 10.35	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	657bp
BC03 Exon4b	F:TGTTATGTGCTAAGCACTGTTGAAGC R:CAGGTGTGGGATATAATCTCATGGTT	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	574bp
BC03 Exon5	F: GGCCTAACTGGCATCTCCCA R: CTCCGGTAGCTGAAGCATGAGA	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-65.7° 1min-72° 1min (16)/96° 1min-57.8° 1min-72° 1min (20)/72° 10min	628bp
BC03 Exon6	F: GTGCTGGGATTACAGCATGAGTG R: ACTATGGGAAATTCTGGCGGAG	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	665bp
BC03 Exon7	F: TTACCCAAATCTCCCTTAAATCCAG R: ACCCATTAGACCCCTGCTATGG	Buffer 2.5/Q buffer 5/dNTP 0.5/Primer F(0.75) R(0.75)/Taq Polymerase 0.15/DNA 5/Water 10.35	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	773bp
BC03 Exon89_1	F: ACCAGCCTGGGCAATATAGTGAG R: GTGGATCCCAATCCGAGATTTC	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	732bp
BC03 Exon89_2	F:GTACTTGAGATGGTCTCCAAAACTGG R: GGATGGTGTCTGTGGGACAGAG	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	642bp
BC03 Exon10	F:GTAGTGAGGGCCTCAGTCTAGATTTG R:TCTTCTAAATATTGGACCATGTTGCA	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	373bp
BC03 Exon11_1	F: GAGCCTCTGCTCATACTGTGGTG R: TCTTTTGCCCAATGGCATAAAG	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	610bp
BC03 Exon11_2	F: TGCAGAAAAGGCTCCCATTG R: CCCGGCATCCTATTGCTTAAG	Buffer 2.5/dNTP 0.5/MgCl 1.0/Primer F(0.2) R(0.2)/Taq Polymerase 0.15/DNA 5/Water 15.45	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	514bp

Table 2.6.3.2b Primers and PCR protocol of *BC042855* gene

Exon	Primer	Mix	Cycle	amplicon
BC04 Exon1	F: ACTTCCGGGTTCTGAGGGAAC R: CCAGGGCTGACGTCACGTT	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	499bp
BC04 Exon3	F: GTTGTATCCCGAACTTTTCCTCAAC R: CCCAGCAAAAGAATAAATGCATG	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68.7° 1min-72° 1min (16)/96° 1min-60.7° 1min-72° 1min (20)/72° 10min	586bp
BC04 Exon4	F:AAAGGGATTCTTCACATGAATTGATG R: TGAATTGATGTGAGCACACCCA	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	262bp
BC04 Exon5	F: CTAGTTTTTTCTCAGGCAGCCAAG R:TTTATCGAATTTACCTCCCTCATGAG	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	742bp

Table 2.6.3.2c Primers and PCR protocol of *FLJ32549* gene

Exon	Primer	Mix	Cycle	amplicon
B03 Exon1	F: GCCCTTATCCGCGTCTTCTCTA R: GTCGAGCCTGCGACTAGAAGTG	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	583bp
B03 Exon2	F: TGCATGTTGGTAGGTGTTTGGTAC R: CTCTGAGCTTGTGTGGGATTGATAG	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	425bp
B03 Exon3	F:TTAGAAAAATGTCAGTGTGGCAAGTG R:GAAAGAGGCATGAATGGATGTAAACT	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	960bp
B03 Exon3a	F: GAAAATATGATGCTGCCAATGTGTC R: TGGTAGCCCATGACCACTTCC	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	636bp
B03 Exon3b	F: CATCCATTCATGCCTCTTTCAGA R: ACCCATGGCTATTAAAGCTGCC	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	558bp
B03 Exon3c	F: GTCAGATTGGTAGCATGACGTAAGG R: ACCCAGTAAAACAGTGGCTCCAC	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	736bp
B03 Exon4	F: ACTCCCCACAAAAGCCCTGTAG R:GCAGAATTCACTCTCTAGGGACAAAG	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	580bp
B03 Exon4a	F: TTAGACGTGCGGCGAGAATG R:GTCATTTGGTCATTAAACTCTTAGAGCC	Buffer 2.5/dNTP 0.5/MgCl 2.0/Primer(10µM) F(0.5) R(0.5)/Amplitaq Gold Polymerase 0.12/DNA 5/Water 13.88	96° 10min/96° 1min-68° 1min-72° 1min (16)/96° 1min-60° 1min-72° 1min (20)/72° 10min	1012bp

2.7 Internal database

The large numbers of individuals with DNA samples, phenotype description, family information and genotype information was handled through an internal SQL database system [107]. Every sample had a unique identification number. The following information was collected for each patient: age, gender, pedigree information, and phenotypic trait. The plates with DNA samples were identified through barcode system. The quantity and concentration information of the samples were also included in the database. Information about diallelic genotyping assays such as oligonucleotide sequences (not available for Assay-on-Demand assays), chromosomal position, and genetic variants were also entered in the database. After genotyping was performed, the results were entered into the database and automatically linked to the individual sample through the barcode of the plate. The database allowed exporting data of individuals, within an analysis population, with their pedigree information; phenotypic trait and genotype information from a set of self selected SNP markers. The export format was the appropriate format for a selection of analysis programs (LINKAGE PREFILE). Furthermore, sets of application tools were used to get the information about the assays and the plates. Additional applications helped to enter and edit sample information, create plate templates, to control robots to make plates and to measure DNA concentration.

Integrated to the database system was a test for Mendelian inheritance errors. This application was part of the SNP genotyping data import to the database. It was utilized on each SNP marker and all populations for which the marker had been genotyped. All inheritance errors were shown in a list with all family members and their genotyping data for that assay. The positions of the family members in the fluorescence diagram of the plate containing the family were seen through the linked plate-view [107]. A few families with Mendel errors were

normal and acceptable. For SNP markers with a high number of inheritance errors, the SNP marker was excluded from the analysis when the assay was evaluated as not reliable. For sample populations of single individuals, genotyping results were subjected to assessment of HWE by a $\chi 2$ -test at the 1% significance level. This test was applied as well to genotypes from families, however a slight imbalance could be expected in this situation.

2.8 Statistical analysis

Before the genetic analysis, each marker was tested for Hardy-Weinberg Equilibrium (HWE) [108] [109] in the control population. Family-based analyses were performed using the transmission disequilibrium test (TDT) [84] in trios, using TRANSMIT [51] and GENEHUNTER [81]. Haplotype frequency estimates among singletons were obtained using an implementation of the EM algorithm (HAPMAX) [110]. Significance testing of haplotype frequency differences was also performed with HAPMAX, making use of the fact that twice the log-likelihood ratio between two nested data models approximately follows a χ^2 distribution with k degrees of freedom, where k is the difference in parameter number between the two models. Significance assessment of associations with or between single locus genotypes was performed using χ^2 or Fisher's exact test for 2×3 contingency tables. All other statistical calculations were performed with SPSS.

Allele and genotype frequencies for each SNP were calculated from all unrelated control individual. In multiplex family, only one affected individual was selected randomly. In monoplex families, each affected individual was selected. Case-control analysis was performed for alleles and genotypes using contingency tables. Pearson's χ^2 was used for

measuring the association between the disease trait and the SNPs. The significance at the 95% confidence interval is given as p-value.

Family based transmission distortion was analyzed by using transmission disequilibrium test (TDT). This test is based on a comparison of alleles transmitted from parents to affected offspring against the alleles that were expected to be transmitted from the population frequency in the parent generation. This was calculated by using GENEHUNTER program, which calculated the TDT by comparing alleles transmitted from parent to offspring with alleles untransmitted.

Linkage disequilibrium (LD) was calculated by employing the HAPMAX program. The measure estimates the difference between the number of two marker haplotypes observed and the one expected under the assumption of independence in segregation of markers using the correlation coefficient.

For the candidate genes with significant association signals, odds ratio (OR) was calculated as a measure of increased risk of affection with a certain SNP variant.

2.9 cDNA amplification

2.9.1 Amplification of AGR2 cDNA

The public databases and Celera Discovery Systems were reviewed in order to establish gene models for the *AGR2* and *AGR3* genes. For *AGR2*, two additional 5' exons were annotated. The presence of the additional 5-prime exons was investigated through RT-PCR in the

Clontech multiple tissue panels I and II (East Meadow Circle, Palo Alto, U.S.A.), using standard supplier's protocols. For the gene model evaluation of *AGR2,* the following primers: were used for the short form: AGRf5: TCA ACT CTG GCC AGG AAC TC; AGRr5: TAC AGC ACC ATA GTC CAG GG and 2) for the long form (designed on the basis of the NCBI and Celera gene models): AGRf11: CGA CTC ACA CAA GGC AGG T; AGRr11: GCT GTA TCT GCA GGT TCG T.

2.9.2 Amplification of cDNA from candidate genes on chromosome 12

Primers were designed according the public database. RT-PCR were performed in the Clontech multiple tissue panels I and II (East Meadow Circle, Palo Alto, U.S.A.). The following primers were used: *LOC115749* gene: BC_03_exF1: CCG AGT CAT CAC GCC TGA AC; BC_03_exR1: TCG GAT TGG GAT CCA CAA AGT; *BC042855* gene: B04_RNA_F1: GTC GAC TTT CCA AAG CGC TG; B04_RNA_R1: TGG ATT GAG ACT GAG CAT GCC and *FLJ32549* gene: B03_RNA_F1: AAG GTC TAT CAC AGC CTC ACC TAC C; B03_RNA_R1: CTT TCA GGA GAT GAC ACA GGA CG. The reaction solution contained the following reagents: 2.5 µl Buffer (10 X concentrate) 0.5 µl dNTP (10 mM), 0.5 µl primer forward; 10 pmol/µl, 0.5 µl primer reverse; 10 pmol/µl 2.0 µl MgCl2 (25 mM) 0.15 µl Amplitaq Gold polymerase (Applied Biosystems; Weiterstadt, Germany) and DDW, to a final amount of 25.0 µl per reaction. Five µl of liquid cDNA were used. The PCR reaction was set up with the following cycle program: 96°C - 10 min, (96°C - 1 min; 68°C - 1 min ; 72°C - 1 min) 16x, (96°C - 1 min; 60°C - 1 min; 72°C - 1 min) 24x, 72°C - 10 min. Five µl PCR products together with 2 µl of 2 X concentrate loading buffer (0.25% bromophenol blue, 0.25% Xylene Cyanol FF, 30% Glycerol in Water) were applied to a 1.5% agarose gel (300

ml Tris borate EDTA (TBE), 3 µl ethidium-bromide). A 100 bp ladder (Invitrogen GmbH, Karlsruhe, Germany) was applied next to the PCR products; electrophoresis conditions were 150 V for 50 min. A picture was taken under UV light.

2.10 Rapid Amplification of cDNA Ends (RACE)

Rapid amplification of cDNA ends (RACE) is a polymerase chain reaction (PCR)-based technique which was developed to facilitate the cloning of full-length cDNA 5'- and 3'-ends after a partial cDNA sequence has been obtained by other methods. Marathon cDNA amplification is a method for performing both 5' and 3' rapid amplification of cDNA ends from the same template (Fig 2.10a). A Marathon-Ready cDNA kit (kidney) was used to confirm if other transcription products existed in the *LOC115749* gene. Marathon-Ready cDNAs are premade "libraries" of adaptor-ligated ds cDNA ready for use as templates in Marathon cDNA amplification. The first and the second strand cDNA synthesis were done by the company (Fig 2.10b). Primers were designed by using standard supplier's protocols. The following primers were used for 5' RACE in *LOC115749* gene: BC_GSP_F1: GCA GGG AAA TCT CGG ATT GGG ATC CAC. The reaction solution contained the following reagents: 5µl 10*cDNA PCR reaction buffer, 1µl dNTP (10mM), 1µl Advantage 2 polymerase mix (50*), 5µl Marathon-ready cDNA (Kidney, 1:5 diluted), 1µl AP1 Primer (10mM), 1µl EB_GSP_F1 primer (10mM) and 36µl DDW. The PCR reaction was set up with the following cycle program: 94°C 30sec, (94°C 5sec, 72°C 4min) 5x, (94°C 5sec, 70°C 4min) 5x, (94°C 5sec, 68°C 4min) 30x.

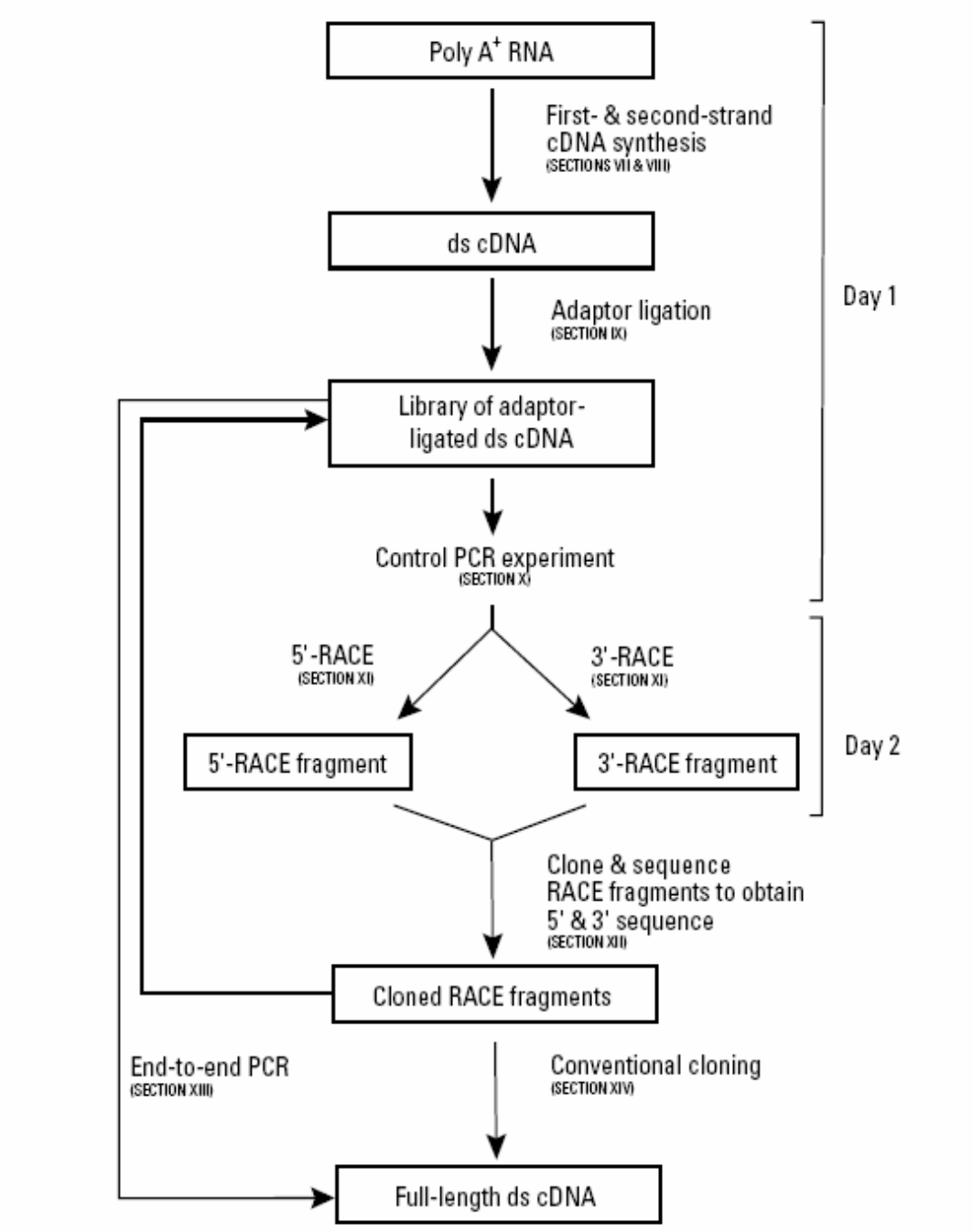


Fig 2.10a Overview of Marathon procedure

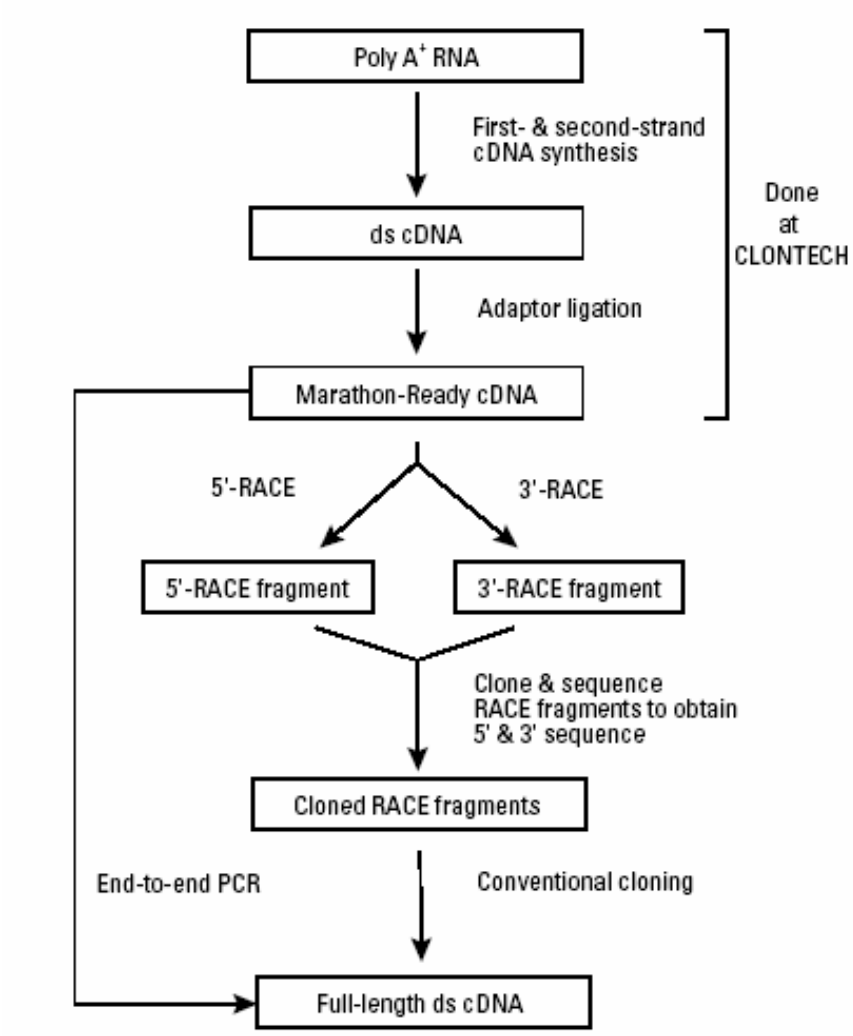


Fig 2.10b Overview of Marathon cDNA ready kit procedure

2.11 Cell culture, reporter gene constructs and dual luciferase reporter gene assay

HEK 293 cells were purchased from the German Collection of Microorganisms and Cell Cultures (DSMZ, Braunschweig, Germany). The cells were cultured in DMEM + 10% fetal calf serum. One day prior to transfection, cells were seeded at a density of 5 x 10^5

cells/2 ml on 6-well plates. Transfections were performed with Fugene 6 (Roche, Switzerland) according to the manufacturer's manual by using 0.08 µg of the target plasmid and 0.02 µg of the pRL-TK reference plasmid (Promega, Mannheim, Germany) for the reporter gene assays. The constructs for the goblet-cell transcription factors *FOXA1* and *FOXA2* have been described elsewhere [111]. Twenty-four hours after transfection, the cells were harvested for reporter gene assay. Transfection efficiency was determined by parallel detection of pRL-TK activity in a dual luciferase reporter gene assay (Promega, Madison, WI, USA). Every single transfection experiment was performed in duplicate and was repeated at least 3 times. The *AGR2* promoter of the short form (NM_006408) from -1542 to -1 was amplified from 100 ng of human genomic DNA by polymerase chain reaction (PCR) under standard conditions with the following primers (restriction sites underlined) pGL_AGR2_N_sense(XhoI): CGC TCG AGA TCT TTA CAG AGG TAA TTA AGT TAA AGT A; pGL_AGR2_N_anti(HindIII): GCA AGC TTG TTG CTA ACT CAG AAA CGA ACC TTC CTT TCC CCA A and cloned into the pGL3-basic plasmid (Promega). All constructs were sequence-verified with an ABI3700 sequencer (ABI, Foster City, CA) before use.

Luciferase activity was determined with a dual luciferase reporter gene kit (DLR) from Promega according to the manufacturer's manual. The cells lysates were analyzed with a MicroLumatPlus LB96V microplate luminometer (EG&G Berthold, Wellesley, MA) after automatic injection of the necessary substrate solutions. All samples were at least measured in duplicate. The results for firefly luciferase activity were normalized to renilla luciferase activity.

2.12 Real-time PCR

RNA transcript levels were measured using quantitative real-time PCR in 138 patient and normal control samples, including 25 normal controls, 56 CD and 57 UC patients. Biopsies were obtained from small and large bowel, with 125 of the 138 samples originating from the sigmoid colon. Patients included in this study consented to the additional research biopsies being taken 24 h prior to endoscopy. The study protocol was approved by the hospital ethical committee prior to the start of the study. Total RNA was isolated from snap-frozen biopsies using a commercial kit (Qiagen, Hilden, Germany). One microgram of total RNA was then reverse-transcribed to cDNA according to the manufacturer's instructions (MultiScribe Reverse Transcriptase, Applied Biosystems, Foster City, CA, USA). The cDNA from each sample was diluted 1:5 and arrayed on 384-well plates for real-time PCR quantitation using an Assays-on-Demand Gene Expression Assay for *AGR2* (Hs00180702_m1; context sequence: GTT TGT TGA CCC ATC TCT GAC AGT T) on the ABI Prism 7900HT Sequence Detection System (Applied Biosystems) according to the manufacturer's instructions. Relative transcript levels were determined using the standard curve quantitation method and β-actin as the endogenous control gene.

3. Results

3.1 *AGR2* and *AGR3* genes

3.1.1 Mutation detection results

The mutation detection experiment identified a total of 30 single nucleotide polymorphisms (SNPs), of which nineteen were not previously known. Twenty-five SNPs were located in the *AGR2* gene and five mapped to the *AGR3* gene. One SNP (hcv111845 – rs4719482) lead to an amino acid exchange in the additional N-terminal sequence of the extended splice variant. An overview of all identified SNPs is given in Table 3.1.1a.

Table 3.1.1a Results of the mutation detection of all exons and the promoters of the *AGR2* and *AGR3* genes.

Gene	Name	Position	SNP	Type	Note
AGR2	07AGR8N1348	1348 in ncbi exon 8	C/T	Intron	Novel*
AGR2	07AGR8N1234	1234 in ncbi exon 8	A/G	Intron	Novel*
AGR2	07AGR8N104	1044 in ncbi exon 8	A/C	3'UTR	hCV26516309
AGR2	07AGR8N707	707 in ncbi exon8	A/G	3'UTR	Novel*
AGR2	07AGR8N392	392 in ncbi exon 8	A/T	3'UTR	Novel*
AGR2	hCV1702536	92 in ncbi exon 7	G/T	Intron	hCV1702536
AGR2	hcv1702535	46 in ncbi exon 7	C/T	Exon	hcv1702535
AGR2	hCV11830196	-60 before ncbi exon 5	A/T	Intron	hCV11830196
AGR2	07AGR1N34	34 in ncbi exon1	A/G	5'UTR	hCV8302356
AGR2	07AGR1N17	17 in ncbi exon1	A/G	5'UTR	rs706073
AGR2	07AGRNP53	-53 before ncbi exon1(in promoter region)	A/C	Intron	hCV27493993
AGR2	07AGRNP199	-199 before ncbi exon1(in promoter region)	C/T	Intron	hcv1170870
AGR2	07AGRNP261	-261 before ncbi exon1(in promoter region)	C/T	Intron	rs17136670
AGR2	07AGRC2_122	-122 before celera exon 2	G/T	Intron	Novel*
AGR2	hcv9501480	428 in celera exon 1	A/G	Intron	hcv9501480
AGR2	hcv111845	144 in celera exon 1	C/T	5'UTR	hcv111845
AGR2	07AGRCP176	-176 before celera exon 1 (in promoter region)	A/G	Intron	Novel*
AGR2	07AGRCP197	-197 before celera exon 1 (in promoter region)	C/T	Intron	Novel*
AGR2	07AGRCP207	-207 before celera exon 1 (in promoter region)	C/T	Intron	Novel*
AGR2	07AGRCP299	-299 before celera exon 1 (in promoter region)	C/T	Intron	Novel*
AGR2	07AGR2C395	-395 before celera exon 1 (in promoter region)	C/T	Intron	rs12674158
AGR2	07AGR2C517	-517 before celera exon 1 (in promoter region)	C/T	Intron	SNP_A-1721592
AGR2	07AGR2CP574	-574 before celera exon 1 (in promoter region)	G/T	Intron	Novel*
AGR2	07AGRCP619	-619 before celera exon 1 (in promoter region)	A/G	Intron	Novel*
AGR2	07AGRCP626	-626 before celera exon 1 (in promoter region)	A/G	Intron	Novel*
AGR3	hCV11170861	58 in ncbi exon 7	A/G	3'UTR	Hcv11170861
AGR3	07AGR3E6_184	-184 before ncbi exon 6	G/C	Intron	Novel*
AGR3	07AGR3E2_89	89 in ncbi exon 2	C/T	Intron	Novel*
AGR3	rs4472406	-54 before ncbi exon 2	C/T	Intron	rs4472406
AGR3	hCV2571858	290 in ncbi exon 1	A/T	Intron	hCV2571858

3.1.2 Results of Data Analyses

The German study cohort and 537 healthy controls of German descent were genotyped for 30 SNPs of the *AGR2* and *AGR3* genes. All markers were in Hardy-Weinberg equilibrium both in the case and control samples. Case-control and TDT tests of association were performed separately for IBD and the sub-phenotypes CD and UC. Results are shown in table 3. Markers passing a nominal threshold of $p<0.05$ in the German screening sample were further investigated in an independent patient and control sample from the UK. Association statistics in the UK, German and combined cohorts for three diagnostic categories (UC, CD, IBD) are presented in Tables 3.1.2a, Table 3.1.2b, Table 3.1.2c and Table 3.1.2d.

The consistency of the TDT and case-control statistics and the strength of replication between populations were used as a pragmatic guide to judge the trustworthiness of the positional signals. The linkage disequilibrium structure (LD) of the human *AGR2* and *AGR3* gene regions is shown in Figure 3.1.2 using the r^2 measure of linkage disequilibrium. Taking the LD data and the association results (Tables 3.1.2a-3.1.2d) together, the association is most consistent in the UC phenotype and localizes to the 5-prime region of the *AGR2* gene (Fig 3.1.2). The association is most pronounced at hcv1702494 (combined sample: $P_{TDT}=0.011$, $P_{case/control}=0.0007$, OR=1.34) and hcv111845 ($P_{TDT}=0.029$, $P_{case/control}=0.005$, OR=1.2837) for the UC phenotype. A haplotype analysis including the markers 07AGRNP53, 07AGRNP261, hcv1702494 and hcv111845 yielded similar results (German population, UC phenotype, $\chi^2=14.8$, df=3, p=0.002) with an odds ratio for the risk haplotype of 1.43. In the UK cohort UC

phenotype, there was no significance (χ^2=8.18, df=4, p=0.085). The association of the CD subphenotype to the marker hcv8302351 is largely limited to the German subpopulation. Further, TDT and case-control test gave disparate results for this marker.

Table 3.1.2a: Overview of single point association statistics in the German cohort. Results that meet nominal p-value criterion of 0.05 are highlighted in bold print.

		IBD				CD				UC			
Gene	SNP	TDT	Case/control			TDT	Case/control			TDT	Case/control		
		P	P	Allele frequencies		P	P	Allele frequencies		P	P	Allele frequencies	
				cases	controls			cases	controls			cases	controls
AGR2	hcv1702558	0.61	**0.024**	0.38	0.33	0.96	**0.046**	0.38	0.33	0.45	0.11	0.38	0.33
AGR2	hcv1702545	0.24	**0.016**	0.37	0.32	0.29	**0.035**	0.37	0.32	0.58	0.08	0.37	0.33
AGR2	07AGR8N392	0.24	0.86	0.02	0.02	0.56	0.25	0.02	0.02	0.16	0.26	0.01	0.02
AGR2	hcv1702537	0.53	0.195	0.22	0.25	0.78	0.26	0.22	0.25	0.48	0.20	0.22	0.25
AGR2	hcv1702535	0.79	0.27	0.49	0.46	0.89	0.25	0.49	0.46	0.49	0.45	0.48	0.46
AGR2	hcv8302351	**0.016**	**0.001**	0.39	0.47	0.19	**0.001**	0.39	0.46	**0.022**	**0.018**	0.40	0.47
AGR2	hcv1702532	0.85	0.66	0.19	0.20	0.42	0.87	0.20	0.20	0.41	0.70	0.19	0.20
AGR2	07AGR1N34	0.2	0.25	0.17	0.19	0.43	0.34	0.17	0.19	0.27	0.24	0.16	0.19
AGR2	07AGR1N17	0.052	0.299	0.17	0.19	0.11	0.63	0.19	0.19	0.27	0.18	0.16	0.19
AGR2	07AGRNP53	0.057	**0.038**	0.21	0.17	0.24	0.11	0.20	0.17	0.11	**0.023**	0.23	0.18
AGR2	07AGRNP199	0.07	0.09	0.05	0.07	**0.03**	0.052	0.05	0.07	0.89	0.38	0.06	0.07
AGR2	07AGRNP261	0.17	0.12	0.21	0.18	0.52	0.27	0.20	0.18	0.15	**0.034**	0.23	0.19
AGR2	hcv1702494	**0.013**	**0.013**	0.43	0.48	0.14	0.06	0.44	0.49	**0.027**	**0.01**	0.41	0.48
AGR2	hcv474914	0.18	0.34	0.31	0.33	0.42	0.67	0.32	0.33	0.23	0.06	0.29	0.33
AGR2	hcv111845	**0.028**	0.11	0.35	0.38	0.23	0.34	0.36	0.38	**0.034**	**0.03**	0.32	0.38
AGR2	07AGRCP176	**0.005**	0.078	0.03	0.05	**0.029**	0.10	0.03	0.05	0.08	0.59	0.04	0.05
AGR2	07AGRCP197	**0.047**	0.42	0.11	0.12	0.24	0.86	0.12	0.12	0.06	0.25	0.09	0.11
AGR2	07AGRCP207	0.21	0.52	0.01	0.001	0.56	0.67	0.001	0.001	0.26	0.14	0.01	0.001
AGR2	07AGRCP299	0.08	0.097	0.22	0.19	0.33	0.21	0.21	0.19	0.1	0.054	0.24	0.19
AGR2	07AGR2C395	0.51	0.82	0.14	0.14	0.46	0.76	0.15	0.15	0.92	0.33	0.13	0.15
AGR2	07AGR2C517	0.56	1	0.14	0.14	0.43	0.54	0.15	0.14	0.92	0.42	0.13	0.14
AGR2	07AGR2CP574	0.76	0.1	0.01	0.001	0.32	0.1	0.01	0.001	0.16	0.52	0.01	0.001
AGR2	07AGRCP619	0.12	0.54	0.14	0.13	0.57	0.18	0.16	0.14	**0.036**	0.66	0.12	0.13
AGR2	07AGRCP626	0.6	0.7	0.14	0.15	0.67	0.80	0.16	0.15	0.77	0.25	0.13	0.15
AGR3	hcv318606	0.84	0.92	0.32	0.32	0.96	0.97	0.32	0.32	0.66	0.74	0.33	0.32
AGR3	hcv11170861	0.68	0.99	0.33	0.33	0.84	0.89	0.32	0.32	0.66	0.84	0.33	0.33
AGR3	hcv2571858	0.94	0.49	0.29	0.28	0.88	0.71	0.29	0.28	0.73	0.41	0.31	0.29
AGR3	hcv2571854	0.29	0.72	0.15	0.15	0.95	0.67	0.16	0.15	0.06	0.5	0.16	0.15
AGR3	hcv2571840	0.42	0.94	0.32	0.32	0.48	0.94	0.33	0.33	0.71	0.96	0.32	0.32
AGR3	hcv2571839	0.93	0.7	0.31	0.30	0.88	0.87	0.31	0.30	0.94	0.69	0.30	0.29

Table 3.1.2b: Replication analysis of significant markers from Table 3 in a UK cohort in the UC subgroup. Significant values at the nominal p-value criterion of 0.05 are given in bold print.

	German				UK				Combined			
SNP	TDT	Case_control			TDT	Case_control			TDT	Case_control		
	P	P	Allele frequence		P	P	Allele frequence		P	P	Allele frequence	
			cases	controls			cases	controls			cases	controls
hcv1702558	0.45	0.11	0.38	0.33	0.41	0.39	0.39	0.36	0.29	0.07	0.38	0.34
hcv1702545	0.58	0.08	0.37	0.33	0.11	0.35	0.38	0.34	0.24	**0.04**	0.38	0.33
hcv8302351	**0.022**	**0.018**	0.40	0.47	0.91	0.86	0.40	0.41	**0.04**	**0.04**	0.40	0.44
07AGRNP53	0.11	**0.023**	0.23	0.18	0.39	0.51	0.13	0.21	0.69	**0.013**	0.23	0.19
07AGRNP199	0.89	0.38	0.06	0.07	0.39	0.27	0.05	0.07	0.72	0.21	0.06	0.07
07AGRNP261	0.15	**0.034**	0.23	0.19	0.33	0.497	0.22	0.21	0.84	**0.023**	0.23	0.19
hcv1702494	**0.027**	**0.01**	0.41	0.48	0.17	**0.045**	0.39	0.45	**0.011**	**0.0007**	0.39	0.47
hcv111845	**0.034**	**0.03**	0.32	0.38	0.41	**0.031**	0.33	0.39	**0.029**	**0.005**	0.33	0.38
07AGRCP176	0.08	0.59	0.04	0.05	0.80	0.88	0.05	0.05	0.14	0.63	0.04	0.05
07AGRCP197	0.06	0.25	0.09	0.11	0.22	0.80	0.12	0.12	**0.03**	0.31	0.11	0.12
07AGRCP619	**0.036**	0.66	0.12	0.13	0.15	0.71	0.14	0.13	**0.01**	0.91	0.13	0.13

Table 3.1.2c: Replication analysis of significant markers from Table 3 in a UK cohort in the CD subgroup. Significant values at the nominal p-value criterion of 0.05 are given in bold print.

	German				UK				Combined			
SNP	TDT	Case_control			TDT	Case_control			TDT			
	P	P	Allele frequence		P	P	Allele frequence		P	P	Allele frequence	
			cases	controls			cases	controls			cases	controls
hcv1702558	0.96	0.46	0.38	0.33	0.87	0.21	0.60	0.65	0.38	**0.024**	0.38	0.34
hcv1702545	0.29	**0.035**	0.37	0.32	0.21	0.20	0.39	0.34	0.13	**0.018**	0.38	0.45
hcv8302351	0.19	**0.001**	0.39	0.46	0.10	0.22	0.38	0.42	0.49	**0.0001**	0.38	0.45
07AGRNP53	0.24	0.11	0.20	0.17	0.49	0.17	0.17	0.21	0.22	0.50	0.19	0.18
07AGRNP199	**0.03**	0.052	0.05	0.07	0.24	0.06	0.05	0.07	**0.014**	**0.013**	0.05	0.07
07AGRNP261	0.52	0.27	0.20	0.18	0.66	0.17	0.17	0.21	0.53	0.8	0.19	0.19
hcv1702494	0.14	0.06	0.44	0.49	**0.04**	0.23	0.41	0.45	**0.028**	**0.038**	0.44	0.48
hcv111845	0.23	0.34	0.36	0.38	0.10	0.29	0.35	0.38	0.14	0.18	0.36	0.38
07AGRCP176	**0.029**	0.10	0.03	0.05	0.72	0.7	0.05	0.05	0.11	0.26	0.04	0.05
07AGRCP197	0.24	0.86	0.12	0.12	0.44	0.5	0.11	0.13	0.12	0.65	0.12	0.12
07AGRCP619	0.57	0.18	0.16	0.14	0.22	0.64	0.13	0.12	0.45	0.15	0.15	0.13

Table 3.1.2d: Replication analysis of significant markers from Table 3 in the UK cohort in for the joint IBD phenotype. Significant values at the nominal p-value criterion of 0.05 are given in bold print.

	German				UK				Combined			
SNP	TDT	Case_control			TDT	Case_control			TDT			
	P	P	Allele frequence		P	P	Allele frequence		P	P	Allele frequence	
			cases	controls			cases	controls			cases	controls
hcv1702558	0.61	**0.024**	0.38	0.33	0.06	0.31	0.61	0.64	0.18	**0.03**	0.38	0.34
hcv1702545	0.24	**0.016**	0.37	0.32	0.14	0.25	0.38	0.34	0.054	**0.016**	0.37	0.33
hcv8302351	**0.016**	**0.001**	0.39	0.47	0.17	0.51	0.39	0.41	**0.005**	**0.005**	0.39	0.41
07AGRNP53	0.057	**0.038**	0.21	0.17	0.28	0.7	0.20	0.21	**0.037**	0.13	0.21	0.18
07AGRNP199	0.07	0.09	0.05	0.07	0.77	0.16	0.05	0.07	0.08	0.06	0.05	0.07
07AGRNP261	0.17	0.12	0.21	0.18	0.34	0.73	0.20	0.21	0.12	0.23	0.21	0.19
hcv1702494	**0.013**	**0.013**	0.43	0.48	**0.015**	0.11	0.41	0.45	**0.001**	**0.006**	0.42	0.47
hcv111845	**0.028**	0.11	0.35	0.38	0.26	**0.013**	0.34	0.39	**0.012**	**0.04**	0.35	0.38
07AGRCP176	**0.005**	0.078	0.03	0.05	0.88	0.97	0.05	0.05	**0.031**	0.20	0.04	0.05
07AGRCP197	**0.047**	0.42	0.11	0.12	0.17	0.61	0.12	0.13	**0.013**	0.35	0.11	0.12
07AGRCP619	0.12	0.54	0.14	0.13	0.55	0.77	0.13	0.13	**0.03**	0.5	0.14	0.13

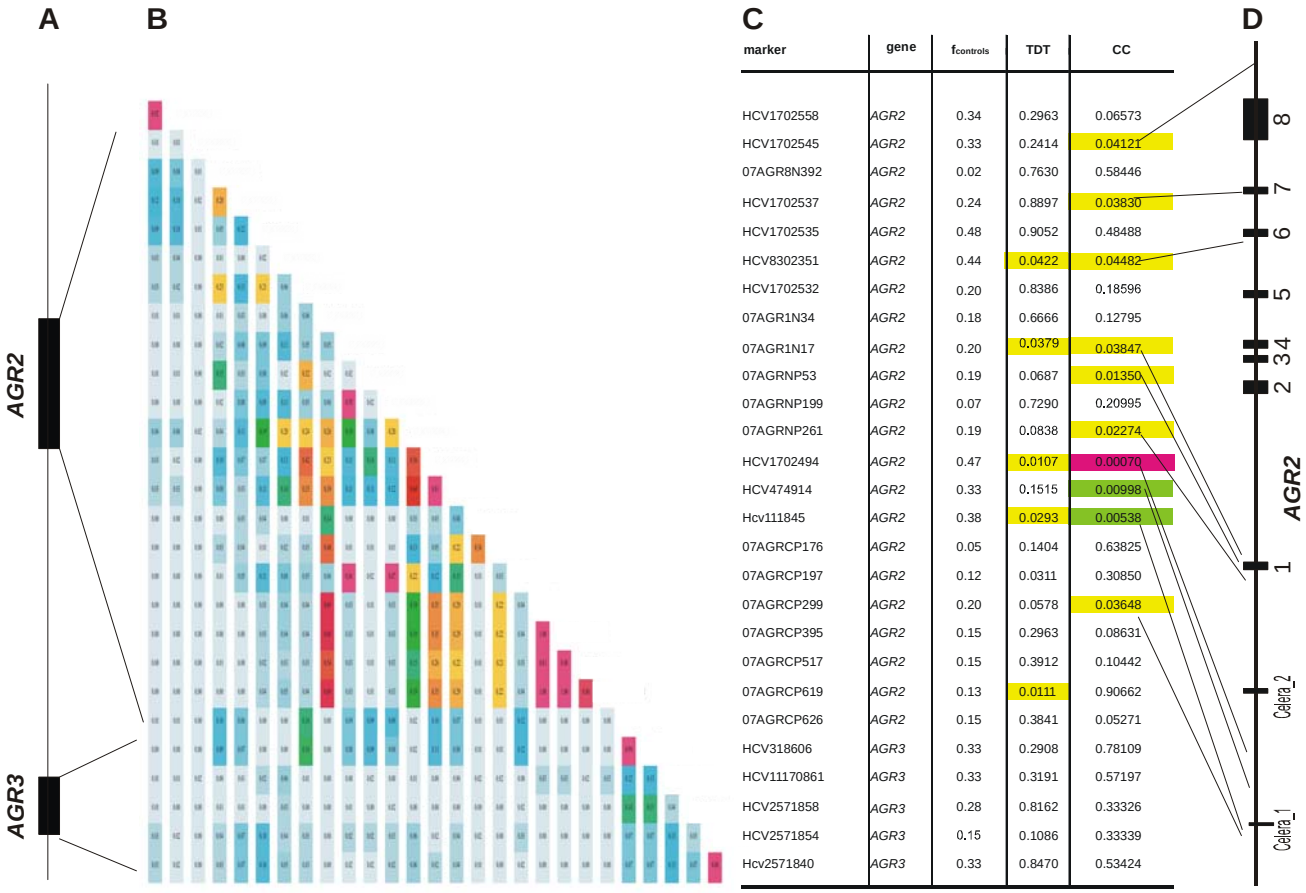


Fig 3.1.2 Overview of the linkage disequilibrium structure of the human AGR2 and *AGR3* region. The panels A and D show the physical map and the exon-intron structure of the *AGR2* gene. The panel B shows the r^2 plot for the region. The level of LD is highlighted in color (red r^2=1, light blue r^2=0). The association results for the UC phenotype in the combined German / UK cohort are given in panel C.

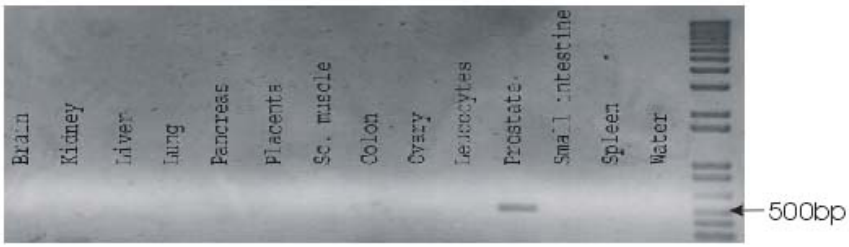
3.1.3 cDNA amplification results

The gene model of *AGR2* was evaluated using RT-PCR in a tissue panel designed on the basis of the NCBI and Celera gene models. The presence of the additional 5' exon predicted in the Celera database was confirmed. The two different transcripts also show a significantly different expression pattern (Fig 3.1.3). The extended form shows a predominant expression in the prostate whereas the shorter form shows a ubiquitous expression pattern. The resultant gene model has been submitted to Genbank.

cDNA of short form of AGR2

cDNA of long form of AGR2

Expression pattern of short form of AGR2



Expression pattern of short form of AGR2

Fig 3.1.3 Evaluation of the expression pattern of the two AGR2 splice variants. RT-PCR was performed in the Human Multiple Tissue Panels using the primers given in the methods.

3.1.4 Real-time PCR results

The relative expression of total *AGR2* was quantified by real-time PCR in 25 normal controls, 56 CD and 57 UC patients. Median expression levels (arbitrary normalization units) of 1.07 for normal controls, 0.57 for CD and 0.67 for UC were observed. This expression difference was statistically significant as tested by non-parametric testing ($p=0.0000001$ for CD versus NC and $p=0.0000001$ for UC versus NC; Mann Whitney U Test, Fig 3.1.4).

For 91 IBD biopsies, genotype data for the *AGR* markers was available. Individuals, who were homozygous for the risk allele at marker hcv111845, showed an overall lower expression relative expression level than the remainder of the samples (relative expression level 0.63 for "22" genotype versus 0.71). However, this difference was not statistically significant due to the limited sample number ($p>0.1$).

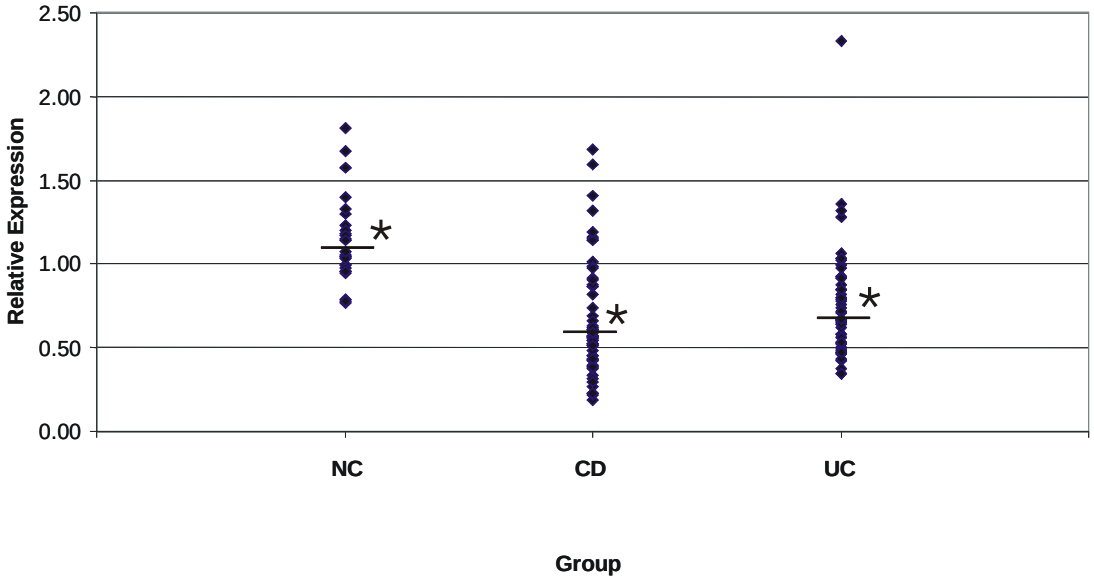


Fig 3.1.4 Relative expression of AGR2 in normal and IBD patient samples. The relative expression of AGR2 was quantitated by real-time PCR in 25 normal controls, 56 CD and 57 UC patients. The expression of AGR2 is significantly different between normal controls and either CD or UC samples (p-value <0.001, as shown by Mann Whitney U Test). * = Median

3.1.5 Expression of Reporter gene construct

A luciferase reporter gene construct (pGL3B-AGR2 [-1542]) driven by the *AGR2* promoter was transfected into HEK 293 cells. Co-transfection of the forkhead box transcription factors *FOXA1* and *FOXA2*, which have been implicated in maintaining goblet cell function [112], led to a significantly increased luciferase activity (Fig 3.1.5). Median luciferase activities (out of seven replicates) of 1023 for the control and 3378 and

2455 for the *FOXA1* and *FOXA2* co-transfection experiments were observed ($p<0.001$, Mann-Whitney U-test).

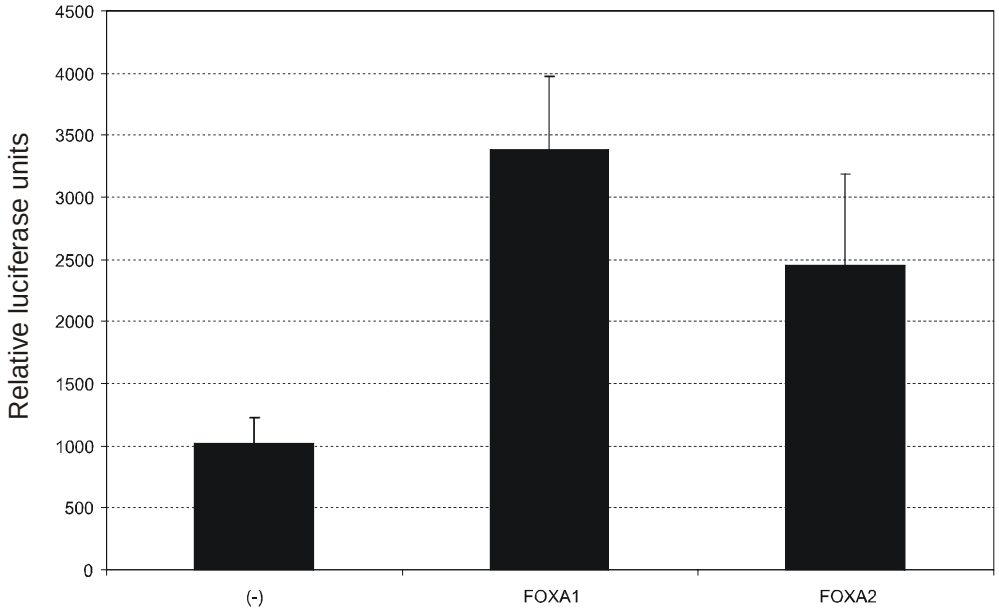


Fig 3.1.5 Luciferase activity observed after stimulation with FoxA1 and FoxA2. (-): transfected with pGL3B-AGR2[-1542] without stimulation; FoxA1: co-transfected with pGL3B-AGR2 [-1542] and FoxA1 gene; FoxA2: transfected with pGL3B-AGR2 [-1542] and FoxA2 gene.

3.2 Association mapping on chromosome 12

3.2.1 Identification of the association lead on chromosome 12

For the association study of inflammatory bowel disease (IBD) susceptibility genes, we selected 15 SNPs covered 62.6-63.2Mb region on chromosome 12q. All of the 15 SNPs

were genotyped in a German cohort, which contained 776 trios with IBD (484 with CD, 292 with UC). All markers were in Hardy-Weinberg equilibrium both in the case and control samples. Case-control and TDT tests of association were performed for IBD and the sub-phenotypes CD and UC. Results are shown in table 3.2.1

Table 3.2.1: Overview of single point association statistics in the German cohort.

	IBD				CD				UC			
SNP	TDT	Case/control			TDT	Case/control			TDT	Case/control		
	P	P	Allele frequencies		P	P	Allele frequencies		P	P	Allele frequencies	
			cases	controls			cases	controls			cases	controls
hcv1521134	0.62	0.82	0.19	0.19	0.32	0.99	0.19	0.19	0.75	0.60	0.20	0.19
hcv2690436	0.23	0.68	0.23	0.23	0.67	0.57	0.22	0.23	0.19	0.93	0.24	0.24
hcv2690456	0.32	0.46	0.28	0.26	0.90	0.44	0.28	0.27	0.18	0.65	0.27	0.26
hcv1403210	0.41	0.69	0.50	0.50	0.86	0.38	0.48	0.50	0.32	0.66	0.48	0.50
hcv8757697	0.97	0.94	0.48	0.50	0.90	0.98	0.49	0.50	0.95	0.86	0.48	0.49
hcv8757644	0.44	0.19	0.25	0.28	0.51	0.99	0.28	0.29	0.06	**0.003**	0.21	0.28
hcv2630409	0.66	0.09	0.37	0.33	0.44	0.25	0.36	0.32	0.84	**0.04**	0.39	0.34
hcv11290387	0.37	0.40	0.22	0.21	0.76	0.86	0.20	0.20	0.12	**0.036**	0.25	0.20
hcv11290388	0.24	0.35	0.22	0.21	0.76	0.94	0.20	0.21	0.055	**0.034**	0.25	0.21
hcv9277491	0.42	0.31	0.22	0.21	0.94	0.98	0.20	0.20	0.24	**0.03**	0.25	0.20
hcv400963	0.44	**0.02**	0.40	0.44	0.51	0.27	0.42	0.46	0.076	**0.0008**	0.35	0.44
hcv12069598	0.81	0.43	0.14	0.12	0.45	0.82	0.13	0.12	0.67	0.19	0.15	0.12
hcv491186	0.55	0.39	0.25	0.25	0.72	0.37	0.25	0.25	0.63	0.60	0.25	0.26
hcv22272931	0.69	0.67	0.24	0.25	0.89	0.88	0.25	0.25	0.67	0.46	0.23	0.25
hcv315870	0.40	0.49	0.45	0.43	0.24	0.35	0.46	0.42	1	0.92	0.44	0.43

Two SNPs (hcv8757644 & hcv400963) show highly significant association ($p<0.005$) with UC sub-group in case-control analyses.

3.2.2 High density SNP mapping in the association region

Further investigation of this region comprised genotyping 35 additional markers in the same samples. These markers were marked as tagging SNPs in HapMap (www.hapmap.org). All markers were in Hardy-Weinberg equilibrium both in the case

and control samples. For the analysis of association in the case control, the population was classified into the diagnostic categories healthy and affected with IBD, which included CD and UC. Cases for each category were taken from the trios and single patients. Pearson's χ^2 were calculated. The results were shown in Fig 3.2.2a. The most significant SNP was rs7955726, which yielded a p-value of 0.0006.

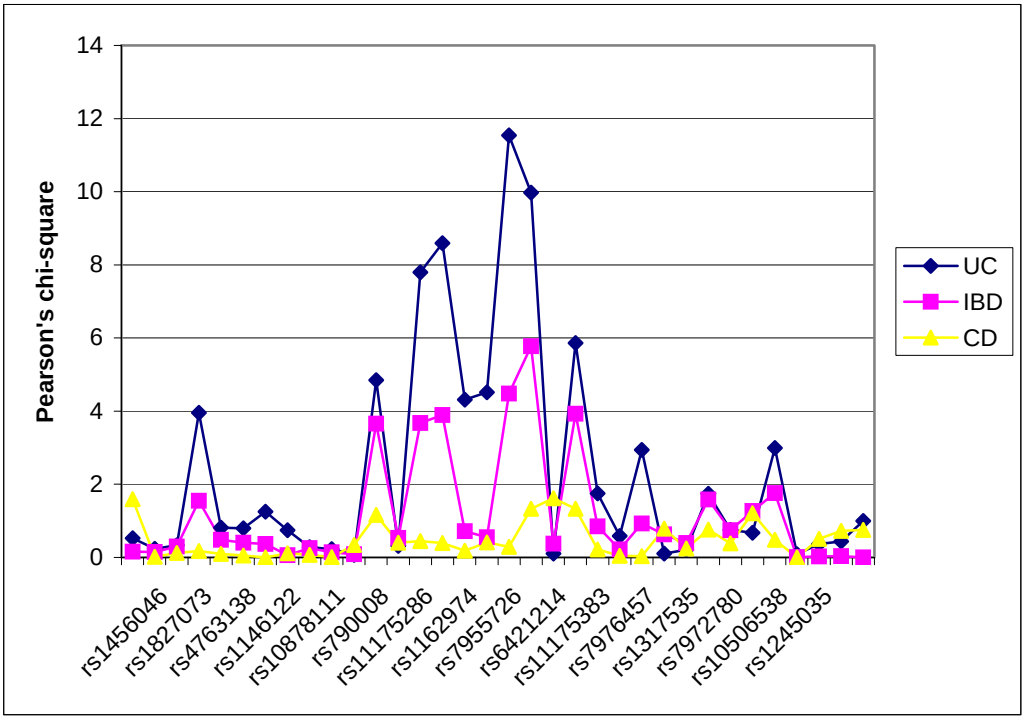


Fig 3.2.2a Pearson's χ^2 in case-control studies. Significance of a peak at the 95% confidence interval is reached at value of 3.9 with 1 df.

Besides the case control study, the association to disease was tested using family-based transmission distortion. In Genehunter program (version 2.1) the number of alleles transmitted from parent to offspring was compared to the number of untransmitted alleles [81]. The significance of the association was calculated using χ^2 test and the p-value at a 95% confidence interval. The results of the single point association with transmission are

shown in Fig 3.2.2b. The highest significant SNP marker was also rs7955726, which gave p-value at 0.002.

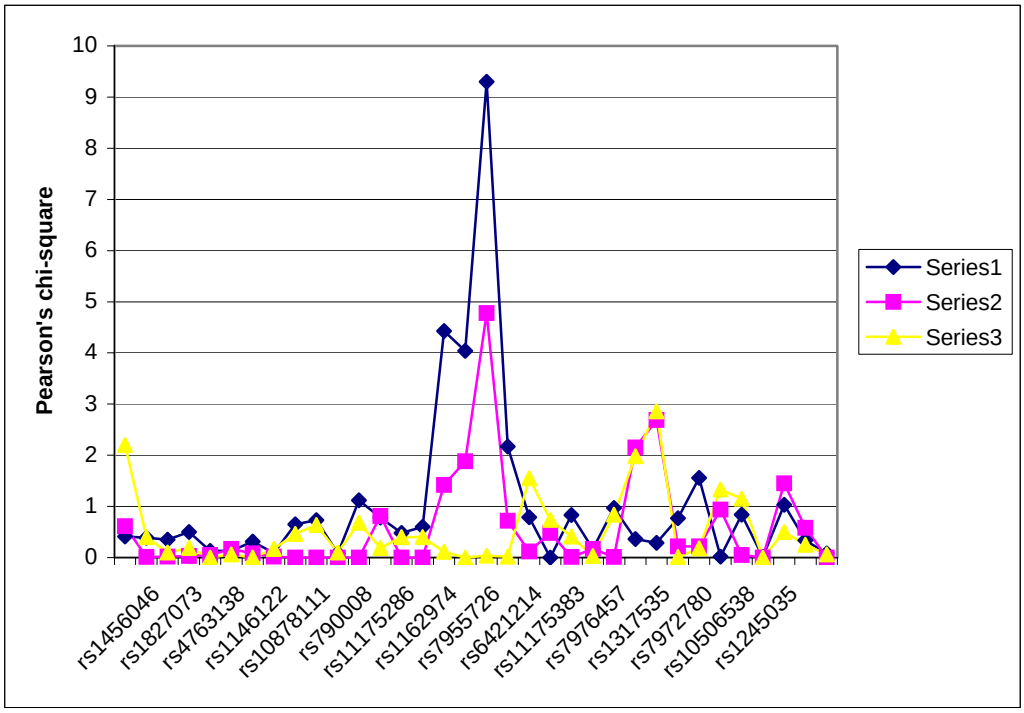


Fig 3.2.2b Pearson's χ^2 in case-control studies. Significance of a peak at the 95% confidence interval is reached at value of 3.9 with 1 df.

Different associations can be identified for each disease category. There is one peak of strong association with UC in case-control studies and TDT that reach a significance level of $p<0.05$. Several markers and resulting peak were mostly carried through UC group. In IBD category, a weak association also can be seen in the same peak. There was no peak of association with CD category. The most significant SNP, rs7955726, showed strong association with UC category both in case-control studies and TDT. The overview is shown in table 3.2.2. The most significant SNP was located in *LOC115749* gene.

Table3.2.2 Overview of single point association statistics in the German cohort

	IBD				CD				UC			
SNP	TDT	Case/control			TDT	Case/control			TDT	Case/control		
	P	P	Allele frequencies		P	P	Allele frequencies		P	P	Allele frequencies	
			cases	controls			cases	controls			cases	controls
rs7311377	0.434	0.709	0.43	0.44	0.138	0.211	0.41	0.45	0.522	0.475	0.46	0.44
rs1456046	0.922	0.706	0.39	0.38	0.527	0.899	0.39	0.39	0.535	0.634	0.40	0.38
rs10878102	0.886	0.596	0.36	0.35	0.755	0.742	0.36	0.35	0.553	0.572	0.36	0.34
rs1827073	0.874	0.218	0.10	0.08	0.666	0.681	0.08	0.08	0.480	**0.047**	0.11	0.08
rs2279666	0.819	0.491	0.47	0.46	1	0.774	0.47	0.46	0.718	0.371	0.48	0.46
rs4763138	0.680	0.530	0.48	0.46	0.811	0.844	0.47	0.46	0.720	0.376	0.49	0.46
rs789722	0.774	0.554	0.47	0.48	0.952	0.984	0.49	0.48	0.576	0.265	0.45	0.49
rs1146122	0.890	0.830	0.48	0.49	0.681	0.746	0.50	0.49	0.766	0.395	0.47	0.49
rs10748002	1	0.623	0.47	0.48	0.498	0.815	0.48	0.48	0.421	0.610	0.47	0.48
rs10878111	0.963	0.711	0.47	.048	0.423	0.922	0.48	0.49	0.394	0.645	0.47	0.48
rs1695105	0.963	0.791	0.48	0.50	0.757	0.570	0.47	0.50	0.775	0.814	0.49	0.49
rs790008	0.954	0.056	0.20	0.25	0.408	0.281	0.22	0.26	0.289	**0.028**	0.19	0.24
rs1464055	0.368	0.468	0.45	0.48	0.665	0.531	0.45	0.48	0.376	0.585	0.45	0.47
rs11175286	0.957	0.055	0.24	0.28	0.530	0.511	0.26	0.29	0.490	**0.005**	0.21	0.27
hcv2690409	0.787	0.075	0.38	0.33	0.671	0.230	0.37	0.32	0.946	**0.039**	0.39	0.34
rs1657050	1	**0.048**	0.24	0.28	0.526	0.539	0.26	0.29	0.439	**0.003**	0.20	0.27
rs1162974	0.234	0.402	0.22	0.22	0.746	0.677	0.19	0.21	**0.035**	**0.038**	0.25	0.21
rs7954838	0.170	0.463	0.22	0.22	1	0.526	0.19	0.21	**0.044**	**0.034**	0.26	0.21
rs7955726	**0.029**	**0.034**	0.32	0.29	0.838	0.603	0.28	0.28	**0.002**	**0.0006**	0.36	0.28
rs2164504	0.395	**0.016**	0.40	0.46	0.902	0.253	0.43	0.47	0.140	**0.0016**	0.37	0.45
rs6421214	0.731	0.550	0.45	0.43	0.213	0.207	0.47	0.43	0.374	0.768	0.42	0.43
rs10878167	0.488	**0.048**	0.21	0.25	0.391	0.252	0.22	0.25	1	**0.015**	0.19	0.25
rs11175383	0.914	0.361	0.26	0.24	0.521	0.652	0.25	0.24	0.361	0.191	0.28	0.24
rs10784408	0.683	0.649	0.14	0.13	0.857	0.855	0.14	0.13	0.677	0.450	0.15	0.13
rs7976457	0.938	0.337	0.11	0.13	0.361	0.876	0.12	0.13	0.325	0.087	0.09	0.12
rs1317532	0.143	0.432	0.48	0.46	0.159	0.379	0.48	0.47	0.551	0.758	0.47	0.46
rs1317535	0.101	0.535	0.32	0.34	0.091	0.632	0.32	0.34	0.590	0.653	0.32	0.34
rs6581575	0.639	0.213	0.48	0.45	0.903	0.389	0.47	0.46	0.379	0.192	0.49	0.45
rs7972780	0.638	0.395	0.50	0.47	0.668	0.542	0.49	0.48	0.211	0.390	0.50	0.47
rs4964110	0.334	0.262	0.09	0.10	0.249	0.274	0.09	0.10	0.900	0.420	0.09	0.10
rs10506538	0.824	0.189	0.11	0.09	0.283	0.491	0.10	0.09	0.359	0.084	0.12	0.09
rs7963840	1	0.917	0.37	0.38	0.950	0.965	0.38	0.38	0.940	0.787	0.37	0.37
rs1245035	0.229	0.890	0.40	0.40	0.479	0.482	0.41	0.40	0.310	0.554	0.38	0.40
rs1520765	0.448	0.871	0.47	0.46	0.616	0.402	0.49	0.47	0.562	0.513	0.45	0.46
rs1619280	1	0.998	0.46	0.47	0.808	0.393	0.44	0.46	0.777	0.320	0.49	0.46

3.2.3 Analyses of LD in the association region

Linkage disequilibrium (LD) is the condition in which the haplotype frequencies in a population deviate from the values they would have if the genes at each locus were combined at random. When there is no such deviation, when linkage disequilibrium equal to zero, then the population is said to be in linkage equilibrium. Linkage Disequilibrium (LD) arises as a consequence of three features of life a) the physical structure of chromosomes; b) the inherent mutations that occur at random during DNA replication; c) the rate of recombination between any two given loci. Taking each in turn this means that markers, which are after all simply mutations, be it SNPs where one base pair has changed or micro-satellites (where replication slippage has occured), do not undergo independent assortment if they are on the same chromosome. This means that when a new mutation arises it will be inherited along with all of the other markers/polymorphisms that occur on that chromosome, unless of course a recombination event occurs between two loci that serve to break the pattern of mutations that are inherited on one chromosome. There are two major parameters, one is D' and another is r^2. D' is scaled version of D (Equivalent to the co-variance between loci) constrainted to lie between 0 and 1. The higher the value is, the stronger LD is. The r^2 is the gametic correlation coefficient.

Due to linkage disequilibrium between genes, a distortion of the true location of a susceptibility gene is likely. Therefore, pair-wise LD analysis in the control population was performed on the set of markers employed for the screening experiment. Markers with a higher frequency than 1% were included into analysis of LD structure in this study. D' values were calculated for all pair-wise inferred two-locus haplotypes for the 49 SNPs. D' values greater than 0.5 were considered as evidence of linkage disequilibrium between markers [113] (Fig 3.2.3). In the analyzed fragment, four blocks were discovered. The

strongest LD was observed between markers in the third region, which was located between rs1162974 and rs7955726. This region includes *LOC115749* gene. This region also demonstrated significant results in the case-control studies and TDT analysis.

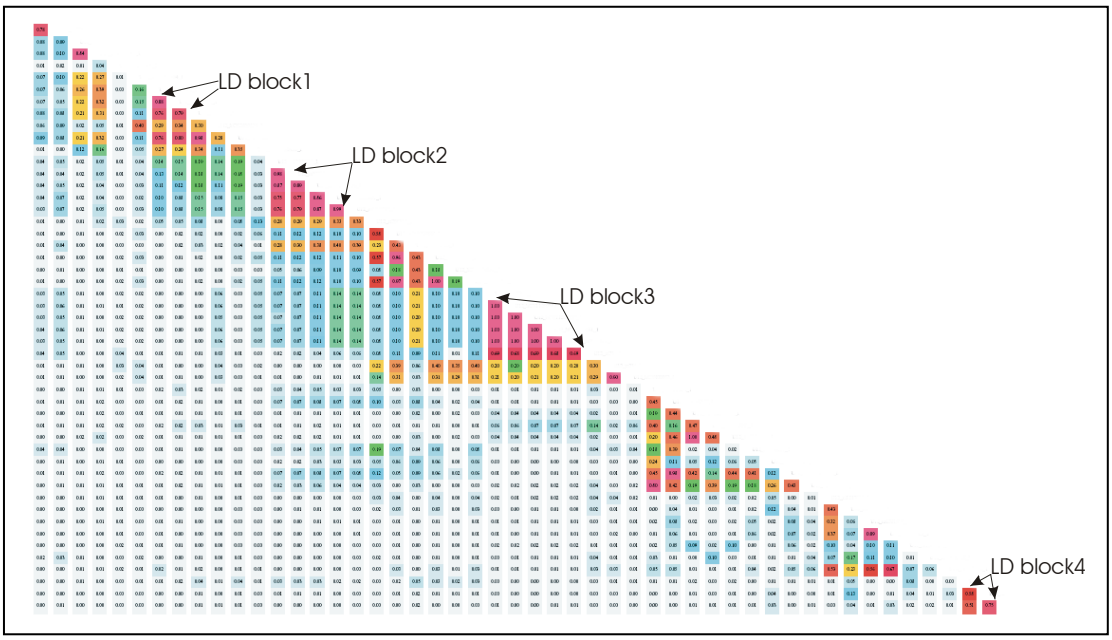


Fig 3.2.3 LD between 49 SNPs markers typed for association on chromosome 12. The colored squares code D' value. LD values marked in with red indicate LD values higher than 0.5.

3.2.4 Candidate gene analyses in association region

3.2.4.1 Mutation detection results

We selected three candidate genes in this region, namely *LOC115749*, *FLJ32549* and *BC042855*. In total, 20 SNPs were identified in the mutation detection experiment. One SNP in *LOC115749* gene (BC03_SNP2) located in exon, which was a synonymous

mutation, and one SNP in *FLJ32549* gene (12BC036N2232) lead to an amino acid exchange from asparagine to serine. Two deletions (12BC039N8134, 12BC036N4343) and one insertion (12BC04N326) were found in those three genes (Table 3.2.4.1). All the other SNPs were located in the intron of these three genes.

Table 3.2.4.1 Results of the mutation detection of the *LOC115749*, *BC042855* and *FLJ32549* genes.

Gene	Name	Position	SNP	Type	Note
LOC115749	BC03_SNP1	-347 before exon 1	C/T	Intron	
LOC115749	BC039_SNP3	-74 before exon 4	A/C	Intron	
LOC115749	rs12582530	-23 before exon 5	C/G	Intron	
LOC115749	rs12581950	118 in exon 6	C/T	Intron	
LOC115749	rs10878132	-118 before exon 8	G/T	Intron	
LOC115749	BC03_SNP2	63 in exon 8	G/T	Exon	Synonymous
LOC115749	12BC039N8134	134 in exon 8	A/-	Intron	One A deletion
LOC115749	rs2010889	-193 before exon 9	C/T	Intron	
LOC115749	rs2010893	-136 before exon 9	C/T	Intron	
LOC115749	rs6581534	97 in exon 9	A/C	Intron	
BC042855	12BC04N1192	192 in exon 1	C/T	Intron	
BC042855	12BC04N338	-38 before exon 3	A/G	Intron	
BC042855	12BC04N326	-26 before exon 3	-/T	Intron	One T insertion
BC042855	rs7489144	365 in exon 5	C/T	Exon	
FLJ32549	12BC036N2232	232 in exon 2	A/G	Exon	Asn to Ser
FLJ32549	12BC036N31123	1123 in exon 3	T/-	3'UTR	One T deletion
FLJ32549	BC036_SNP2	343 in exon 4	A/G	3' UTR	
FLJ32549	rs2643665	449 in exon 4	A/G	3'UTR	
FLJ32549	12BC036N4553	553 in exon 4	A/G	3'UTR	
FLJ32549	rs1133242	1235 in exon 4	C/T	3'UTR	

3.2.4.2 Statistical analyses results

Ten SNPs from the mutation detection in these three genes were genotyped. Two of them were significant in case-control studies in the UC subgroup. One of the SNPs, BC03_SNP2, was located in exon 8 and didn't cause amino acid change. The other significant was intronic. Three markers showed significant results in TDT. Two of them (rs6581534 and BC039_SNP3), which were intronic SNPs, were located within the LOC115749 gene. The third one (rs7489144) was located within *BC042855* gene. The overview is shown in table 3.2.4.2.

Table 3.2.4.2 Overview of single point association statistics in the German cohort. Results that meet nominal p-value criterion of 0.05 are highlighted in bold print

		IBD				CD				UC			
Gene	SNP	TDT	Case/control			TDT	Case/control			TDT	Case/control		
		P	P	Allele frequencies		P	P	Allele frequencies		P	P	Allele frequencies	
				cases	controls			cases	controls			cases	controls
FLJ32549	rs1133242	0.693	0.121	0.21	0.24	0.605	0.412	0.22	0.25	0.223	0.066	0.19	0.23
FLJ32549	rs2643665	1	0.977	0.49	0.50	0.616	0.858	0.49	0.50	0.560	0.831	0.50	0.50
FLJ32549	BC036_SNP2	0.366	0.621	0.01	0.01	0.317	0.887	0.01	0.01	1	0.321	0.003	0.01
LOC115749	rs6581534	0.184	0.540	0.22	0.22	1	0.607	0.19	0.21	**0.048**	0.071	0.25	0.21
LOC115749	rs2010893	0.312	0.445	0.22	0.22	0.690	0.658	0.19	0.21	0.050	**0.047**	0.25	0.21
LOC115749	rs2010889	0.342	0.517	0.22	0.22	0.629	0.587	0.19	0.21	0.054	0.059	0.25	0.21
LOC115749	BC03_SNP2	0.347	0.400	0.22	0.21	0.936	0.726	0.20	0.21	0.172	**0.025**	0.26	0.21
LOC115749	BC039_SNP3	0.176	0.837	0.22	0.23	1	0.449	0.20	0.22	**0.043**	0.175	0.25	0.22
LOC115749	BC03_SNP1	0.694	0.497	0.43	0.42	0.546	0.343	0.44	0.41	0.950	0.932	0.42	0.42
BC042855	rs7489144	0.921	0.156	0.06	0.04	0.043	**0.025**	0.08	0.04	**0.039**	0.916	0.05	0.04

3.2.4.3 cDNA amplification results

The gene model of *LOC115749 gene, BC042855 gene* and *FLJ32549 gene* were evaluated using RT-PCR in a tissue panel designed on the basis of the NCBI gene models. The *LOC115749* gene was highly expressed in testis and also expressed in kidney, prostate, placenta and thymus. Two splicing variants were seen in kidney, prostate, placenta and thymus (Fig 3.2.4.3a). The *FLJ32549* gene was expressed in all tissues.

BC042855 gene was expressed only in muscle, heart and brain. *BC042855* gene has two different variants expressed in brain (Fig 3.2.4.3b).

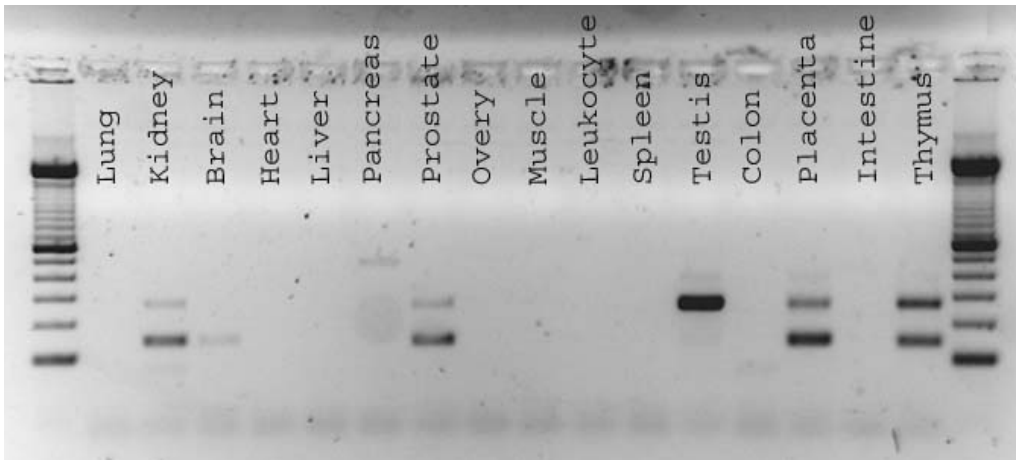


Fig 3.4.2.3a Evaluation of the expression pattern of the LOC115749 gene. RT-PCR was performed in the Human Multiple Tissue Panels using the primers given in the methods.

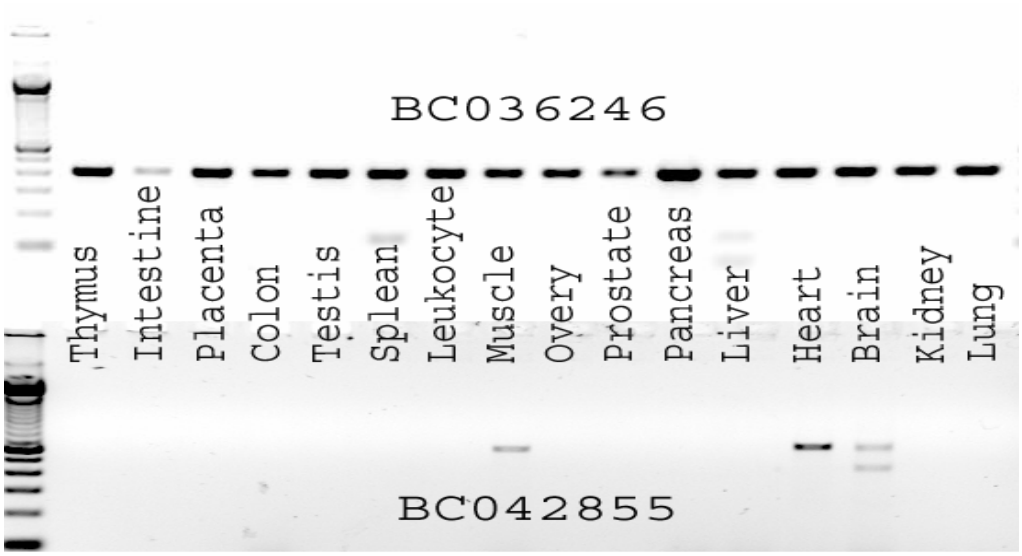


Fig 3.2.4.3b Evaluation of the expression pattern of the BC042855 and BC036246 gene. RT-PCR was performed in the Human Multiple Tissue Panels using the primers given in the methods.

3.2.4.4 Rapid amplification of cDNA ends (RACE) results

Five prime rapid amplification of cDNA ends (RACE) was performed in kidney cDNA to evaluate the gene model of 5' end of the *LOC115749* gene. Two transcripts were found in the experiment, and no new 5' end exons were found according to the NCBI database (Fig 3.2.4.4). This result confirmed the gene model that already exists in the NCBI database.

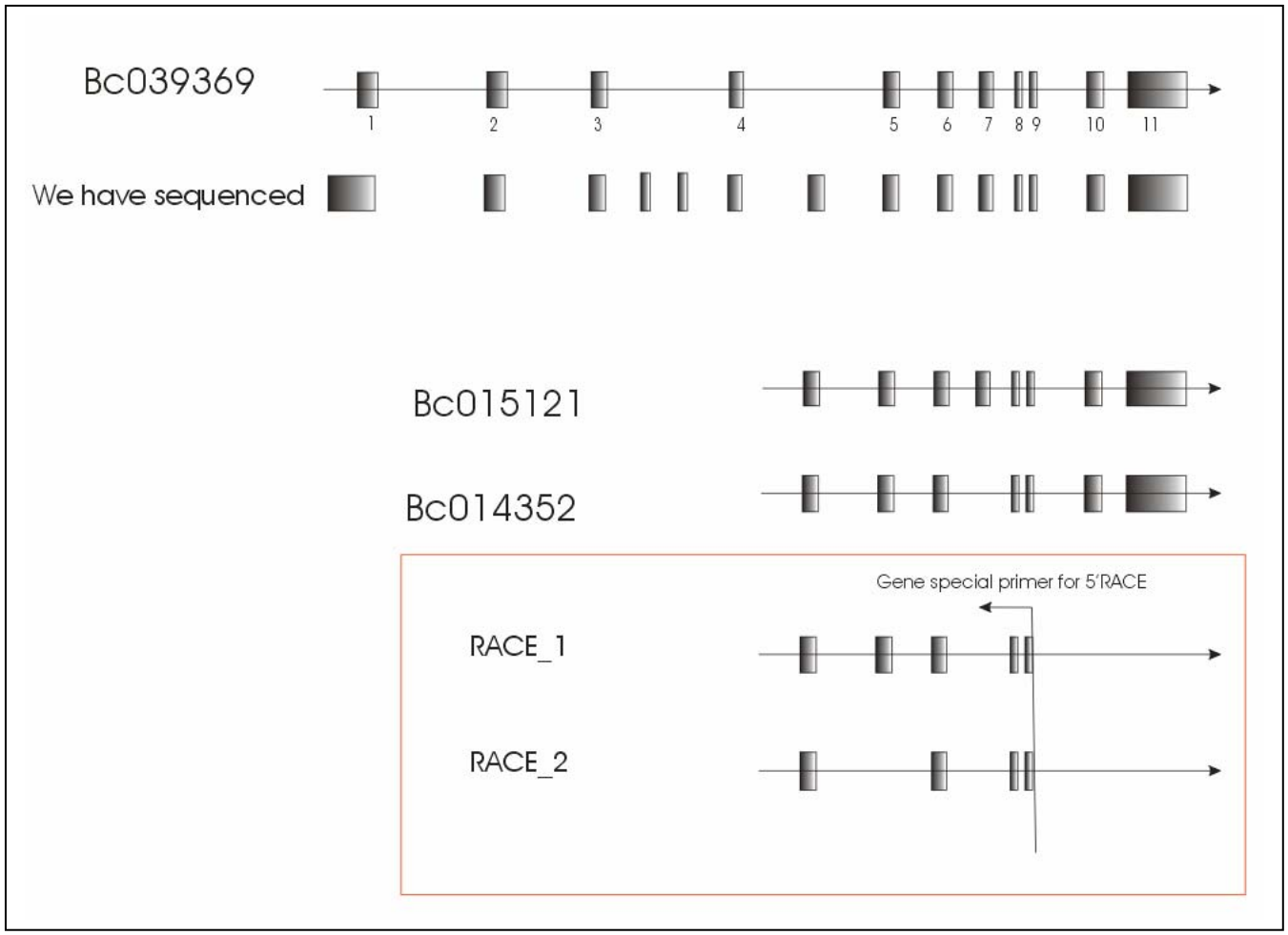


Fig 3.2.4.4 RACE results of the BC039369 gene. The RACE products obtained from kidney tissue are consistent with the gene model in the NCBI database.

4. Discussion

4.1 *AGR2* gene

In Xenopus, XAG2 has been shown to act as a signaling molecule and plays an important role in cement gland differentiation and ectodermal patterning [74]. XAG2 is highly expressed in the cement gland, which consists of mucus-secreting cells functioning as endocrine organs. Over-expression of XAG2 induces both cement gland differentiation and expression of other anterior neural marker genes [74]. The *AGR2* gene is the human homologue of the *Xenopus laevis* cement gland gene *XAG*, mapping to chromosome band 7p21.3 [114]. This gene is highly expressed in the trachea, lung, stomach, colon, prostate and small intestine [76]. In *Xenopus laevis*, the *XAG* families of genes are expressed in a gradient in the ectoderm during early development of the cement gland, and appear to be important factors during differentiation of this organ [115]. The cement gland arises from the outer layer of embryonic ectoderm and forms a cone of columnar epithelium. *AGR2* has been found to be expressed in tissues that contain mucous secreting cells and/or function as endocrine organs. Thus, from an evolutionary perspective, the human *AGR* genes may be involved in the epithelial barrier function.

Goblet cells reside throughout the length of the small and large intestine and are responsible for the production and maintenance of the protective mucus blanket by synthesizing and secreting high-molecular-weight glycoproteins known as mucins. Mucus is secreted by the epithelial surfaces throughout the gastrointestinal tract from the stomach to the colon. It forms a gel adherent to the surface that provides a protective barrier between the underlying

epithelium and the lumen. The mucus layer provides a protective barrier against pathogens by acting as a physical barrier, since the mucus layer provides binding sites for the bacterial adhesions. At the same time in the colon, the mucus layer provides an essential environment for the enteric microflora [116]. The involvement of goblet cell mucins in the pathophysiology of intestinal neoplasia and ulcerative colitis are presented [64].

Foxa transcription factors comprise a subfamily of forkhead transcription factors. The so-called forkhead box encodes a winged-helix DNA-binding motif, the name of which describes the structure of the domain when bound to DNA. The three Fox (forkhead box) group A genes, Foxa1, Foxa2 and Foxa3, are expressed in embryonic endoderm, the germ layer that gives rise to the digestive system. The Fox group A genes also contribute to the specification of the pancreas and the regulation of glucose homoeostasis. The transactivation domains in the C-terminal and N-terminal regions of Foxa1 and Foxa2 share structural similarity [117 118]. In mouse, Foxa2 is expressed first in the primitive streak at E6.5, shortly after the onset of gastrulation [119]. Thereafter it is expressed in the notochord, gut endoderm and ventral midline of the central nervous system. Later in embryonic development, Foxa2 is expressed in endodermally derived tissues including liver, lung, pancreas and intestine [120 121]. Foxa2 is an important regulator and deletion of the Foxa2 gene in pancreatic beta-cells in mice results in a phenotype resembling PHHI (persistent hyperinsulinaemic hypoglycaemia of infancy) [122]. Foxa1 was shown to be an essential activator of glucagon gene expression in vivo [123].

The hypothesis of involvement of the *AGR2* gene in epithelial barrier function is further supported by the regulation of the *AGR2* promoter by transcription factors typical for

epithelial goblet cells. Luciferase reporter gene assays show an activation of human *AGR2* promoter by *FOXA1* and *FOXA2*. *FOXA1* contributes to pancreatic beta-cell function [124] and both regulate signaling and transcriptional programs required for morphogenesis and goblet cell differentiation during formation [112 125]. There is a binding site for Hepatic Nuclear Factor 1(HNF1) in the *AGR2* promoter region at SNP 07AGRNP53. HNF1 and *FOXA1* and *FOXA2* belong to the same family.

We observed association of markers in the 5' region of *AGR2* primarily with the UC phenotype in two independent cohorts (UK and German extraction). All four significant SNPs were located in 5' end of the *AGR2* gene (Table 2). Additionally, 07AGRNP53 and 07AGRNP261, which are part of the risk haplotype, were located in promoter region. The expression level of *AGR2* in UC patients was significantly lower than in healthy controls. Also, a trend towards lower expression of *AGR2* in carriers of the risk alleles was observed. The link between down-regulation of the *AGR2* transcript in risk allele carriers and in disease has not yet fully been explored. None of the individual promoter SNPs identified in this study (Table 2) fully defined the disease haplotype, but also the more distant markers hcv1702494 and hcv111845 are needed for the definition of the risk haplotype. This suggests that further, unidentified private mutations contribute to the down-regulation of *AGR2* in disease.

4.2 Chromosome 12

4.2.1 Association mapping

Association mapping employing diallelic SNP markers was performed in the linkage region on chromosome 12q. In the beginning, 15 SNP markers within this region were genotyped in a German cohort, which included 776 trios with IBD (484 with CD, 292 with UC) and 360 unrelated healthy control individuals. The SNP markers were tested for allele and genotype association with the disease phenotypes IBD, CD and UC in a case-control study and a TDT association. Two SNP markers showed association with UC subgroup in the case-control study. However, none of markers was significant in the TDT association study. The two markers were hcv8757644 (P_{TDT}=0.088; $P_{case/control}$=0.0036) and hcv400963 (P_{TDT}=0.076; $P_{case/control}$=0.00057). For the SNP hcv8757644, no significant signal was seen in IBD and CD population. For the SNP hcv400963, no significant result was seen in CD subgroup and a weak significant result was seen in IBD population (P_{TDT}=0.44; $P_{case/control}$=0.02). These results showed that this region associated with UC subgroup.

For further study, an additional 35 SNP markers were selected within this region. The selected SNP markers covered the entire LD block in this region, and all SNPs were tagging SNPs according to the HapMap website (www.hapmap.org). The selected SNP markers were genotyped in the German cohort described above. Together with the 15 SNP markers that were described above, an overview of the whole region was obtained (Table 4.2.1, Fig 4.2.1a, Fig 4.2.1b). Two association leads were observed in case-control studies in UC subgroup. One lead was located between marker rs7955726 ($P_{case/control}$=0.00068) and marker hcv400963 ($P_{case/control}$=0.00056). This region was located around the *LOC115749* gene. Another lead was located between marker rs1657050 ($P_{case/control}$=0.0034) and marker hcv8757644 ($P_{case/control}$=0.0036). *FLJ32549* gene was located around this region.

One association peak was found in TDT studies in UC subgroup. This peak was located at marker rs7955726 (P_{TDT}=0.00229). This is the same marker that showed significance in the case-control studies, located around *LOC115749* gene. The overlap of the association results gave a good indication toward the potential location of a disease gene. By examining the LD between physically close markers, association to a disease-causing mutation could be identified through an SNP that is located nearby if it is positioned on the same haplotype.

Table 4.2.1 Overview of single point association statistics in the German cohort. Results that meet nominal p-value criterion of 0.05 are highlighted in bold print.

	IBD				CD				UC			
SNP	TDT	Case/control			TDT	Case/control			TDT	Case/control		
	P	P	Allele frequencies		P	P	Allele frequencies		P	P	Allele frequencies	
			cases	controls			cases	controls			cases	controls
rs7311377	0.434	0.709	0.43	0.44	0.138	0.211	0.41	0.45	0.522	0.475	0.46	0.44
rs1456046	0.922	0.706	0.39	0.38	0.527	0.899	0.39	0.39	0.535	0.634	0.40	0.38
rs10878102	0.886	0.596	0.36	0.35	0.755	0.742	0.36	0.35	0.553	0.572	0.36	0.34
rs1827073	0.874	0.218	0.10	0.08	0.666	0.681	0.08	0.08	0.480	**0.047**	0.11	0.08
hCV1521134	0.398	0.797	0.19	0.19	0.080	0.861	0.19	0.19	0.576	0.798	0.19	0.19
rs2279666	0.819	0.491	0.47	0.46	1	0.774	0.47	0.46	0.718	0.371	0.48	0.46
rs4763138	0.680	0.530	0.48	0.46	0.811	0.844	0.47	0.46	0.720	0.376	0.49	0.46
rs789722	0.774	0.554	0.47	0.48	0.952	0.984	0.49	0.48	0.576	0.265	0.45	0.49
hCV2690436	0.282	0.670	0.23	0.24	0.827	0.652	0.23	0.23	0.194	0.788	0.23	0.24
rs1146122	0.890	0.830	0.48	0.49	0.681	0.746	0.50	0.49	0.766	0.395	0.47	0.49
hCV2690456	0.292	0.344	0.28	0.26	0.947	0.400	0.28	0.27	0.144	0.430	0.28	0.26
rs10748002	1	0.623	0.47	0.48	0.498	0.815	0.48	0.48	0.421	0.610	0.47	0.48
rs10878111	0.963	0.711	0.47	.048	0.423	0.922	0.48	0.49	0.394	0.645	0.47	0.48
hCV1403120	0.358	0.709	0.50	0.50	0.662	0.407	0.48	0.50	0.391	0.715	0.48	0.50
BC036_SNP2	0.366	0.621	0.01	0.01	0.317	0.887	0.01	0.01	1	0.321	0.003	0.01
hcv8757697	0.930	0.882	0.48	0.50	0.799	0.885	0.48	0.50	0.903	0.913	0.48	0.49
rs1695105	0.963	0.791	0.48	0.50	0.757	0.570	0.47	0.50	0.775	0.814	0.49	0.49
rs790008	0.954	0.056	0.20	0.25	0.408	0.281	0.22	0.26	0.289	**0.028**	0.19	0.24
hcv8757644	0.351	0.188	0.25	0.28	0.734	0.942	0.28	0.29	0.088	**0.0036**	0.21	0.28
rs1464055	0.368	0.468	0.45	0.48	0.665	0.531	0.45	0.48	0.376	0.585	0.45	0.47
rs11175286	0.957	0.055	0.24	0.28	0.530	0.511	0.26	0.29	0.490	**0.005**	0.21	0.27
hcv2690409	0.787	0.075	0.38	0.33	0.671	0.230	0.37	0.32	0.946	**0.039**	0.39	0.34
rs1657050	1	**0.048**	0.24	0.28	0.526	0.539	0.26	0.29	0.439	**0.003**	0.20	0.27
rs1162974	0.234	0.402	0.22	0.22	0.746	0.677	0.19	0.21	**0.035**	**0.038**	0.25	0.21
hcv11290387	0.390	0.430	0.22	0.21	0.576	0.687	0.19	0.20	0.094	**0.03**	0.25	0.21
hcv11290388	0.261	0.361	0.22	0.21	0.695	0.828	0.20	0.21	0.057	**0.028**	0.25	0.21
hcv9277491	0.448	0.328	0.22	0.21	0.871	0.870	0.20	0.20	0.243	**0.025**	0.25	0.20
rs7954838	0.170	0.463	0.22	0.22	1	0.526	0.19	0.21	**0.044**	**0.034**	0.26	0.21
rs7955726	**0.029**	**0.034**	0.32	0.29	0.838	0.603	0.28	0.28	**0.002**	**0.0006**	0.36	0.28
rs2164504	0.395	**0.016**	0.40	0.46	0.902	0.253	0.43	0.47	0.140	**0.0016**	0.37	0.45
hcv400963	0.385	**0.019**	0.40	0.44	0.577	0.308	0.42	0.46	0.077	**0.0005**	0.35	0.44
rs6421214	0.731	0.550	0.45	0.43	0.213	0.207	0.47	0.43	0.374	0.768	0.42	0.43
rs10878167	0.488	**0.048**	0.21	0.25	0.391	0.252	0.22	0.25	1	**0.015**	0.19	0.25
rs11175383	0.914	0.361	0.26	0.24	0.521	0.652	0.25	0.24	0.361	0.191	0.28	0.24
hcv12069598	0.808	0.400	0.14	0.12	0.490	0.685	0.13	0.12	0.733	0.241	0.15	0.12
rs10784408	0.683	0.649	0.14	0.13	0.857	0.855	0.14	0.13	0.677	0.450	0.15	0.13
rs7976457	0.938	0.337	0.11	0.13	0.361	0.876	0.12	0.13	0.325	0.087	0.09	0.12
hcv491186	0.546	0.398	0.25	0.25	0.824	0.414	0.25	0.25	0.537	0.540	0.25	0.26
hcv22272931	0.654	0.610	0.24	0.25	0.835	0.855	0.24	0.25	0.668	0.404	0.23	0.25
hcv315870	0.415	0.500	0.45	0.43	0.250	0.342	0.46	0.42	1	0.933	0.44	0.43
rs1317532	0.143	0.432	0.48	0.46	0.159	0.379	0.48	0.47	0.551	0.758	0.47	0.46
rs1317535	0.101	0.535	0.32	0.34	0.091	0.632	0.32	0.34	0.590	0.653	0.32	0.34
rs6581575	0.639	0.213	0.48	0.45	0.903	0.389	0.47	0.46	0.379	0.192	0.49	0.45
rs7972780	0.638	0.395	0.50	0.47	0.668	0.542	0.49	0.48	0.211	0.390	0.50	0.47
rs4964110	0.334	0.262	0.09	0.10	0.249	0.274	0.09	0.10	0.900	0.420	0.09	0.10
rs10506538	0.824	0.189	0.11	0.09	0.283	0.491	0.10	0.09	0.359	0.084	0.12	0.09
rs7963840	1	0.917	0.37	0.38	0.950	0.965	0.38	0.38	0.940	0.787	0.37	0.37
rs1245035	0.229	0.890	0.40	0.40	0.479	0.482	0.41	0.40	0.310	0.554	0.38	0.40
rs1520765	0.448	0.871	0.47	0.46	0.616	0.402	0.49	0.47	0.562	0.513	0.45	0.46
rs1619280	1	0.998	0.46	0.47	0.808	0.393	0.44	0.46	0.777	0.320	0.49	0.46

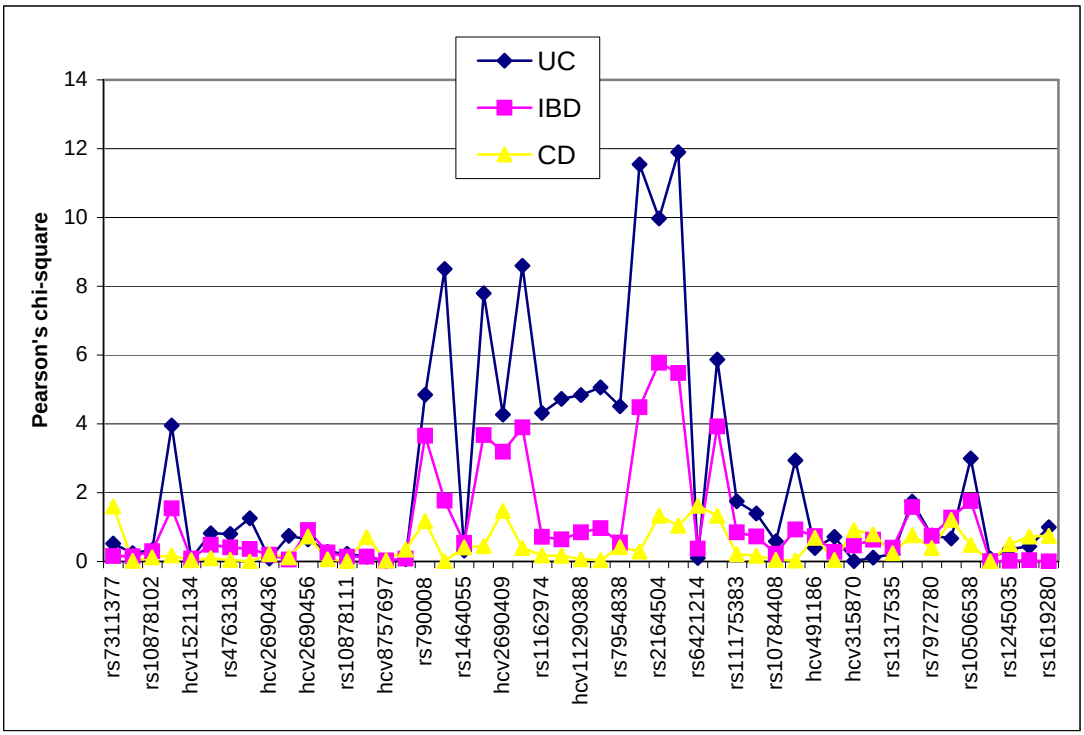


Fig 4.2.1a Pearson's $\chi 2$ in case-control studies. Significance of a peak at the 95% confidence interval is reached at value of 3.9 with 1df.

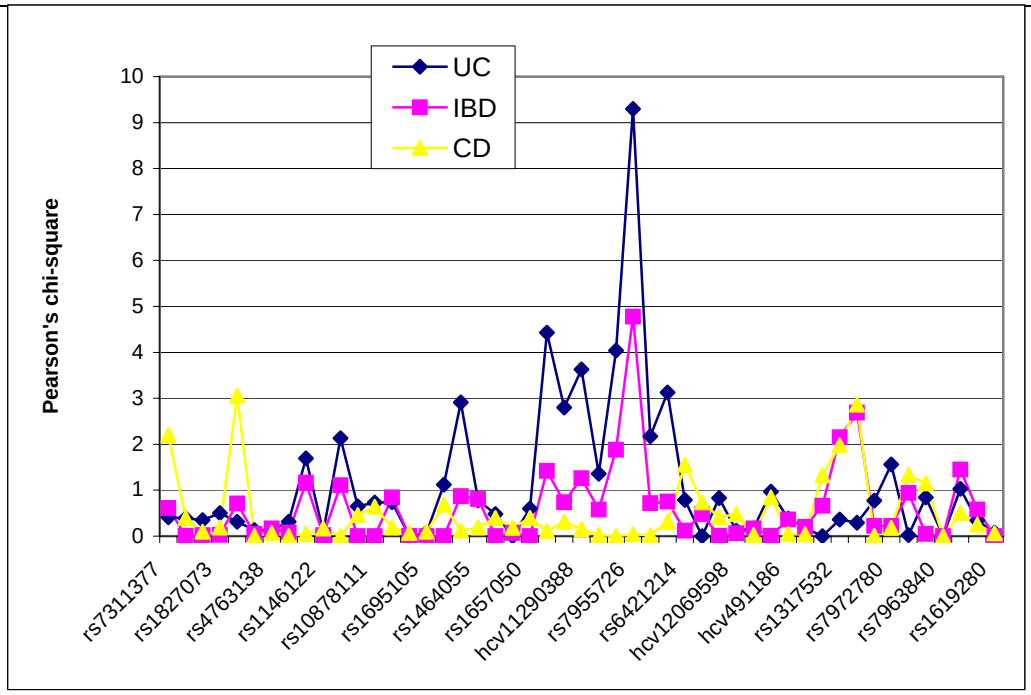


Fig 4.2.1b Pearson's $\chi 2$ in TDT association studies. Significance of a peak at the 95% confidence interval is reached at value of 3.9 with 1df.

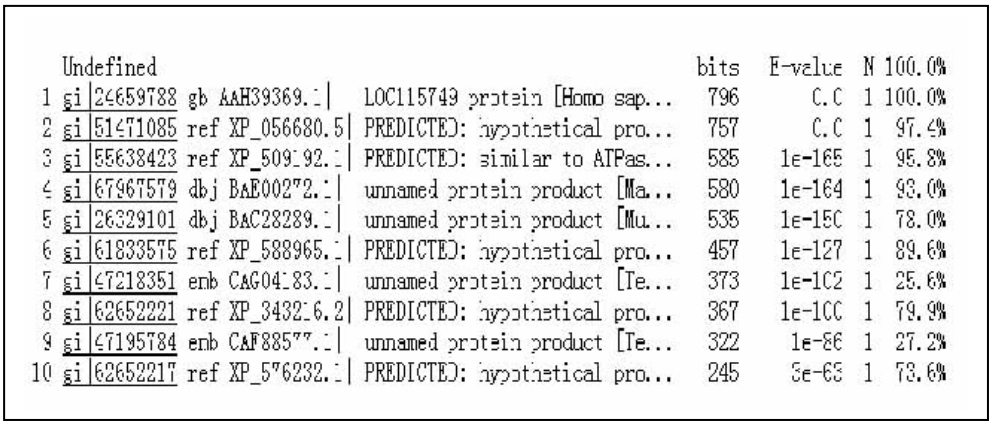
4.2.2 Candidate genes on chromosome 12

LOC115749 gene is one of the genes that could be a candidate gene due to the association in the surrounding markers. In both case-control studies and TDT association studies, the most significant SNP marker (rs7955726) was located in this gene.

The intron-exon structure of *LOC115749* gene known and yields an open reading frame of 444 amino acid. However, the functional of the protein is unknown. After using the Scanprosite program (http://us.expasy.org), several functional sites were predicted: five casein kinase II phosphorylation sites; eight protein kinase C phosphorylation sites; one tyrosine kinase phosphorylation site; one N-glycosylation site and one N-myristoylation

site. These potential functional sites may give some indication about the function of the *LOC115749* gene.

Another website (http://www.sbg.bio.ic.ac.uk/phyre) also used for predicting the function of *LOC115749* gene. This protein was 95.8% identity with the XP_509192 protein, which similar to the ATPase, H^+ transporting lysosome (Fig 4.2.2a). The ATPase plays an important role in IBD [126 127 128].



	Undefined		bits	E-value	N	100.0%
1	gi|24659788 gb AAH39369.1|	LOC115749 protein [Homo sap...	796	C.C	1	100.0%
2	gi|51471085 ref XP_056680.5|	PREDICTED: hypothetical pro...	757	C.C	1	97.4%
3	gi|55638423 ref XP_509192.1|	PREDICTED: similar to ATPas...	585	1e-165	1	95.8%
4	gi|67967579 dbj BAE00272.1|	unnamed protein product [Ma...	580	1e-164	1	93.0%
5	gi|26329101 dbj BAC28289.1|	unnamed protein product [Mu...	535	1e-150	1	78.0%
6	gi|61833575 ref XP_588965.1|	PREDICTED: hypothetical pro...	457	1e-127	1	89.6%
7	gi|47218351 emb CAG04183.1|	unnamed protein product [Te...	373	1e-102	1	25.6%
8	gi|62652221 ref XP_343216.2|	PREDICTED: hypothetical pro...	367	1e-100	1	79.9%
9	gi|47195784 emb CAF88577.1|	unnamed protein product [Te...	322	1e-86	1	27.2%
10	gi|62652217 ref XP_576232.1|	PREDICTED: hypothetical pro...	245	3e-63	1	73.6%

Fig 4.2.2a Protein prediction by www.sbg.bio.ic.ac.uk/phyre website.

According to the database, this gene has 11 exons. We sequenced all the exons in a set of 47 unrelated UC patients. The only exon polymorphism that was found was a synonymous mutation. This SNP showed weak association (P_{TDT}=0.172; $P_{case/control}$=0.025) with the UC patients in case-control studies. The most significant SNP (rs7955726) was located in the intron between exon 3 and exon 4 (an overview was shown in Fig 4.2.2b). It associated with two binding sites, one for NF-E2 p45 and another for activator protein 1 in its wild type.

The transcription factor NF-E2 is expressed in erythroid cells, megakaryocytes, and mast cells, and it has been shown to be a heterodimer formed between the large subunit (p45) and the small subunit (p18) [129] [130]. P45 belongs to a family of basic leucine-zipper protein [131]. The NF-E2 p45 is highly expressed in the erythroid and megakaryocytic lineages [129] [132] [133]. Micheal McMahon et al [134] demonstrated that the small intestine and stomach are organs where regulation of antioxidant responsive element (ARE)-driven genes may be particularly dependent on NF-E2.

The transcription factors activator protein (AP) 1 and nuclear factor (NF) κB have been reported to be crucial for the induction of genes involved in inflammation, as well as in a wide range of diseases originating from chronic activation of the immune system, such as inflammatory bowel disease [135] [136] [137] [138]. The transcription factor AP1 is encoded by protooncogenes and regulates various aspects of cell proliferation and differentiation [139] [140] [141]. The regulation of AP1 activity is complex. Two genes (Jun and Fos) are involved [142] [143] [144]. Many stimuli, such as physiological agents, bacterial and viral infections, pharmacological compounds and cellular stress can induce AP1 activity.

The most significant SNP (rs7955726) can influence these two binding sites. In the genotype of wild-type (G), the binding sites for NF-E2 p45 and AP1 were intact. But for the other genotype (T), those two binding sites were disrupted.

The RT-PCR of a tissue panel showed that this gene is highly expressed in thymus. The thymus plays crucial role in immune system. This also might be a hint to function studies.

FLJ32549 gene is another candidate gene within this region. It is a known gene with hypothetical protein product. It contains 4 exons. One significant SNP marker (hcv8757644, P_{TDT}=0.088; $P_{case/control}$=0.0036) associated with UC patients in case-control study was located between exon1 and exon2. However, the TDT association studies did not yield the same result. All the other SNP markers in this gene did not show significant results. *FLJ32549* gene may only have a weak effect on UC patients.

BC042855 gene is the third candidate gene within this region. It has 5 exons without protein product. Two SNPs (rs11175286, P_{TDT}=0.48951; $P_{case/control}$=0.0053; rs1657050, P_{TDT}=0.43858; $P_{case/control}$=0.0034) in this gene showed significant results in UC patients in case-control studies. TDT gave disparate results for these two markers.

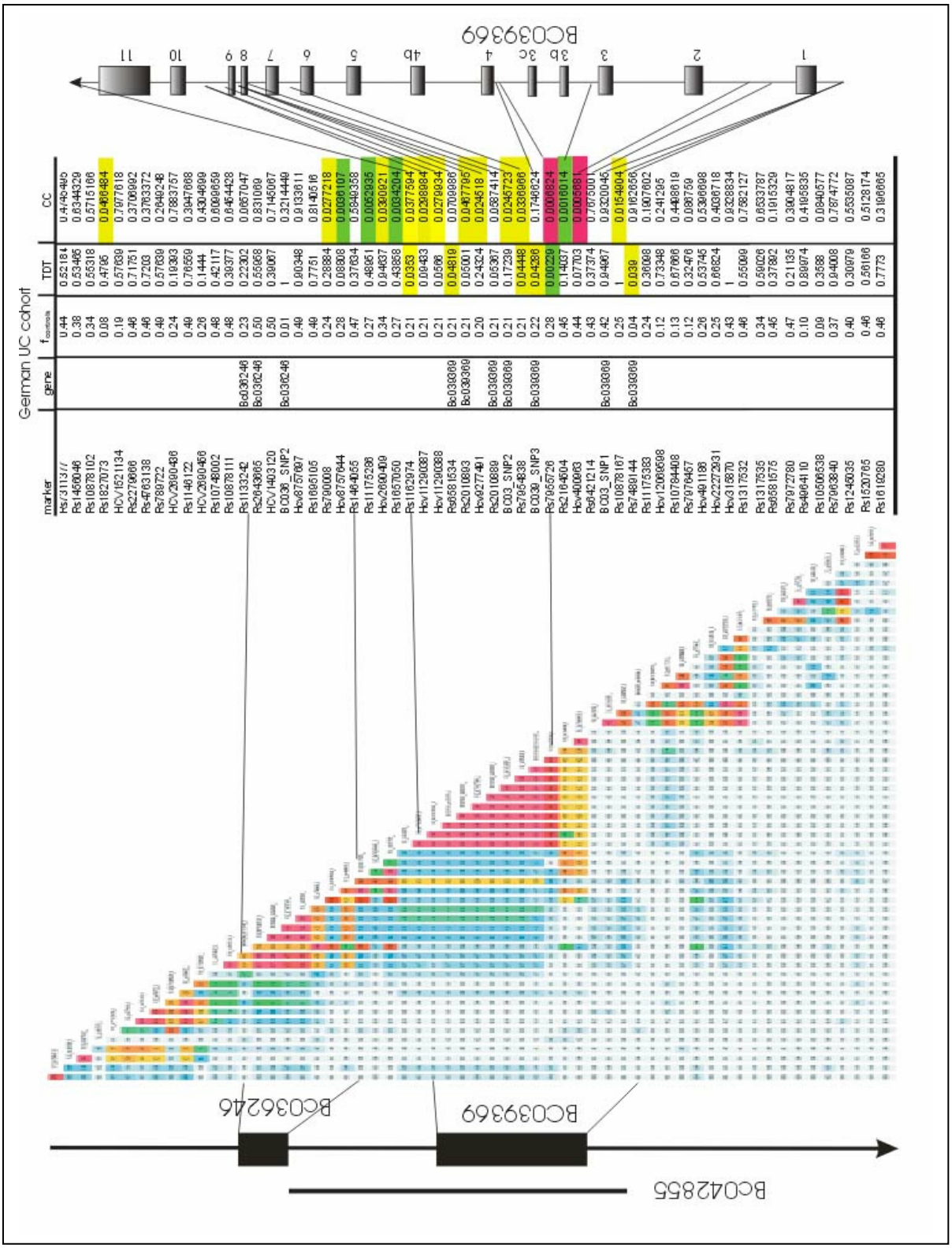


Fig 4.2.2b Overview of the linkage disequilibrium structure of the chromosome 12q12 region.

5. Conclusions

We use association mapping with diallelic SNP markers to refine a previously identified linkage region on chromosome 12. Candidate gene analysis was used for identification of susceptibility gene for inflammatory bowel disease on chromosome 12 and chromosome 7. Two kinds of statistical methods were used. The case-control study is based on identity by state of the disease allele and the marker used for analysis in the group of independent cases, while the TDT association study identifies the transmission of alleles from parents to offspring.

The AGR gene is located on chromosome 7p2.1, which is a linkage region for inflammatory bowel disease. Thirty SNP markers of *AGR2* and *AGR3* gene were genotyped in a German cohort and the significant SNP markers were verified in UK cohort. The association is most pronounced at hcv1702494 and hcv111845, which is located on the 5' region of *AGR2* gene for UC phenotype. Functional studies were performed on the *AGR2* gene. Those data demonstrate association of the 5'region of the *AGR2* gene to the UC phenotype in two independent populations. Functionally, the gene may be involved in the maintenance of epithelial integrity based on the mouse model, phylogenetic background and activation by transcription factors, which are characteristic for epithelial goblet cells. The disease effect is likely to be mediated through down-regulation of the *AGR2* transcript in disease, as suggested by association to individual promoter SNPs. The mechanistic risk profile of the risk haplotype is functionally not yet fully explored and possibly includes further private mutations in more distant regulatory elements. The impact of the risk mutations on the overall phenotype is moderate, as

indicated by allelic odds ratios in the range of 1.3 – 1.4. Overall, the *AGR2* represents an interesting new avenue into the etiopathophysiology of IBD and warrants further evaluation in additional, independent populations.

Fifty diallelic SNP markers were used to identify the inflammatory bowel disease linkage region on chromosome 12q. Significant results both in case-control study and TDT association study were obtained. These results showed strong association with UC subgroup.

Three candidate genes were located in this region. The major disease-associated SNP rs7955726 was located in the intron between exon3 and exon 4 of *LOC115749* gene. The function of *LOC115749* gene is unknown. The significant SNP marker (rs7955726) influenced the binding site for NF-E2 p45 and AP1. This gene is highly expressed in thymus and the protein was 95.8% identity with the XP_509192 protein, which similar to the ATPase, H^+ transporting lysosome. The ATPase plays an important role in IBD The other two candidate genes only show weak association with UC phenotype. It seems that *LOC115749* gene might be a disease gene for UC but the mechanistic details were not known yet.

6. Summary

Genome wide linkage analyses have implicated chromosome 7p21.3 and chromosome 12q14 as a susceptibility region for IBD. Recently, the mouse phenotype with diarrhea and goblet cell dysfunction caused by anterior gradient protein 2 dysfunction was reported (European patent WO2004056858). The genes encoding for the human homologues anterior gradient proteins 2 and 3 (*AGR2* and *AGR3*) are located on chromosome 7p21.3. The gene structures of human *AGR2* and *AGR3* were verified and exhaustive mutation detection was performed in 46 individuals with IBD. Thirty SNPs were tested for association to ulcerative colitis (UC, N=317) and Crohn's disease (CD, N=631) in a German cohort and verified in a (UK) cohort of 384 CD and 311 UC patients. An association signal was identified in the 5' region of the *AGR2* gene. *AGR2* was down-regulated in UC patients as compared to normal controls. Luciferase assays of the *AGR2* promoter showed regulation by the goblet-cell specific transcription factors *FOXA1* and *FOXA2*. In summary, *AGR2* represents an interesting new avenue into the aetiopathophysiology of IBD and the maintenance of epithelial integrity. Fifty SNPs within chromosome 12q14 region were genotyped in a German cohort (484 with CD, 292 with UC) to confine the region where a potential disease susceptibility gene could be located. The most significant SNP marker was rs7955726, which located between exon 3 and exon 4 of *LOC115749* gene. *LOC115749* gene was highly expressed in thymus and highly identity to the ATPase proteins. Mutation detection and genotyping of *LOC115749* gene only gave weak significant results. The rs7955726 influenced the binding site of NF-E2 and AP1. The *LOC115749* gene might be a candidate gene for UC patients but the mechanistic details were not known yet.

7. Zusammenfassung

Genomweite Kopplungsstudien haben Chromosom 7p21.3 und Chromosom 12q14 als Kandidatenregionen für chronisch entzündliche Darmerkrankungen (CED) identifiziert. Vor kurzem wurde ein Mausphänotyp mit Diarrhö und Becherzellenfehlfunktion beschrieben (Europäisches Patent WO2004056858), was durch eine Fehlfunktion im Anterior Gradienten Protein 2 verursacht wird. Die Gene, welche für die menschlichen Homologe der Anterior Gradienten Proteine 2 und 3 kodieren (*AGR2* and *AGR3*), liegen auf Chromosom 7p21.3. Die Genstrukturen von *AGR2* und *AGR3* wurden verifiziert, und es wurde eine Mutationsdetektion in 46 Individuen mit CED durchgeführt. Die Assoziation von 30 SNPs mit Ulcerativer Colitis (UC, N=317) und Morbus Crohn (CD, N=631) wurde in einer deutschen Kohorte getestet und und in einer englischen (UK) Kohorte von 384 CD und 311 UC Patienten verifiziert. Es konnte ein Assoziationssignal in der 5'-Region des *AGR2* Genes identifiziert werden. *AGR2* ist in UC-Patienten herunterreguliert, verglichen mit gesunden Kontrollen. Luciferase Analysen des *AGR2* Promotors zeigten eine Regulation durch die Becherzell-spezifischen Transkriptionsfaktoren *FOXA1* und *FOXA2*. AGR2 liefert neue Einblicke in die Pathogenese von CED und in die Aufrechterhaltung des epithelialem Zusammenhalts. In der Region auf Chromosom 12q14 wurden 50 SNPs in einer deutschen Kohorte genotypisiert (484 mit CD, 292 mit UC), um die Region auf ein potentielles Krankheitsgen einzuschränken. Der am stärksten signifikante SNP (rs7955726), lokalisierte zwischen Exon 3 und Exon 4 des BC039369 Genes. Das BC039369 Gen wird am stärksten im Thymus exprimiert und hat hohe Ähnlichkeit mit ATPase Proteinen. Die Mutationsdetektion und Genotypisierung von BC039369 ergab nur schwach signifikante

Ergebnisse. Der SNP rs7955726 beeinflusst die Bindungsstelle von NF-E2 und AP1. BC039369 könnte ein Kandidatengen für Patienten mit UC sein, die zugrunde liegenden Mechanismen konnten aber bisher nicht aufgeklärt werden.

8. References

1. Shivananda, S. et al. Incidence of inflammatory bowel disease across Europe: is there a difference between north and south? Results of the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD). *Gut* **39**, 690-7. (1996).

2. Probert, C.S., Jayanthi, V., Rampton, D.S. & Mayberry, J.F. Epidemiology of inflammatory bowel disease in different ethnic and religious groups: limitations and aetiological clues. *Int J Colorectal Dis* **11**, 25-8. (1996).

3. Podolsky, D.K. Inflammatory bowel disease (1). *N Engl J Med* **325**, 928-37. (1991).

4. Fiocchi, C. Inflammatory bowel disease: etiology and pathogenesis. *Gastroenterology* **115**, 182-205. (1998).

5. Tysk, C., Lindberg, E., Jarnerot, G. & Floderus-Myrhed, B. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut* **29**, 990-6. (1988).

6. Orholm, M. et al. Familial occurrence of inflammatory bowel disease. *N Engl J Med* **324**, 84-8. (1991).

7. Munkholm, P. Crohn's disease--occurrence, course and prognosis. An epidemiologic cohort-study. *Dan Med Bull* **44**, 287-302. (1997).

8. Curran, M.E. et al. Genetic analysis of inflammatory bowel disease in a large European cohort supports linkage to chromosomes 12 and 16. *Gastroenterology* **115**, 1066-71. (1998).

9. Hugot, J.P. et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599-603. (2001).

10. Ogura, Y. et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603-6. (2001).

11. Hampe, J. et al. Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet* **357**, 1925-8. (2001).

12. Schreiber, S. Genetics of inflammatory bowel disease: a puzzle with contradictions? *Gut* **47**, 746-7. (2000).

13. Sartor, R. Enteric microflora in IBD: pathogens or comensals? *Inflamm Bowel Dis* **3**, 230-235 (1997).

14. Shanahan, F. Probiotics and inflammatory bowel disease: is there a scientific rationale? *Inflamm Bowel Dis* **6**, 107-15. (2000).

15. Elson, C.O. Commensal bacteria as targets in Crohn's disease. *Gastroenterology* **119**, 254-7. (2000).

16. Beutler, B. Autoimmunity and apoptosis: the Crohn's connection. *Immunity* **15**, 5-14. (2001).

17. Papadakis, K.A. & Targan, S.R. The role of chemokines and chemokine receptors in mucosal inflammation. *Inflamm Bowel Dis* **6**, 303-13. (2000).

18. Papadakis, K.A. & Targan, S.R. Role of cytokines in the pathogenesis of inflammatory bowel disease. *Annu Rev Med* **51**, 289-98. (2000).

19. Fuss, I.J. et al. Disparate CD4+ lamina propria (LP) lymphokine secretion profiles in inflammatory bowel disease. Crohn's disease LP cells manifest increased secretion of IFN-gamma, whereas ulcerative colitis LP cells manifest increased secretion of IL-5. *J Immunol* **157**, 1261-70. (1996).

20. Plevy, S.E. et al. A role for TNF-alpha and mucosal T helper-1 cytokines in the pathogenesis of Crohn's disease. *J Immunol* **159**, 6276-82. (1997).

21. Stallmach, A., Strober, W., MacDonald, T.T., Lochs, H. & Zeitz, M. Induction and modulation of gastrointestinal inflammation. *Immunol Today* **19**, 438-41. (1998).

22. Murch, S.H., Braegger, C.P., Walker-Smith, J.A. & MacDonald, T.T. Location of tumour necrosis factor alpha by immunohistochemistry in chronic inflammatory bowel disease. *Gut* **34**, 1705-9. (1993).

23. Nikolaus, S. et al. Increased secretion of pro-inflammatory cytokines by circulating polymorphonuclear neutrophils and regulation by interleukin 10 during intestinal inflammation. *Gut* **42**, 470-6. (1998).

24. Brandt, E. et al. Enhanced production of IL-8 in chronic but not in early ileal lesions of Crohn's disease (CD). *Clin Exp Immunol* **122**, 180-5. (2000).

25. Schreiber, S. et al. Tumour necrosis factor alpha and interleukin 1beta in relapse of Crohn's disease. *Lancet* **353**, 459-61. (1999).

26. Vecchi, M. et al. Antibodies to neutrophil cytoplasm in Italian patients with ulcerative colitis: sensitivity, specificity and recognition of putative antigens. *Digestion* **55**, 34-9. (1994).

27. Terjung, B., Spengler, U., Sauerbruch, T. & Worman, H.J. "Atypical p-ANCA" in IBD and hepatobiliary disorders react with a 50-kilodalton nuclear envelope

protein of neutrophils and myeloid cell lines. *Gastroenterology* **119**, 310-22. (2000).

28. Bonen, D.K. & Cho, J.H. The genetics of inflammatory bowel disease. *Gastroenterology* **124**, 521-36. (2003).

29. Kuster, W., Pascoe, L., Purrmann, J., Funk, S. & Majewski, F. The genetics of Crohn disease: complex segregation analysis of a family study with 265 patients with Crohn disease and 5,387 relatives. *Am J Med Genet* **32**, 105-8. (1989).

30. Satsangi, J., Rosenberg, W.M. & Jewell, D.P. The prevalence of inflammatory bowel disease in relatives of patients with Crohn's disease. *Eur J Gastroenterol Hepatol* **6**, 413-416 (1994).

31. Meucci, G. et al. Familial aggregation of inflammatory bowel disease in northern Italy: a multicenter study. The Gruppo di Studio per le Malattie Infiammatorie Intestinali (IBD Study Group). *Gastroenterology* **103**, 514-9. (1992).

32. Schreiber, S., Rosenstiel, P., Albrecht, M., Hampe, J. & Krawczak, M. Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat Rev Genet* **6**, 376-88. (2005).

33. Thompson, N.P., Driscoll, R., Pounder, R.E. & Wakefield, A.J. Genetics versus environment in inflammatory bowel disease: results of a British twin study. *Bmj* **312**, 95-6. (1996).

34. Binder, V. Genetic epidemiology in inflammatory bowel disease. *Dig Dis* **16**, 351-5. (1998).

35. Satsangi, J. et al. Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nat Genet* **14**, 199-202. (1996).

36. Hugot, J.P. et al. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* **379**, 821-3. (1996).

37. Hampe, J. et al. Linkage of inflammatory bowel disease to human chromosome 6p. *Am J Hum Genet* **65**, 1647-55. (1999).

38. Rioux, J.D. et al. Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *Am J Hum Genet* **66**, 1863-70. Epub 2000 Apr 21. (2000).

39. Ma, Y. et al. A genome-wide search identifies potential new susceptibility loci for Crohn's disease. *Inflamm Bowel Dis* **5**, 271-8. (1999).

40. Duerr, R.H., Barmada, M.M., Zhang, L., Pfutzer, R. & Weeks, D.E. High-density genome scan in Crohn disease shows confirmed linkage to chromosome 14q11-12. *Am J Hum Genet* **66**, 1857-62. Epub 2000 Apr 3. (2000).

41. Rioux, J.D. et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat Genet* **29**, 223-8. (2001).

42. Cho, J.H. et al. Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and IBD1. *Proc Natl Acad Sci U S A* **95**, 7502-7. (1998).

43. Vermeire, S. et al. Evidence for inflammatory bowel disease of a susceptibility locus on the X chromosome. *Gastroenterology* **120**, 834-40. (2001).

44. Koutroubakis, I., Manousos, O.N., Meuwissen, S.G. & Pena, A.S. Environmental risk factors in inflammatory bowel disease. *Hepatogastroenterology* **43**, 381-93. (1996).

45. Orholm, M. et al. Investigation of inheritance of chronic inflammatory bowel diseases by complex segregation analysis. *Bmj* **306**, 20-4. (1993).

46. Hampe, J. et al. Evidence for a NOD2-independent susceptibility locus for inflammatory bowel disease on chromosome 16p. *Proc Natl Acad Sci U S A* **99**, 321-6. Epub 2001 Dec 18. (2002).

47. Satsangi, J. et al. Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease. *Lancet* **347**, 1212-7. (1996).

48. Akolkar, P. et al. Fine mapping of regions linked to Crohn's disease on chromosomes 12 and 16. *Am J Hum genet Suppl* **63**, A279 (1998).

49. Duerr, R.H. et al. Linkage and association between inflammatory bowel disease and a locus on chromosome 12. *Am J Hum Genet* **63**, 95-100. (1998).

50. Hampe, J. et al. A genomewide analysis provides evidence for novel linkages in inflammatory bowel disease in a large European cohort. *Am J Hum Genet* **64**, 808-16. (1999).

51. Clayton, D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* **65**, 1170-7. (1999).

52. Brant, S.R. et al. American families with Crohn's disease have strong evidence for linkage to chromosome 16 but not chromosome 12. *Gastroenterology* **115**, 1056-61. (1998).

53. Rioux, J.D. et al. Absence of linkage between inflammatory bowel disease and selected loci on chromosomes 3, 7, 12, and 16. *Gastroenterology* **115**, 1062-5. (1998).

54. Vermeire, S. et al. Exclusion of linkage of Crohn's disease to previously reported regions on chromosomes 12, 7, and 3 in the Belgian population indicates genetic heterogeneity. *Inflamm Bowel Dis* **6**, 165-70. (2000).

55. Parkes, M. et al. The IBD2 locus shows linkage heterogeneity between ulcerative colitis and Crohn disease. *Am J Hum Genet* **67**, 1605-10. Epub 2000 Nov 10. (2000).

56. Yang, H. et al. Linkage of Crohn's disease to the major histocompatibility complex region is detected by multiple non-parametric analyses. *Gut* **44**, 519-26. (1999).

57. Dimon, C., Allen, M. & Van Heel, D. Family based association studies of STAT6, a positional candidate gene for IBD. *Gastroenteroloty* **Suppl**, 2333 (2001).

58. Brant, S.R. et al. MDR1 Ala893 polymorphism is associated with inflammatory bowel disease. *Am J Hum Genet* **73**, 1282-92. Epub 2003 Nov 7. (2003).

59. Martin, K., Radlmayr, M., Borchers, R., Heinzlmann, M. & Folwaczny, C. Candidate genes colocalized to linkage regions in inflammatory bowel disease. *Digestion* **66**, 121-6. (2002).

60. Ishihara, K. & Hirano, T. IL-6 in autoimmune disease and chronic inflammatory proliferative disease. *Cytokine Growth Factor Rev* **13**, 357-68. (2002).

61. Atuma, C., Strugala, V., Allen, A. & Holm, L. The adherent gastrointestinal mucus gel layer: thickness and physical state in vivo. *Am J Physiol Gastrointest Liver Physiol* **280**, G922-9. (2001).

62. Dekker, J., Rossen, J.W., Buller, H.A. & Einerhand, A.W. The MUC family: an obituary. *Trends Biochem Sci* **27**, 126-31. (2002).

63. Verdugo, P. Goblet cells secretion and mucogenesis. *Annu Rev Physiol* **52**, 157-76. (1990).

64. Specian, R.D. & Oliver, M.G. Functional biology of intestinal goblet cells. *Am J Physiol* **260**, C183-93. (1991).

65. Podolsky, D.K. Mechanisms of regulatory peptide action in the gastrointestinal tract: trefoil peptides. *J Gastroenterol* **35**, 69-74. (2000).

66. Verdugo, P. Mucin exocytosis. *Am Rev Respir Dis* **144**, S33-7. (1991).

67. Nadel, J.A. Role of epidermal growth factor receptor activation in regulating mucin synthesis. *Respir Res* **2**, 85-9. Epub 2001 Feb 21. (2001).

68. van Den Brink, G.R., de Santa Barbara, P. & Roberts, D.J. Development. Epithelial cell differentiation--a Mather of choice. *Science* **294**, 2115-6. (2001).

69. Yang, Q., Bermingham, N.A., Finegold, M.J. & Zoghbi, H.Y. Requirement of Math1 for secretory cell lineage commitment in the mouse intestine. *Science* **294**, 2155-8. (2001).

70. Corfield, A.P., Carroll, D., Myerscough, N. & Probert, C.S. Mucins in the gastrointestinal tract in health and disease. *Front Biosci* **6**, D1321-57. (2001).

71. Einerhand, A.W. et al. Role of mucins in inflammatory bowel disease: important lessons from experimental models. *Eur J Gastroenterol Hepatol* **14**, 757-65. (2002).

72. Jass, J.R. & Walsh, M.D. Altered mucin expression in the gastrointestinal tract: a review. *J Cell Mol Med* **5**, 327-51. (2001).

73. Komiya, T., Tanigawa, Y. & Hirohashi, S. Cloning of the gene gob-4, which is expressed in intestinal goblet cells in mice. *Biochim Biophys Acta* **1444**, 434-8. (1999).

74. Aberger, F., Weidinger, G., Grunz, H. & Richter, K. Anterior specification of embryonic ectoderm: the role of the Xenopus cement gland-specific gene XAG-2. *Mech Dev* **72**, 115-30. (1998).

75. Higgins, D.G., Bleasby, A.J. & Fuchs, R. CLUSTAL V: improved software for multiple sequence alignment. *Comput Appl Biosci* **8**, 189-91. (1992).

76. Thompson, D.A. & Weigel, R.J. hAG-2, the human homologue of the Xenopus laevis cement gland gene XAG-2, is coexpressed with estrogen receptor in breast cancer cell lines. *Biochem Biophys Res Commun* **251**, 111-6. (1998).

77. Liang, F. et al. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat Genet* **25**, 239-40. (2000).

78. Roest Crollius, H. et al. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat Genet* **25**, 235-8. (2000).

79. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* **25**, 232-4. (2000).

80. Morton, N.E. Sequential tests for the detection of linkage. *Am J Hum Genet* **7**, 277-318. (1955).

81. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* **58**, 1347-63. (1996).

82. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11**, 241-7. (1995).

83. Kruglyak, L. & Lander, E.S. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* **57**, 439-54. (1995).

84. Spielman, R.S., McGinnis, R.E. & Ewens, W.J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**, 506-16. (1993).

85. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-7. (1996).

86. Spielman, R.S. & Ewens, W.J. The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* **59**, 983-9. (1996).

87. Ott, J. Statistical properties of the haplotype relative risk. *Genet Epidemiol* **6**, 127-30. (1989).

88. Terwilliger, J.D. & Ott, J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered* **42**, 337-46. (1992).

89. Zhao, H. et al. Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* **67**, 936-46. Epub 2000 Aug 31. (2000).

90. Perou, C.M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747-52. (2000).

91. Scherf, U. et al. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* **24**, 236-44. (2000).

92. Welsh, J.B. et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc Natl Acad Sci U S A* **98**, 1176-81. (2001).

93. Long, A.D. & Langley, C.H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* **9**, 720-31. (1999).

94. Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A. & Contopoulos-Ioannidis, D.G. Replication validity of genetic association studies. *Nat Genet* **29**, 306-9. (2001).

95. Sachidanandam, R. et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-33. (2001).

96. Lalouel, J.M. & Rohrwasser, A. Power and replication in case-control studies. *Am J Hypertens* **15**, 201-5. (2002).

97. Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847-56. (2000).

98. Lennard-Jones, J.E. Classification of inflammatory bowel disease. *Scand J Gastroenterol Suppl* **170**, 2-6; discussion 16-9. (1989).

99. Stoll, M. et al. Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nat Genet* **36**, 476-80. Epub 2004 Apr 11. (2004).

100. Chomczynski, P., Mackey, K., Drews, R. & Wilfinger, W. DNAzol: a reagent for the rapid isolation of genomic DNA. *Biotechniques* **22**, 550-3. (1997).

101. Livak, K.J., Flood, S.J., Marmaro, J., Giusti, W. & Deetz, K. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl* **4**, 357-62. (1995).

102. Dieffenbach, C.W., Lowe, T.M. & Dveksler, G.S. General concepts for PCR primer design. *PCR Methods Appl* **3**, S30-7. (1993).

103. Lowe, T., Sharefkin, J., Yang, S.Q. & Dieffenbach, C.W. A computer program for selection of oligonucleotide primers for polymerase chain reactions. *Nucleic Acids Res* **18**, 1757-61. (1990).

104. Saiki, R.K. et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487-91. (1988).

105. Williams, J.F. Optimization strategies for the polymerase chain reaction. *Biotechniques* **7**, 762-9. (1989).

106. Manaster, C. et al. InSNP: a tool for automated detection and visualization of SNPs and InDels. *Hum Mutat* **26**, 11-9. (2005).

107. Hampe, J. et al. An integrated system for high throughput TaqMan based SNP genotyping. *Bioinformatics* **17**, 654-5. (2001).

108. Hardy, G. Mendelian proportions in a mixed population. *Science* **28**, 49-50 (1908).

109. Weinberg, W. On the demonstration of heredity in man. *Naturkunde in Wurttemberg, Stuttgart* **64**, 368-382 (1908).

110. Krawczak, M. et al. Allelic association of the cystic fibrosis locus and two DNA markers, XV2c and KM19, in 55 German families. *Hum Genet* **80**, 78-80. (1988).

111. Rausa, F.M., Tan, Y. & Costa, R.H. Association between hepatocyte nuclear factor 6 (HNF-6) and FoxA2 DNA binding domains stimulates FoxA2 transcriptional activity but inhibits HNF-6 DNA binding. *Mol Cell Biol* **23**, 437-49. (2003).

112. Wan, H. et al. Foxa2 regulates alveolarization and goblet cell hyperplasia. *Development* **131**, 953-64. (2004).

113. Reich, D.E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199-204. (2001).

114. Petek, E., Windpassinger, C., Egger, H., Kroisel, P.M. & Wagner, K. Localization of the human anterior gradient-2 gene (AGR2) to chromosome band 7p21.3 by radiation hybrid mapping and fluorescencein situ hybridisation. *Cytogenet Cell Genet* **89**, 141-2. (2000).

115. Sive, H.L., Hattori, K. & Weintraub, H. Progressive determination during formation of the anteroposterior axis in Xenopus laevis. *Cell* **58**, 171-80. (1989).

116. Macfarlane, G.T., Allison, C., Gibson, S.A. & Cummings, J.H. Contribution of the microflora to proteolysis in the human large intestine. *J Appl Bacteriol* **64**, 37-46. (1988).

117. Qian, X. & Costa, R.H. Analysis of hepatocyte nuclear factor-3 beta protein domains required for transcriptional activation and nuclear targeting. *Nucleic Acids Res* **23**, 1184-91. (1995).

118. Pani, L., Quian, X.B., Clevidence, D. & Costa, R.H. The restricted promoter activity of the liver transcription factor hepatocyte nuclear factor 3 beta involves a cell-specific factor and positive autoactivation. *Mol Cell Biol* **12**, 552-62. (1992).

119. Sasaki, H. & Hogan, B.L. Differential expression of multiple fork head related genes during gastrulation and axial pattern formation in the mouse embryo. *Development* **118**, 47-59. (1993).

120. Ang, S.L. et al. The formation and maintenance of the definitive endoderm lineage in the mouse: involvement of HNF3/forkhead proteins. *Development* **119**, 1301-15. (1993).

121. Monaghan, A.P., Kaestner, K.H., Grau, E. & Schutz, G. Postimplantation expression patterns indicate a role for the mouse forkhead/HNF-3 alpha, beta and gamma genes in determination of the definitive endoderm, chordamesoderm and neuroectoderm. *Development* **119**, 567-78. (1993).

122. Lantz, K.A. et al. Foxa2 regulates multiple pathways of insulin secretion. *J Clin Invest* **114**, 512-20. (2004).

123. Sharma, S.K. et al. Characterization of a novel Foxa (hepatocyte nuclear factor-3) site in the glucagon promoter that is conserved between rodents and humans. *Biochem J* **389**, 831-41. (2005).

124. Lantz, K.A. & Kaestner, K.H. Winged-helix transcription factors and pancreatic development. *Clin Sci (Lond)* **108**, 195-204. (2005).

125. Wan, H. et al. Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis. *J Biol Chem* **24**, 24 (2005).

126. Musch, M.W. et al. T cell activation causes diarrhea by increasing intestinal permeability and inhibiting epithelial Na+/K+-ATPase. *J Clin Invest* **110**, 1739-47. (2002).

127. Allgayer, H. et al. Inverse relationship between colonic (Na+ + K+)-ATPase activity and degree of mucosal inflammation in inflammatory bowel disease. *Dig Dis Sci* **33**, 417-22. (1988).

128. Scheurlen, C., Allgayer, H., Hardt, M. & Kruis, W. Effect of short-term topical corticosteroid treatment on mucosal enzyme systems in patients with distal inflammatory bowel disease. *Hepatogastroenterology* **45**, 1539-45. (1998).

129. Andrews, N.C., Erdjument-Bromage, H., Davidson, M.B., Tempst, P. & Orkin, S.H. Erythroid transcription factor NF-E2 is a haematopoietic-specific basic-leucine zipper protein. *Nature* **362**, 722-8. (1993).

130. Igarashi, K. et al. Regulation of transcription by dimerization of erythroid factor NF-E2 p45 with small Maf proteins. *Nature* **367**, 568-72. (1994).

131. Mohler, J., Vani, K., Leung, S. & Epstein, A. Segmentally restricted, cephalic expression of a leucine zipper gene during Drosophila embryogenesis. *Mech Dev* **34**, 3-9. (1991).

132. Kuroha, T. et al. Ablation of Nrf2 function does not increase the erythroid or megakaryocytic cell lineage dysfunction caused by p45 NF-E2 gene disruption. *J Biochem (Tokyo)* **123**, 376-9. (1998).

133. Nagai, T. et al. Regulation of NF-E2 activity in erythroleukemia cell differentiation. *J Biol Chem* **273**, 5358-65. (1998).

134. Nagai, T. et al. Regulation of NF-E2 activity in erythroleukemia cell differentiation The Cap'n'Collar basic leucine zipper transcription factor Nrf2 (NF-E2 p45-related factor 2) controls both constitutive and inducible expression of intestinal detoxification and glutathione biosynthetic enzymes. *J Biol Chem* **273**, 5358-65. (1998).

135. Perkins, N.D. The Rel/NF-kappa B family: friend and foe. *Trends Biochem Sci* **25**, 434-40. (2000).

136. Baldwin, A.S., Jr. Series introduction: the transcription factor NF-kappaB and human disease. *J Clin Invest* **107**, 3-6. (2001).

137. Tak, P.P. & Firestein, G.S. NF-kappaB: a key role in inflammatory diseases. *J Clin Invest* **107**, 7-11. (2001).

138. Lawrence, T., Gilroy, D.W., Colville-Nash, P.R. & Willoughby, D.A. Possible new role for NF-kappaB in the resolution of inflammation. *Nat Med* **7**, 1291-7. (2001).

139. Karin, M., Liu, Z. & Zandi, E. AP-1 function and regulation. *Curr Opin Cell Biol* **9**, 240-6. (1997).

140. Shaulian, E. & Karin, M. AP-1 as a regulator of cell life and death. *Nat Cell Biol* **4**, E131-6. (2002).

141. Herrlich, P. Cross-talk between glucocorticoid receptor and AP-1. *Oncogene* **20**, 2465-75. (2001).

142. Kyriakis, J.M. Activation of the AP-1 transcription factor by inflammatory cytokines of the TNF family. *Gene Expr* **7**, 217-31. (1999).

143. Wisdom, R. AP-1: one switch for many signals. *Exp Cell Res* **253**, 180-5. (1999).

144. Karin, M. The regulation of AP-1 activity by mitogen-activated protein kinases. *J Biol Chem* **270**, 16483-6. (1995).

9. Index of figures and tables

10. Curriculum Vitae

PERSONAL INFORMATION

Name: Weiyue Zheng
Date of birth: May 13th, 1972
Place of birth: Henan, P.R. China
Nationality: People's Republic of China
Marital status: Married

EDUCATION

November 2002-present:	PhD student in Mucosa immunology group, Institute for Clinical Molecular Biology, Hospital of Christian-Albrechts University of Kiel, Germany
August 1999 – July 2002:	Master of biochemistry and molecular biology, Peking Union Medical College (PUMC). Focus on dilated cardiomyopathy
September 1990 – July 1995:	M.D. in Clinical Medicine, Department of Clinical Medicine, Shandong Medical University, Shandong, China. Completed all the required courses and got good scores.

PREFESSIONAL EXPERIENCE AND RESEARCH EXPERIENCE

November 2002 – present:	Institute for Clinical Molecular Biology, University Clinic Schleswig-Holstein, Campus Kiel, Germany (Prof. Dr. Stefan Schreiber): *Evaluation of susceptibility genes for inflammatory bowel disease by association study and candidate gene analyses*
August 1999 – July 2002:	Peking Union Medical College (PUMC), Beijing, China. *Candidate gene associated with family dilated cardiomyopathy with melignant phenotype.*
September 1995 – August 1999:	Resident. China Great Wall Aluminium Corporation Hospital

11. Declaration (Erklärung) and publication list

I acted as a primary scientist in the evaluation of susceptibility genes for inflammatory bowel disease on chromosome 7 and chromosome 12. The chromosome 12 project has been performed in collaboration with Denmark (Prof. Zeynep Tümer). The collaborators contributed with the cDNA experiments on relate region on chromosome 12. No part of it has been submitted to any other board for another qualification. Some of the results have already been published.

Kiel, ……. (Weiyue Zheng)

PUBLICATIONS

1. Weiyue Zheng, Philip Rosenstiel, Klaus Huse et al. Evaluation of AGR2 and AGR3 as candidate genes for inflammatory bowel disease. Genes and Immunity. In press.

2. Carl Manaster, Weiyue Zheng, Markus Teuber et al. InSNP – a visual tool for automation of targeted mutation and InDel detection. Human mutation. 2005 Jul; 26(1): 11-19

3. Carl Manaster, RutaValentonyte, Markus Teuber, Weiyue Zheng et al. SGCaller – Automated assistance for genotypes measured by sequencing. Biotechniques. 2005 Apr; 38(4): 544-546

12. Acknowledgements

I deeply appreciated the expertise and advice I received from Prof. Dr. Stefan Schreiber who has given me the opportunity to work under excellent conditions and with access to large resources of patient material and laboratory, equipment fundamental for the success of this work. His conception of an applied genetic epidemiology and his continuous encouragement and incentive has opened many prospects to me.

I would like to express my sincerest thanks to my doctorfather, Prof. Thomas Bosch, who supported and advised me straightforward. I am very appreciated the critical discussion and advises from him.

My special thanks go to Dr. Jochen Hampe, who created the laboratory platform structure, developed internal database system, statistical analysis software and provided the scientific basis of this thesis. His enthusiasm, new ideas, advices and motivation helped me much doing this work.

I thank the head of the department of General Internal Medicine of the University Clinic Schleswig-Holstein Campus Kiel, Prof. Dr. Ulrich R Fölsch for the opportunity to work in his hospital.

No thesis can be finalized without being read and approved by a thesis committee. I want to thank my thesis committee members for their guidance and assistance.

I would like to thank to Dr. Philip Rosenstiel for the AGR2 promoter functional study experiments. I also appreciate to Dr. Klaus Huse for the AGR2 expression, splice variant definition experiments.

This work would not have been possible without the willingness of people from all over the world to donate blood and genetic material for scientific work, and the medical workers who helped with their clinical expertise and time to collect the material and the medical information relevant for the analyses. I am grateful for the help and commitment I have experienced.

The expert assistance from the laboratory technicians is most appreciated. I want to thank Tanja Henke and Illona Urbach for high quality work of preparing DNA samples for the use in the high throughput format; thank Tanja Wesse, Birthe Petersen and Lena Bossen in the genotyping of many SNPs; thank Tam Wendt, Anita Dietsch, Meike Barche, Hebke Hinz and Melanie Friskovec for their qualified support in the sequencing process; thank Tanja Kaacksteen, Yasmin Brodtmann for the cell culture and luciferase measurement.

I am especially grateful to Dr. Annette Stenzel, who is a great example of scientific work, experiment design and management of different stuff. I am very grateful for her to introduce a general statistical analysis and the first knowledge in the experiment management.

I thank Marcus Will for his never ending work to keep the server-based computer network running and his comments on various software problems. I also thank Carl Manaster, Tim Lu, Rainer Vogler, Marcus Teuber and Michael Wittig, who created our internal databases and analysis software.

I am very grateful to Nancy Mah, who spends much time correcting my English Language in this thesis. I wish to thank Stephan Ott, who always helps me when I met some difficulties.

I would like to express my thanks to Zeynep Tümer, our collaborator from the Wilhelm Johannsen Centre for Functional Genome Research, Department of Medical Biochemistry and Genetics, The Panum Institute, University of Copenhagen, Denmark, who provide us many information about the cDNA cloning on chromosome 12q14 region.

I thank to the colleagues who work in the main office for a pleasant working and social atmosphere during the writing of thesis: Almut Nebel, Ruta Kwiatkowski, Andre Franke, Friederike Flachsbart, Abdou El Sharawy and Janina Heinze. In addition, I am very thankful to Andre who helped me to translate the “Summary’’ chapter into German language.

I would like to thank all the persons who work in the Institute of Clinic Molecular Biology group provide me with a pleasant research atmosphere.

I am very grateful to my parents and my sister who supported me in all my intentions with time, love and trust, and provided me with the foundation to dare and to enjoy life.

My final thanks go to my wife Li Ding, who is more to me and did more for me than words can express.