A systematic genome-wide association analysis for inflammatory bowel diseases (IBD)

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

vorgelegt von

Dipl.-Biol. Andre Franke

Kiel, im September 2006

A systematic genome-wide association analysis for inflammatory bowel diseases (IBD)

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

vorgelegt von

Dipl.-Biol. ANDRE FRANKE

Kiel, im September 2006

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

Koreferent: ..... Prof. Dr. Stefan Schreiber

Tag der mündlichen Prüfung: .....

Zum Druck genehmigt: .....

.....
der Dekan

"After great pain a formal feeling comes."

(Emily Dickinson)

To my wife and family

Table of contents

Abbreviations, units, symbols, and acronyms

°C	degree Celsius
µg	microgram
µl	microliter
µM	micromolar (µmol/l)
aa	amino acid
AbD	Assay-by-Design
AIF	assay information file
AoD	Assay-on-Demand
ASP	affected sibling pair
ATG	autophagy
ATG16L	autophagy-related protein 16
bp	base pairs
BTNL2	butyrophilin-like 2
c	concentration
χ^2	chi-square, measure of association or independence
Caco-2	Human colonic, epithelial-like adenocarcinoma cell line
CARD	caspase recruitment and activation domain
cc	case-control
CCD	charge-coupled device
CD	Crohn's disease
cDNA	complementary DNA
cds	coding sequence
CEPH	Centre d'Etude du Polymorphisme Humain
chr.	chromosome
CI	confidence interval
cM	centiMorgan
CNV	copy number variations
conc.	concentration
CR	call rate
cSNP	coding SNP
D'	D prime or delta prime (measure of LD)

dATP	2'-deoxyadenosine-5'-triphosphate
DCCV	Deutsche Morubs Crohn/Colitis ulcerosa Vereinigung
dCTP	2'-deoxycytidine-5'-triphosphate
ddNTP	dideoxynucleotide triphosphate
DDW	double distilled water
dGTP	2'-deoxyguanosine-5'-triphosphate
DNA	deoxyribonucleic acid
dNTP	2'-deoxynucleoside-5'-triphosphate
dsDNA	double-stranded DNA
dTTP	2'-deoxythymidine-5'-triphosphate
DVD	digitale versatile disc (formerly: digital video disc)
E value	expect value (for BLAST searches)
e.g.	exempli gratia
EDTA	ethylenediaminetetraacetic acid
EGF	epidermal growth factor
EM	expectation maximization
EST	expressed sequence tag
Exo I	exonuclease I
f	frequency
F	forward
FAM	6-carboxyfluorescein
fig.	figure
figs.	figures
FRET	Förster resonance energy transfer
g	gram
g	relative centrifugal force (RCF)
GC	Guanine/Cytosine
GIMPS	Great Internet Mersenne Prime Search
gt	genotype
GW	genome-wide
GWS	genome-wide scan
h	hour
H_0	null hypothesis
HeLa	epithelial-like malignant cells derived from the cervix of Henrietta Lacks
HLA	human leukocyte antigen
HSP	heat shock protein

ht	haplotype tagging
HT-29	a specific cell line
HWE	Hardy-Weinberg equilibrium
i.e.	id est
ibd	identity by descent
IBD	inflammatory bowel diseases
ibs	identity by state
IEC	intestinal epithelial cells
IFN	interferon
IκB	inhibitor of NF-κB
IL	interleukin
kb	kilobase
kD or kDa	kiloDalton
KORA	Cooperative Health Research in the Region Augsburg
l	liter
LD	linkage disequilibrium
LDU	linkage disequilibrium units
LiHa	liquid handling arm
LIMS	laboratory information management system
LOD	logarithm of odds
$\log_{10}$	decadic logarithm
LPS	lipopolysaccharide
LRR	leucine-rich repeat
λ_s	relative risk for siblings
M	molar (mol/l)
MAF	minor allele frequency
Mb	mega base
MDA	multiple displacement assay
mg	milligram
MGB	minor groove binder
$MgCl_2$	magnesium chloride
MHC	major histocompatibility complex
MIM	Mendelian inheritance in man
min	minute
ml	milliliter
mM	millimolar (mmol/l)

mRNA	messenger RNA
MTP	microtiter plate
mut	mutation
NCBI	National Center for Biotechnology Information
NELL	nel-like
NF-κB	nuclear factor κB
ng	nanogram
nM	nanomolar (nmol/l)
NOD	nucleotide oligomerization domain
NPL	non-parametric linkage
nsSNP	non-synonymous SNP
OCTN	organic cation transporter
OMIM	Online Mendelian Inheritance in Man
OR	odds ratio
p	probability
p_{CCA}	allelic p-value
p_{CCG}	genotypic p-value
PCR	polymerase chain reaction
pfSNP	putative functional SNP
pH	potentia hydrogenii
pmol	picomol
PSC	primary sclerosing cholangitis
p_{TDT}	p-value for TDT
PTGER4	prostaglandin E receptor 4
p-value	probability measure in statistical hypothesis testing
R	reverse
r^2	measure of linkage disequilibrium
RFLP	restriction fragment length polymorphism
Rn	fluorescent emission of the normalized reporter dye
RNA	ribonucleic acid
RoMA	robotic manipulation arm
ROX	6-carboxy-X-rhodamine, succinimidyl ester
rpm	rotations per minute
RR	relative risk
RT	room temperature (roughly 21-23°C)
s	second

SAP	shrimp alkaline phosphatase
SD	standard deviation
SHIP	Study of health in Pomerania
SNP	single nucleotide polymorphism
SPSS	statistical package for the social sciences
ssDNA	single-stranded DNA
STR	short tandem repeat
TAE	tris acetate EDTA
TAMRA	6-carboxytetramethylrhodamine, succinimidyl ester
Taq	Thermophilus aquaticus
TBE	tris borate EDTA
TDT	transmission disequilibrium test
TE	tris EDTA
Te-MO	Tecan multipipetting option
TET	5'-Tetrachloro-Fluorescein
TGF-β	tranforming growth factor beta
THP-1	a specific cell line
T_m	melting temperature
TNFα	tumor necrosis factor alpha
Tris	tris-(hydroxymethyl)-aminomethane
UC	ulcerative colitis
UV	ultraviolet
VIC	trade name for fluorscent dye
w/	with
w/o	without
WB	western blot
WGA	whole genome amplification
WT	wildtype
yrs	years

Amino acid symbols

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
X		Stop codon
Y	Tyr	Tyrosine

DNA base nomenclature

A	Adenine
C	Cytosine
G	Guanine
T	Thymine

IUPAC ambiguity code

IUPAC Code	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
N	G or A or T or C	N

List of figures

List of tables

1 Introduction

1.1 Inflammatory bowel diseases, a complex disorder

Numerous mucosal inflammatory diseases which, before the 20th century, were either rare or completely absent in populations of industrialized countries (Bach *et al.*, 2002), have significantly increased over recent decades due to different life-style and environmental factors. Most prominent examples are CD, which was unknown until 1920 (Crohn *et al.*, 1984), atopic eczema (OMIM 603165), and asthma (OMIM 600807). This increase in barrier organ diseases suggests infectious or toxic agents associated with the concomitant changing living conditions. An altered composition of the commensal bacterial flora that is due to hygiene, antibiotics, and different nutrition, is also thought to be a key trigger of an inflammatory barrier disease but the precise mechanisms at work are not clear yet. Known and proposed barrier disease susceptibility factors involve barrier integrity (Palmer *et al.*, 2006), immunoregulatory (Valentonyte *et al.*, 2005), and pathogen defense related genes (Hugot *et al.*, 2001). In contrast to Mendelian or near-Mendelian inherited diseases, complex diseases, such as inflammatory bowel diseases, result from many predisposing variants across the genome. Diseases are said to be complex, if their etiology is based on the complex interplay of several predisposing genetic factors with the environment and the microbial metagenome and they are common, if the prevalence is higher than 0.1% (100 cases per 100,000) in the general population.

Similar histological appearance, cellular activation patterns, and common susceptibility loci identified by genome-wide linkage studies (Becker *et al.*, 1998), all indicate that there are common mechanisms at work in inflammatory disorders, which make them particularly suited to joint analyses.

The most intimate contact with the outside world takes place at the 5,000 square meter surface of the intestines. Considering the often hostile environment, this contact usually seems to take place without any troubles. Inflammatory bowel diseases (IBD) is characterized by perturbed control of inflammation in the gut, and in its interaction with bacteria, and damage to the gut wall (Schreiber *et al.*, 1998). IBD (OMIM 601458) is divided into the the main subphenotypes CD (CD; OMIM 266600) and ulcerative colitis (UC; OMIM 191390) and as many as 1.4 million persons in the United States and 2.2 million persons in Europe suffer from these diseases (Loftus *et al.*, 2004). Although mortality is low, morbidity associated with these diseases is substantial. IBD is also considered a prototype, which will help in the understanding of other chronic inflammatory disorders, e.g. rheumatoid arthritis (RA), where the phenotype is less clear and the disease organ less accessible.

1.1.1 Pathogenesis and pathophysiology

IBD is the result of an abnormal immune response of the gut mucosa triggered by one or more environmental risk factors in people with predisposing gene variations. The median age of onset is 26.5 years for sporadic cases of CD and 22.0 years for familial CD, as determined by Colombel and colleagues (1996). Given this low age of onset and the unpredictable fluctuating symptons, CD imposes a substantial personal burden, with time off work, need for expensive drugs, or surgery and multidisciplinary care. In addition, annual costs for inflammatory bowel diseases drugs are a major economical factor for public health systems, e.g. yearly 7.0 million US dollars in Sweden (Blomqvist *et al.*, 2001).

CD and UC have a combined prevalence of 200–300 per 100,000 in the United States and the life-time risk has been calculated to be 0.15% for CD and 0.3% for UC. With almost twice as many people suffering from UC than from CD, incidence of IBD is about 3–20 new cases per 100,000 each year (Probert *et al.*, 1996; Shivananda *et al.*, 1996) and is rising worldwide (see also fig. 1-1). Furthermore, a north-south gradient across Europe was reported, with the highest prevalence being in Iceland and Stockholm and the lowest in Portugal and Greece (Shivanda *et al.*, 1996). Significant differences in prevalence exist among different ethnic groups living in the same geographic region, displayed by the two- to eightfold higher prevalence in Ashkenazi Jews versus non-Jews (Roth *et al.*, 1989).

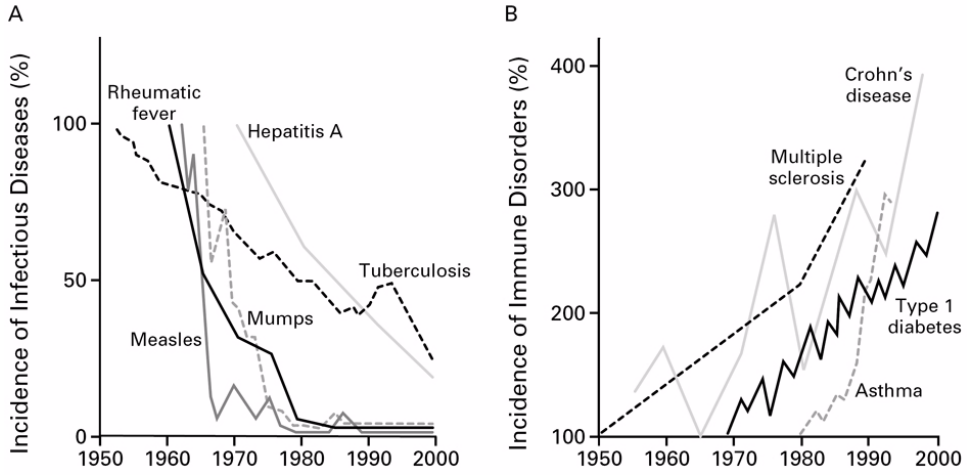

Fig. 1-1 Inverse relation between the Incidence of prototypical infectious diseases (A) and the incidence of immune disorders (B) after World War II. Reprinted from Bach *et al.* (2002).

Clinical features include abdominal pain, chronic diarrhea, rectal bleeding, weight loss, intestinal stenoses, fistulae, growth retardation, fever, and anaemia. A severe course of UC is the formation of a toxic megacolon, a sudden sessation of bowel function leading to toxic dilatation and eventual perforation of the bowel. While some patients develop a chronically active disease, others come to a complete clinical remission between active episodes. IBD is also associated with extraintestinal manifestations, such as arthritis and uveitis (Nordgren *et al.*, 1992). Patients of UC have an increased risk to develop primary sclerosing cholangitis and vice-versa (Cullen *et al.*, 2003), an inflammatory disorder of the bile duct. In addition, IBD is frequently found in well-defined genetic diseases, such as Turner's syndrome (Price *et al.*, 1979; Kohler *et al.*, 1981; Scarpa *et al.*, 1996), Hermansky-Pudlak syndrome (Shanahan *et al.*, 1988; Schinella *et al.*, 1980; Mahadeo *et al.*, 1991), and glycogen storage disease type Ib, which is characterized by neutropenia and abnormal neutrophil function (Yang *et al.*, 1995). IBD, in particular UC, is also associated with several other immune-mediated disorders such as ankylosing spondylitis thyroiditis and multiple sclerosis (Satsangi *et al.*, 1997; Parkes *et al.*, 1997).

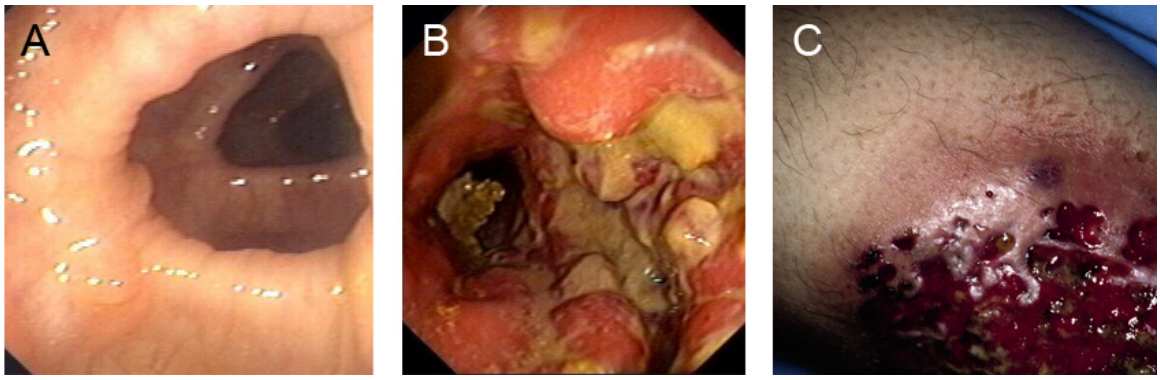

Fig. 1-2 Clinical characteristics of Crohn's disease. (A) Bowel of healthy individual; (B) heavily inflamed bowel with ulcers of a patient; (C) extraintestinal inflammation of the skin of a patient. *By courtesy of Dr. Susanna Nikolaus.*

Some investigators have argued that CD and UC are at opposite ends of a continuous range of disease, but more and more evidence suggests that they are distinct disorders that share some genetic and environmental risk factors and differ in others (Ferguson *et al.*, 1994; Shanahan *et al.*, 2001). In approx. 10% of cases, definitive classification of CD or UC cannot be made and these are designated as "indeterminate colitis". The discrimination between CD and UC is based on clinical, endoscopic, radiological, and histopathological features (Podolsky *et al.*, 1991; Lennard-Jones *et al.*, 1989; Truelove *et al.*, 1976). Recently, gene expression profiling has added further evidence to define CD and UC as distinct molecular entities (Warner *et al.*, 2002). The following table briefly summarizes the important differences between CD and UC:

Crohn's disease	Ulcerative colitis
inflammation is transmural and discontinuous ("skip lesions")	continuous inflammation starting from the rectum that is restricted to superficial layers (mucosa)
any part of the gastrointestinal tract may be involved, most frequently the terminal ileum and colon	inflammation is limited to rectal (proctitis) and colonic mucosal layers
presence of granulomas, stenoses, intestinal/perianal fistulas	no fistulaes, stenoses, or granulomas observed
more often associated with psoriasis and thrombotic vascular complications	more often associated with ankylosing spondylitis, arthritis, primary sclerosing cholangitis, increased risk of colon carcinoma

Table 1-1 Major differences between the two IBD subphenotypes CD and UC. *According to Yang et al. (2001).*

Since its first description in 1913, the cause of CD is still largely unknown. Except for cigarette smoking (Andus *et al.*, 2000), uncertainty still exists around other environmental (risk) factors, such as diet (e.g. fiber, sugar, milk, fast food; see also Persson *et al.*, 1992), stress (for review see Mawdsley *et al.*, 2005), and domestic hygiene (e.g. hot running water). Endogenous modifiers of disease activity – such as the effect of the brain-gut axis and psychological stress – are still under investigation. Breastfeeding is supposed to provide protection against CD and UC according to Klement *et al.* (2004). In contrast to CD, tobacco smoking is protective for UC (Shanahan *et al.*, 2002). In one study, the relative risks of developing UC in heavy ex-smokers, all ex-smokers, non-smokers, and smokers were 4.4, 2.5, 1.0, and 0.6, respectively (Lindberg *et al.*, 1988). Nicotine is probably the main active ingredient in this association (Birrenbach *et al.*, 2004), but the mechanisms remain unknown.

An interesting hypothesis was made by Hugot *et al.* (2003), who described an association of CD with refrigeration. This cold chain hypothesis is based on the observation that psychotrophic bacteria *Yersinia* spp and *Listeria* spp can exist and grow at temperatures between -1°C and 10°C and that cold-chain development paralleled the outbreak of CD during the 20th century. Chronic ingestion of psychotrophic bacteria by genetically pre-disposed individuals might invoke an over-active immune reaction, assuming the loss of tolerance owing to the lack of exposure to microbial antigens during childhood. Several arguments can be made to emphasize the critical role of the commensal microflora. First, the bacterial flora is modified by behavioral changes influencing immune responses, not only locally, but also systemically. Second, animal models of intestinal inflammation demonstrate that germ-free conditions can completely abrogate the development of "spontaneous" IBD. A third argument is the close association between exposure to microbial matter and disease development. The earlier the exposure to microbes, the higher the expression levels of PAMP receptors and the lower the risk of disease development and "overreactions" (Shanahan *et al.*, 2002; Bach *et al.*, 2002). The physical barrier of the intestine consists of a thick glykocalyx

mucous layer coating the apical surface of the epithelial cells. Epithelial cells are laterally joined by tight and adherens junctions. Defects in this protective lining allows the influx of outside components, such as bacteria, which can result in an exagerrated immune response. The bacteria within the enteric lumen have a complex open ecosystem that is continuous with the external environment, and there are up to ten times more bacteria (10^{12} bacteria/g faeces in the colon) than there are cells in the human body. With over 400 bacterial species, identified by 16S rRNA, accounting for over one kilogram of intestinal contents (Farrell *et al.*, 2002; Shanahan *et al.*, 2002) and anaerobes predominating, this dynamic complex has been described as a neglected organ (Bocci *et al.*, 1992). However, progress in such investigations has been hampered by gaps in our knowledge of the normal flora. At least half of the bacterial species cannot be cultured. Reports of qualitative changes have been inconsistent or conflicting (Sartor *et al.*, 1997; Shanahan *et al.*, 2000).

Finally, a possible link exists between de-worming and the emergence of immunological disease (Weinstock *et al.*, 2005; Korzenik *et al.*, 2005). Recent studies suggest that it is possible to downregulate aberrant intestinal inflammation in humans with the porcine whipworm *Trichuris suis* (Summers *et al.*, 2003; Summers *et al.*, 2005). Eggs of this helminth are orally administered to patients. As especially Crohn's disease probably results from failure to downregulate a chronic Th1 intestinal inflammatory process, induction of a Th2 immune response by intestinal helminths diminishes Th1 responsiveness.

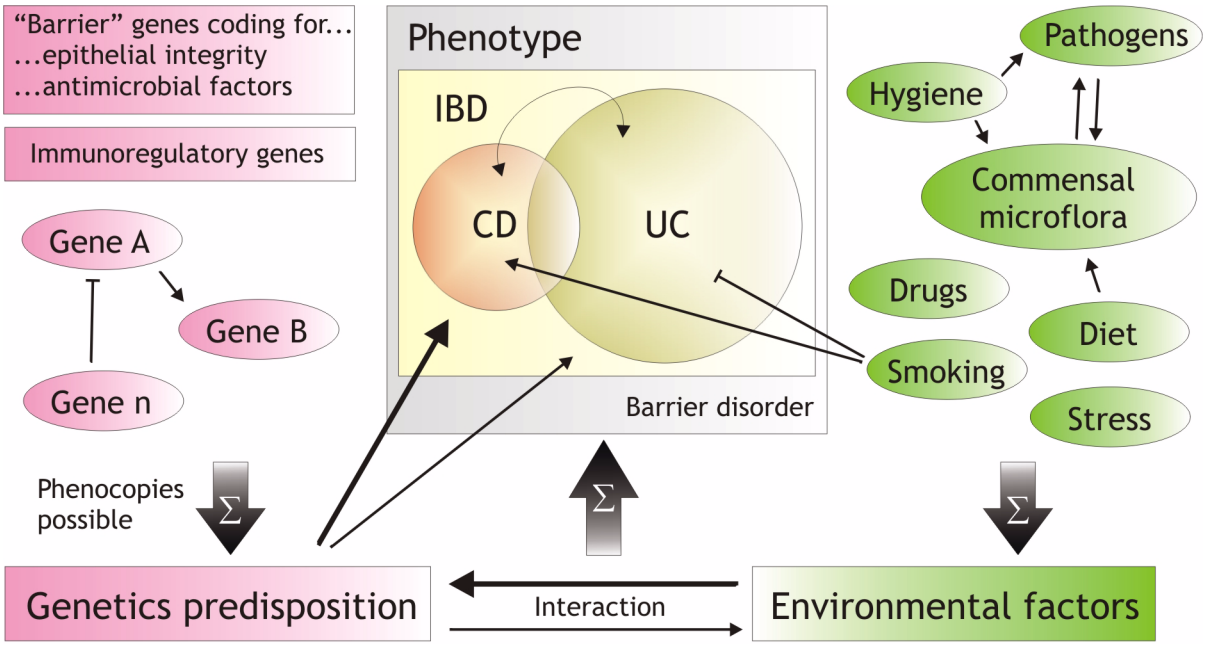

Fig. 1-3 Current knowledge of the etiology of IBD. CD and UC have an overlapping phenotype, and changes in diagnosis are quite frequent. Disease-causing factors can either lead to an increased risk or protect individuals from an outbreak. It is widely accepted that different combinations of susceptibility factors can result in the same disease ("phenocopies").

1.2 Genetics basis of inflammatory bowel diseases

1.2.1 Genetic evidence from family and twin studies

Strong familial clustering has been observed for CD and UC in the late 1980s (Kuster *et al.*, 1989; Mayberry *et al.*, 1989; Colombel *et al.*, 1996; Peeters *et al.*, 1996) and systematic investigations have yielded relative sibling risks of 6–9 for UC (Orholm *et al.*, 1991; Meucci *et al.*, 1992) and 15–50 for CD (Tysk *et al.*, 1988; Satsangi *et al.*, 1994; Schreiber *et al.*, 2005; Fielding *et al.*, 1986).

Besides familial risk data and segregation analysis (Duerr *et al.*, 2002; Orholm *et al.*, 1993), strong genetic support was corroborated by twin studies. Tysk and colleagues (1988) used the Swedish twin registry and inpatient hospital records to identify twins affected by inflammatory bowel diseases. Monozygotic twins with UC had a proband concordance of 6.3%. None of the dizygotic twins with UC were concordant. Monozygotic twins with CD had proband concordance of 58.3% while dizygotic twins had proband concordance of 3.9%. A total of 80 twin pairs were considered (including 34 monozygotic) and despite discordance for presence or absence of inflammatory bowel diseases, no pair was affected by both UC and CD.

Thomson *et al.* (1996) traced 144 twin pairs with inflammatory bowel diseases from 16,000 members of the National Association for Colitis and Crohn's Disease. Six of 38 monozygotic twins with UC and five of 25 with CD were concordant for the disease. The relative risk for an unaffected identical twin to develop inflammatory bowel diseases compared with that for a non-identical twin was 3.49 ($p = 0.03$). Once again, in the majority of cases only one twin had developed overt inflammatory bowel diseases; however, there was no pair of twins with mixed inflammatory bowel diseases.These studies and case reports all indicate a higher concordance for inflammatory bowel diseases in monozygotic than in dizygotic twins, suggesting that genetic factors rather than environmental factors play the primary role in disease pathogenesis. Since Tysk *et al.* calculated the heritability of liability for UC to be 0.53 and that for CD to be 1.0, this suggests a much larger genetic influence in CD. The existing difference in disease prevalence among distinct ethnical groups in some populations and, therefore, the same exposure to environmental factors, also hints at a genetic component in the etiology of IBD (Jayanthi *et al.*, 1992; Probert *et al.*, 1996). The relative contribution of genetic factors to the pathogenesis of CD may be greater than in schizophrenia, asthma, hypertension, or longevity, and at least equivalent to that in insulin-dependent diabetes.

1.2.2 Single nucleotide polymorphisms (SNPs)

The human DNA sequence differs only by approx. 1.2% from the chimpanzee genome (Chen *et al.*, 2001) and there is an 0.1% difference between two individual human genomes. The single nucleotide polymorphisms and other variants contribute mainly to these differences and they have become the toolkit of the geneticists. Polymorphic markers, which are not necessarily involved in a specific trait themselves, are often used as proxies for susceptibility loci that contribute to the disease. The four main types of DNA markers are restriction fragment length polymorphisms (RFLPs), variable number of tandem repeat polymorphisms (VNTRs), microsatellites (short tandem repeats, STRs), and single nucleotide polymorphisms. Only recently, a major role for copy number polymorphisms (CNV) has been suggested.

SNPs are single base-pair differences between genomic sequences and are highly abundant in the human genome. In 2001, the International SNP Map Working Group published a SNP map consisting of 1.42 million SNPs, or approximately one SNP per 1.9 kb (Sachidanandam *et al.*, 2001). In August 2006, the same database (dbSNP build 126) contained a total of 11.96 million SNPs, from which 5.65 million were annotated as validated. However, not all SNPs in public databases are truly polymorphic for a given population. Botstein and Risch (2003) argue that there are probably more than 15 million SNPs with a frequency of at least 1%, which corresponds to almost 1 SNP per 200 base-pairs. SNPs are mostly biallelic and, therefore, less informative than microsatellite markers. However, as they are much more frequent, less prone to undergo mutation, and can also be scored more easily with automatic methods than microsatellites, SNPs are regarded as the state-of-the-art tool for fine mapping in complex diseases.

There is a clear relationship between the severity of amino acid replacement and the likelihood of clinical observation. As compared with a conservative amino acid substitution, a nonsense change is 9.0 times more likely to present clinically (Krawczak *et al.*, 1998). Corresponding ratios for radical, moderately radical, and moderately concervative changes are 3.0, 2.3, and 1.8, respectively. This useful classification of SNPs according to their codon replacement was first described by Grantham *et al.* (1974), and is therefore termed the Grantham scale. Other used terms to classify SNPs are nonsense (generates stop codon), missense, intronic, non-synonymous (amino acid sequence changes), synonymous (no sequence change of gene product), silent, neutral, coding, and non-coding. Missense and nonsense SNP are significantly less abundant than synonymous or intronic SNPs, reflecting evolutionary selection against radical changes (Stephens *et al.*, 2001).

1.2.3 Linkage studies

Once a genetic background is assumed for a disease, the next important step is to reduce the vast number of genes, approximately 28,000 to 35,000 genes (Roest Crollius *et al.*, 2000; Ewing *et al.*, 2000), to a few susceptibility regions in the genome. One method that was developed for already in the 1980s for that purpose is genome-wide linkage analysis.

Alleles close to each other on a chromosome tend to be inherited together, i.e. they are linked. Linkage analysis tests if co-segregation between a phenotype and a genetic marker exists within a pedigree of many generations, or in many independent families. If only one disease locus exists for a given disease, it will be located in a region of the genome that is shared by all affected individuals. The most common approach in genetically complex diseases is the affected sib-pair linkage analysis. Many (hundreds of) pairs of siblings who are affected by a specific disease are genotyped to determine the proportion of parental alleles shared at each marker. If the marker allele sharing between affected siblings is significantly different from the expected ratio (under the assumption of no linkage) of 25%, 50%, and 25% sharing of 2, 1, and 0 parental alleles respectively, the region surrounding the markers is linked to the disease. The resolution relies on detecting crossovers between markers and the disease locus within families and there will be only few such events close to the locus among affected sib-pairs; hence the candidate interval will be large. On the other hand, it is possible to pick up the effect of a disease locus at longer distances and, therefore, the spacing between markers in a linkage scan can be relatively wide (typically 5–10 cM with 1 cM = 1% recombination rate). This is what made whole genome linkage scans possible. A whole genome linkage scan usually includes typing of about 300–800 evenly-distributed microsatellites. These short tandem repeats are di-, tri-, and tetranucleotide repeats (e.g. $[CA]_n$ or $[CAG]_n$) that are abundant in the human genome (approx. 10,000). The LOD (logarithm of odds) score, which is the favoured measure for linkage analysis, is the decadic logarithm of the probability that two markers are linked with a given recombination value divided by the probability that they are unlinked (see also Lander and Kruglyak criteria, 1995).

Numerous independent linkage scans have been performed for inflammatory bowel diseases by different groups, giving rise to several identified susceptibility loci, with some replicating more consistently than others. *IBD1* on chromosome 16 is the most widely and consistently replicated linkage region among the detected regions. A summary of the most important linkage regions is shown in table 1-2, page 9 and a detailed review of the the IBD linkage literature can be found at the end of this thesis (9.1 on page 173) or in Zheng *et al.* (2003). Although there are striking discrepancies among the genome-wide scans in respect

of linkage loci and examined populations, almost all studies detected more than three linkage loci, indicating that multiple genes are involved in the pathogenesis of IBD.

Name	Locus	Known disease genes
IBD1	16p12–q13	*CARD15*
IBD2	12p13.2–q24.1	–
IBD3	6p	MHC region
IBD4	14q11–12	–
IBD5	5q31–33	*SLC22A4/SLC22A5*
IBD6	19p13	–
IBD7	1p32–36	–
IBD8	16p12	–
IBD9	3p21–26	–

Table 1-2 IBD linkage regions.

Because Hugot *et al.* (1996) and Satsangi *et al.* (1996) detected strong evidences for linkage on chromosomes 16, 12, 6, and 3, subsequent studies mainly focused on these candidate regions. Thus, later identified, though being attractive loci as well, such as 14q, 19p, and 1p, were less investigated. Collectively, genome scans have shown that CD and UC are distinct disorders, which is most apparent in CD-specificity of the 16q12 (*IBD1*) locus. An additional complexity is the variable role that different susceptibility loci might have in different ethnogeographic populations, reflecting true genetic heterogeneity. More than three studies (Rioux *et al.*, 2000; Brant *et al.*, 2000; Akolkar *et al.*, 2001) reported significant differences between Jewish and non-Jewish patients. Mwantembe *et al.* (2001) noted also that IBD is more prevalent among South African whites than among blacks, a pattern observed elsewhere as well. They concluded that the inflammatory process leading to IBD may be distinct in different population groups. Besides genetic heterogeneity, the discrepancies between the results of the various screens and follow-up linkage studies may be related to the difficulties in detection of genes of modest effects in complex traits (Risch *et al.*, 1996). Positive linkage results are typically observed over very broad regions for genes of major effect (Cho *et al.*, 2000). Mapping resolution depends on both sample size and marker density, and even the largest studies typically have a rather poor resolution of 5–10 cM, corresponding to approx. 5–10 Mb (Clark *et al.*, 2003), although a high-resolution linkage study was reported just recently (McKnight *et al.*, 2006). Once the region of interest has been narrowed down to a sufficiently small region, fine mapping is usually carried out in cases and controls by association studies.

1.2.4 Association studies

Classical linkage analysis and positional cloning clearly remain the methods of choice for identifying rare, high-risk, disease-associated mutations, owing to the clear inheritance patterns they display. Although more than 1,200 genes (Botstein *et al.*, 2003) have been characterized for highly penetrating Mendelian phenotypes, only a small number of human genes that contribute to complex diseases were identified by this classical "positional cloning" approach. Greater statistical power and finer resolution is achieved by SNP association studies in a case-control setting. By using unrelated individuals, the information of numerous meiotic events throughout the multigenerational pedigree is indirectly extracted, thus leading to significant narrowing of the region of interest (Nordborg *et al.*, 2002; Botstein *et al.*, 2003). Owing to this physical narrowing and the lower level of informativity of biallelic SNPs, hundreds of thousand SNPs are necessarys to carry out a genome-wide scan of sufficient coverage compared to max. 800 STRs for a genome-wide linkage scan. Precise SNP numbers are still under debate and vary from 100,000 to 1,000,000 SNPs (Botstein *et al.*, 2003). Discussions will probably end with the availability of the 1000$ genome (Bennett *et al.*, 2005) that extracts genetic variation *in toto*. As millions of validated SNPs (see 1.2.2 on page 7) are annotated in public databases, merely methodological problems limit coverage, albeit throughput continuously increases (see fig. 1-4). Furthermore, a standard genome-wide scan for 400 cases and 400 controls still costs more than half a million Euro, posing a major obstacle for many research groups.

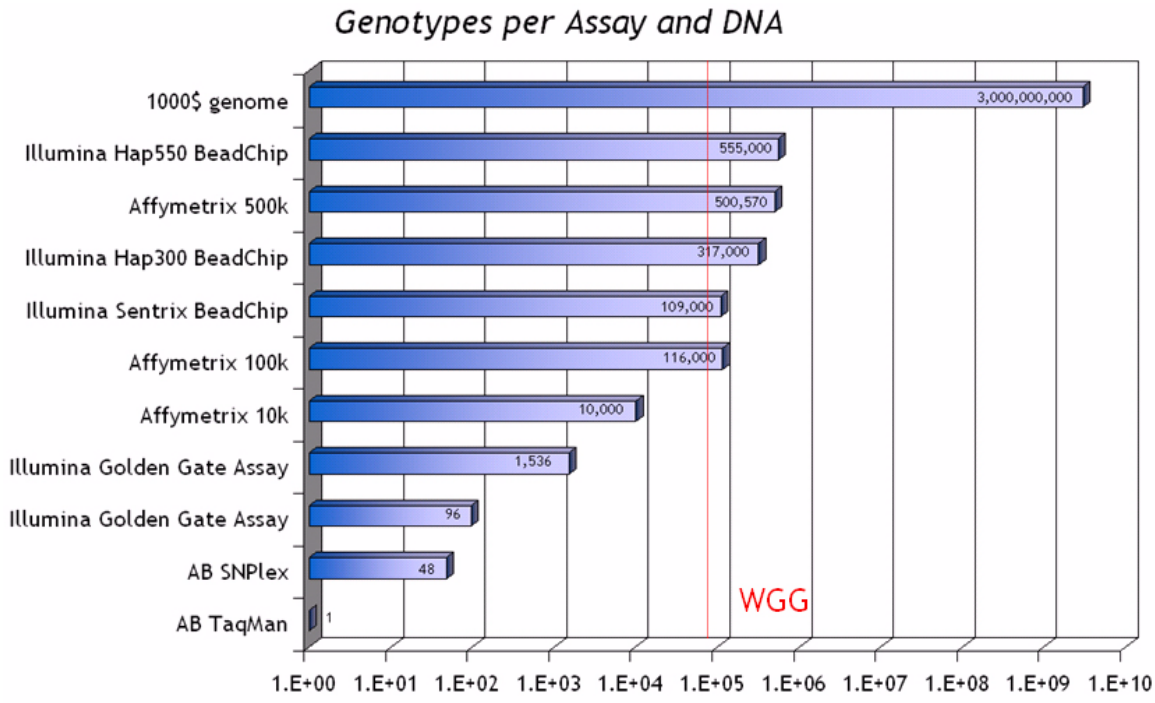

Fig. 1-4 SNP genotyping method market overview. Whole genome SNP genotyping (WGG) methods exist since 2003 and the throughput is increasing with time (y-axis). Genotypes per assay and DNA are shown on a logarithmical scale along the x-axis.

An association is said to exist between genotype and phenotype when they occur together more often than expected by chance. To investigate whether phenotypes are associated or not, one usually collects two groups of individuals, e.g. patients and a control sample for case-control design. In each group the proportion of interest is then determined. The usual chi-square test for 2 x 2 tables is subsequently applied to determine statistical significance. The strength of an association is often measured by the relative incidence or relative risk (RR) or the correlation coefficient. However, potential effects that are unrelated to disease, such as ethnic, social, or geographical stratification, could generate systematic differences between patients and controls. This problem of population stratification is generally overcome by using a family-based association design, e.g. the transmission disequlibrium test (TDT; Spielmann *et al.*, 1993). Unfortunately, this design is expensive and computationally inefficient (Morton *et al.*, 1998).

Botstein and Risch (2003) distinguish between two different types of genome-wide association studies, namely a map-based (LD-based, indirect) and sequence-based (functional SNPs, direct) approach, both being unbiased in terms of involved genes. A third and commonly employed association study is the so-called candidate gene approach. To this end, genes are selected for genotyping based on their possible role in disease-associated pathways, animal models, or other characteristics.

Until recently, only few genome-wide association studies have been reported. A significant number of large multi-center association studies are ongoing and a flood of related publications is expected within the next years. Early successes of genome-wide association studies, showing the feasibility and power of this approach, are listed in the following table:

Reference	Disease	No. of SNPs	Identified gene(s)
Martins Silva *et al.*, 2003	Multiple sclerosis	3,974	MHC region
Smyth *et al.*, 2006	Type I diabetes	6,500	*IFIH1*
Hu *et al.*, 2005	Esophageal cancer	11,555	*GASC1, EPHB1, PIK3C3*
Namkung *et al.*, 2005	Alcoholism	15,878	*GABRA1*
Nakamura *et al.*, 2005	Crohn's disease	72,738	*TNFSF15*
Ozaki *et al.*, 2005	Myocardial infarction	92,788	*LTA*
Klein *et al.*, 2005	Age-related macular degeneration	116,204	*CFH*
Herbert *et al.*, 2006	Obesity	116,204	*INSIG2*
Maraganore *et al.*, 2005	Parkinson disease	198,345	*SEMA5A, PARK10+11* locus

Table 1-3 Early successes in disease-finding utilizing genome-wide association studies. Ordered by number of genotyped SNPs.

The currently tremendous interest in using diseases association mapping in humans has "slopped over" to other organisms as well, e.g. mice (Liu *et al.*, 2006).

1.2.5 Known susceptibility genes

Early segreation analyses suggested a major recessive gene for CD and a major dominant gene for UC. However, CD and ulcerative are now considered polygenic traits. Support comes from an increasing amount of published susceptiblity genes and several replicated linkage loci. In IBD, disease genes involved in the regulation of the innate immune system, mucosal integrity, and cell-cell interactions are all clearly plausible candidate genes. The most important IBD genes and regions are briefly described in the next sections. For review see also Noble *et al.* (2006) and Yamada *et al.* (2005).

1.2.5.1 CARD15

The first susceptibility locus for CD, which has received greatest support in replicative studies worldwide, is on the pericentromeric region of chromosome 16, and designated *IBD1* locus. Mutations of the *CARD15* (caspase activating recruitment domain, 15) gene in this region have been conclusively associated with CD (Hugot *et al.*, 2001; Ogura *et al.*, 2001; Hampe *et al.*, 2001). NOD2 (nucleotide oligomerization domain), the protein product of *CARD15*, functions as an intracellular sensor of muramyl dipeptide (MDP), a highly conserved peptidoglycan (PGN) motif, common to many intraluminal bacteria. Therefore, NOD2 is part of the pattern-associated molecular pathogen (PAMP) recognition system (Hugot *et al.*, 2001) and it is further a member of the CATERPILLER (CARD, transcription enhancer, R[purine]-binding, pyrin, lots of leucine repeats) protein family. Thus, *CARD15* is part of the evolutionary ancient innate immune system. Besides intestinal epithelial cells, it is expressed in monocytes and activates nuclear factor κB (NF-κB), which is a key transcriptional factor involved in initiation of immunoinflammatory responses (van Heel *et al.*, 2001; Mahida *et al.*, 2001).

Besides a frameshift mutation 1007insC ("SNP 13") that leads to a truncation of the transcript in the leucine-rich repeat region, Hugot and others (2001) identified the two missense mutations Arg702Trp ("SNP 8") and Gly908Arg ("SNP 12"). Of the patients with CD, 10–30% were heterozygous and 3–15% homozygous or compound heterozygous for the *CARD15* mutations; the corresponding proportions from the control population were 8–15% and 0–1% respectively. Hugot and colleagues (2001) determined the relative risk of CD for heterozygous, homozygous, or compound heterozygous individuals for the mutations to be 3-fold, 38-fold, and 44-fold higher than for normal controls, respectively. Later, Croucher *et al.* (2003) confirmed that the disease-associated SNPs occur independently and share a common back-

ground haplotype. This suggests a common origin and the possibility of a yet undiscovered and more strongly predisposing muation.

"SNP"	dbSNP ID	nt change	aa change	Exon	f_{Co}	f_{Ca}	OR
8	rs2066844	C14772T	R702W	4	4.7%	10.5%	2.33 (95%CI: 1.81 – 3.00)
12	rs2066845	G25386C	G908R	8	1.3%	4.0%	3.12 (95%CI: 2.03 – 4.78)
13	rs2066847	32629insC	1007insC	11	3.9%	15.5%	4.17 (95%CI: 3.22 – 5.39)

Table 1-4 Summary of the three major mutations in the *CARD15* coding sequence. Minor allele frequencies for cases and controls were calculated based on 1091 healthy German controls and 1150 German CD patients. Odds ratios and 95% confidence intervals were calculated for carriership of the risk allele, i.e. 11 + 12 vs. 22 (*unpublished data*).

In 2002, Lesage *et al.* sequenced all *CARD15* exons plus flanking splice sites in 453 CD patients, 159 UC patients, and 103 healthy controls (1,430 chromosomes in total). They identified 67 SNPs, of which 31 were potential disease-causing mutations (DCM). SNP 8, 12, and 13 represented 32%, 18%, and 31% respectively, whereas the total of the rare mutations represented 19% of DCMs, thus they were categorized as "private mutations". Interestingly, 93% of all mutations were located in the distal third of the gene, once more highlighting the crucial role of the leucine-rich repeat domain (LRR). *CARD15* also plays a role in the susceptibility to Blau syndrome (Miceli-Richard *et al.*, 2001), Psoriasis (Lee *et al.*, 1990; Rahman *et al.*, 2003), early-onset sarcoidosis (Kanazawa *et al.*, 2005), although sometimes different sequence variants are relevant than in CD (see fig. 1-5). Sarcoidosis is a granulomatous relapsing inflammatory disease that mainly affects the lung, while Blau syndrome is a rare disease typically defined by granulomatous arthritis, skin eruption, and uveitis occurring in the absence of lung or other visceral involvement.

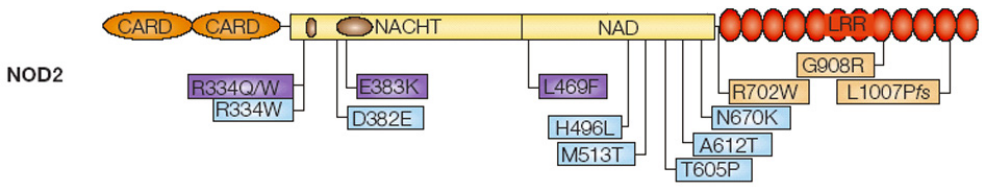

Fig. 1-5 Different mutations in *CARD15* contribute to distinct inflammatory disorders. Although CD-associated variants are mainly found within the LRR domain, mutations linked to other diseases are predominantly situated in the nucleotide-binding domain found in NACHT (NAIP, CIITA, HET-E and TP1), and the NAD (NACHT-associated domain). The CARD domains have been implicated in apoptosis and NF-kB activation, the NBD (NACHT + NAD) domain in oligomerization, and the LRR in bacterial recognition. Purple: Blau syndrome; light blue: early-onset sarcoidosis; yellow: CD. Brown ovals depict the binding site of the magnesium-nucleotide complex. *From Schreiber et al., 2005.*

CARD15 variants seem to account for less than 20% of CD and they have not been found to be associated with collagenous colitis, another rare subphenotype of IBD (Madisch *et al.*, 2006). The CD-associated mutations arose about 40,000 years ago (Schreiber *et al.*, 2005), suggesting that positive selective evolutionary pressure (Akey *et al.*, 2002) is missing, highlighting the role of just recently changed living conditions due to "Westernization".

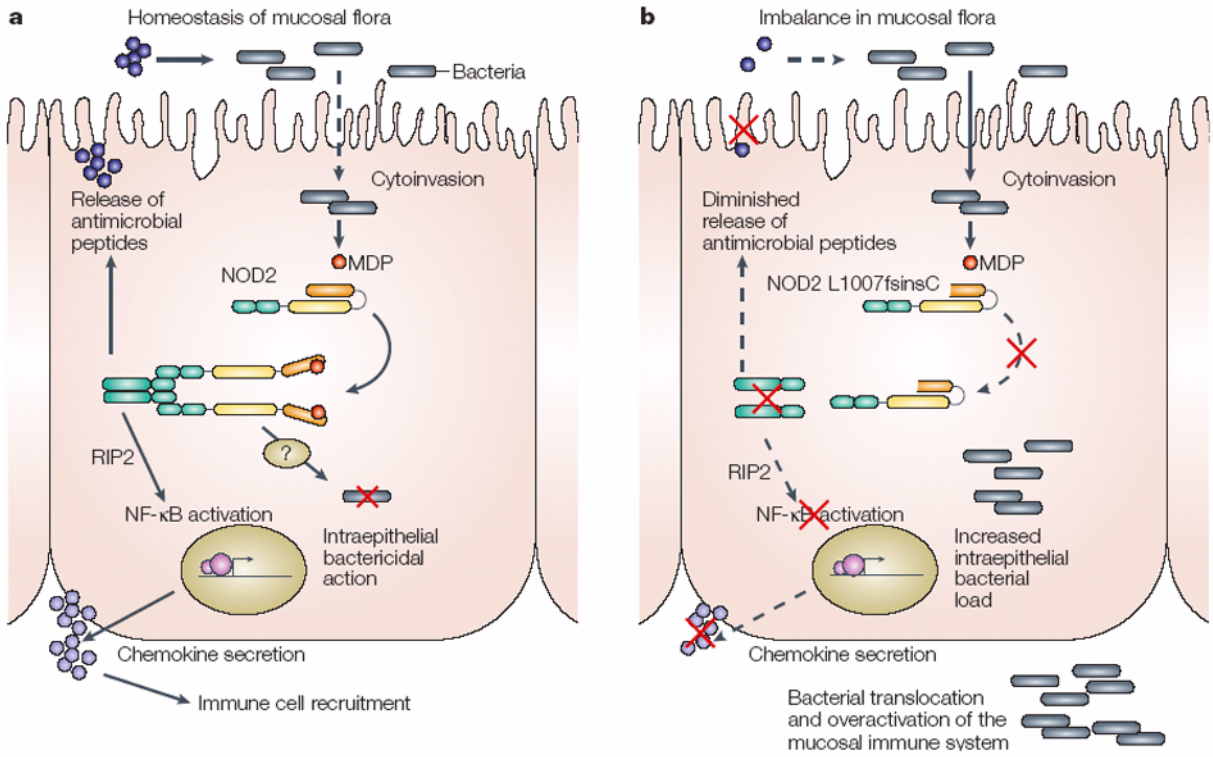

Fig. 1-6 Effects of *CARD15* mutations on the NOD2 protein and the intestinal epithelial barrier. (A) MDP is recognized intracellularly by the LRR of NOD2. This leads to self-oligomerization and RIP2-mediated NF-κB activation. Regulated chemokine secretion and defensin release contribute to mucosal homeostasis. (B) A defect in the MDP-sensing LRR, as caused by the CD-associated variants, leads to loss-of-barrier function, as expression of protective chemokines and defensins is reduced. *Illustration from Schreiber et al., 2005.*

As Yamazaki *et al.* (2002) showed that *CARD15* variants do not play a significant role in the pathogenesis of Japanese CD patients (see also Inoue *et al.*, 2002) and it is further known that mutations are absent or very rare in Chinese (Leong *et al.*, 2003) and Koreans (Lee *et al.*, 2005), once more an ethnic difference in genetic susceptibility to CD is suggested. Low frequencies of *CARD15* risk alleles have also been observed in the New Zealand Maori (Gearry *et al.*, 2006), a population in which IBD is uncommon (Wigley *et al.*, 1962; Schlup *et al.*, 1986) and in African colored patients (Zaahl *et al.*, 2005; Kugathasan *et al.*, 2005). The contribution of *CARD15* variants to disease susceptibility is also much lower in northern Europe (Finland, Ireland, Scotland, Sweden, Iceland, and Norway) than elsewhere in Europe (Arnott *et al.*, 2004; Idestrom *et al.*, 2005; Medici *et al.*, 2006). A meta-analysis of 42 case-control studies examined the ORs for development of CD for carriers of one mutant allele (Economou *et al.*, 2004) of SNP 8, 12, and 13 and risk was increased by a factor of 4, 3, or 2, respectively.

In summary, the discovery of NOD2/*CARD15* has dramatically changed the focus of IBD research to the role of the innate immune system (Schreiber *et al.*, 2005). Was the disease pathophysiology initially considered a merely T cell-driven process, it is now mainly viewed as a barrier disorder.

1.2.5.2 CARD4

The protein NOD1, a structural homologue of NOD2, is encoded by the *CARD4* gene on chromosome 7p14. Because of its homology to NOD2, its expression in epithelial cells, its ability to recognize *Shigella flexneri* LPS (Girardin *et al.*, 2001) and to activate NF-κB, plus a known linkage peak on 7p14–p15 (Satsangi *et al.*, 1996), association of *CARD4* with IBD was exhaustively tested by several groups. Zouali *et al.* (2003) found five non-conservative changes in the *CARD4* coding sequence, of which E266K (MAF = 0.28) was the only non-private mutation. This nsSNP was not associated with UC, CD, or IBD in their sample, though. In 2005, Hysi and others found an insertion-deletion polymorphism (ND_1+32656) near the beginning of intron 9 that was associated with the presence of asthma and elevated immunoglobulin E levels. They further showed that the variant influences the binding of an unknown nuclear factor. In contrast to asthma, the more common allele of ND_1+32656 was found to be associated with IBD in a British panel (McGovern *et al.*, 2005). ND_1+32656 was further significantly associated with early age of onset CD and IBD. McGovern *et al.* (2005) also observed the presence of a strong protective two-marker haplotype (ND_1+32656*2/rs2907748*1).

1.2.5.3 *TNF-α*

The pro-inflammatory cytokine TNF-α plays a crucial role in mucosal inflammation and is likely to be at the apex of the inflammatory cascade in CD (Murch *et al.*, 1993; Satsangi *et al.*, 1998), as TNF-α can induce apoptosis and activate NF-kB through signaling cascades emanating from TNFR1. The TNF locus, which maps to the *IBD3* locus on chromosome 6 within the class III region of MHC, between HLA-B and HLA-DR, is a good site to look for genotype-phenotype relationships (Sashio *et al.*, 2002). Many polymorphisms described in the *TNF* gene have already been studied in IBD (Fowler *et al.*, 2005; Bouma *et al.*, 1996 and 1998; Louis *et al.*, 2000). TNF levels are elevated in the serum, mucosa, and stool of IBD patients, and infusion of monoclonal anti-TNF antibody is a highly efficacious IBD therapy (Komatsu *et al.*, 2001; Braegger *et al.*, 1992; Breese *et al.*; 1994; D'Haens *et al.*, 1999). In 2002, van Heel *et al.* identified a polymorphism in the *TNF-α* promoter as a marker of IBD susceptibility in the UK population. However, it remains unclear whether the TNF_{-857} variant is a true disease allele or a marker allele in LD with a neighbouring functional polymorphism. Interestingly enough, TNF-α has been found to upregulate NOD2 in epithelial cell lines (Rosenstiel *et al.*, 2003).

1.2.5.4 5q31 haplotype

In 2001, John Rioux and colleagues described a systematic approach for LD mapping in the *IBD5* linkage region (Rioux *et al.*, 2000; Ma *et al.*, 1999) for CD on chromosome 5q31. This led to the finding of a strongly disease-associatied haplotype, spanning 250 kb and comprising 11 discrete haplotype blocks (Daly *et al.*, 2001). However, analogous to the HLA region, tight LD across the region hindered them to identify a single causative mutation. Instead, Rioux *et al.* (2001) identified 11 significant SNPs that carry equivalent genetic information and are strongly associated with CD, a finding that was later replicated in several populations (Giallourakis *et al.*, 2003; Mirza *et al.*, 2003; Negoro *et al.*, 2003). Three years later, Peltekova *et al.* (2004) detected 10 new SNPs by resequencing the *IBD5* interval, of which two were found to have functional effects. The first is a C→T missense substitution in exon nine of solute carrier family 22 member four (*SLC22A4*) that causes the amino acid substitution L503F at an evolutionarily conserved position in the 11th transmembrane domain of OCTN1. The second SNP is a G→C transversion in the *SLC22A5* promoter (-207G→C), which disrupts a functional promoter element. *SLC22A4* and *SLC22A5* are within a single haplotype block (block 7; see fig 1-7, page 17) and encode for the organic cation transporter proteins OCTN1 and OCTN2, respectively. The protein OCTN1 has a length of 551 amino acids, is strongly expressed in kidney, trachea, bone marrow, and to a lesser extent in the small bowel, and has been characterized as a carnitine transporter (Tokuhiro *et al.*, 2003). OCTN2, that is 557 aa long, is 75.8% homologous to OCTN1, and functional studies have shown it to be a high affinity sodium carnitine transporter, expressed in kidney, smooth muscle, and heart tissue (Tamai *et al.*, 1998).

In contrast to most of the studied populations, where the two-allele risk haplotype “TC” is prevalent (e.g. 54% in affected vs. 42% in unaffected; Peltekova *et al.*, 2004), the previously identified variants of *IBD5* have been found to be extremely rare in the Japanese population (Negoro *et al.*, 2003; Yamazaki *et al.*, 2004). An intronic SNP rs2268277 of *SLC22A4* has been reported to be associated with another chronic inflammatory disorder, namely rheumatoid arthritis (Tokuhiro *et al.*, 2003).

The observation that the frequencies of L503F and -207G→C in individuals who do not carry the general *IBD5* risk haplotype are not significantly different in cases and controls (Torok *et al.*, 2005; Fisher *et al.*, in press), led to an ongoing debate whether the real disease-causing mutation(s) remain(s) to be identified. Though Peltekova *et al.* (2004) showed an altered function or expression of the OCTN1 and OCTN2 cation transporterts, it is generally agreed that the link between aberrant OCTN function and CD remains to be established.

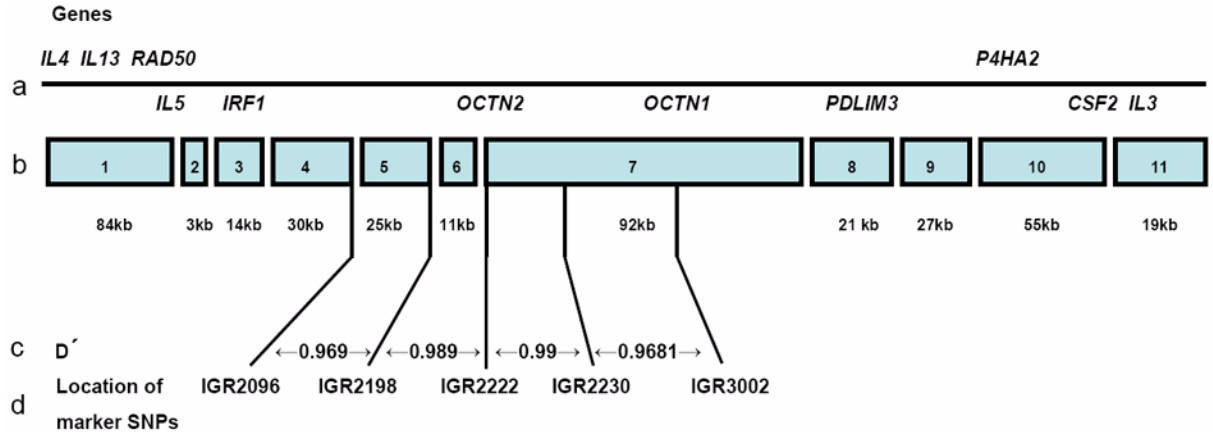

Fig. 1-7 *IBD5* locus (5q31) with high-resolution haplotype structure. (A) Candidate genes above the relevant haplotype blocks. Genes above the line are transcribed from left to right and those below the line are transcribed from right to left. (B) 11 blocks numbered 1 to 11 (between 3 kb and 92 kb) each of limited genetic diversity are punctuated by sites of recombination. (C) D' scores are shown to demonstrate the tight linkage disequilibrium between the SNPs that were analyzed. The lowest D' (D' = 0.959) was observed between IGR2096 and the OCTN2 variant (-207G→C). (D) Location of the SNPs that were analysed, IGR2222 representing the OCTN2 variant (-207G→C) and IGR3002 representing the OCTN1 variant (1672C→T). IGR2078, which was used by Peltekova and colleagues (2004) to represent the extended *IBD5* haplotype, is located in block 4. *Illustration from Noble et al. (2005).*

1.2.5.5 DLG5

Refining the linkage peak of Hampe *et al.* (1999) on chromosome 10q23, led to the association finding between the discs large homolog 5 (*DLG5*; 79,220,557–79,356,354 bp) and IBD (Stoll *et al.*, 2004). Since then, several investigations have led to controversial findings and ongoing discussions concerning the validity of the initial finding.

SNP 113G→A (rs1248696), which results in the amino acid substitution R30Q in the DUF622 domain, is the only haplotype tagging SNP of the significantly overtransmitted haplotype D in the *DLG5* gene. R30Q is completely absent, i.e. monomorphic, in Greek (Gazouli *et al.*, 2005) and Japanese (Yamazaki *et al.*, 2004) patients. Failure of replication of the positive association of R30Q and of the negative association with haplotype A with IBD and CD was reported by Torok *et al.* (2005), Noble *et al.* (2006), and Vermeire *et al.* (2005) for other European populations. Although not seeing the negative haplotypic association with IBD, Daly *et al.* (2005) replicated the positive association of the nsSNP R30Q in a Candian and Italian patient panel, but not in a panel from the United Kingdom. In 2006, Friedrichs *et al.* determined *DLG5* as a male-specific CD susceptibility locus by using multivariate logistic regression analyses on the previous datasets of Stoll *et al.* (2004) and Daly *et al.* (2005). This finding is consistent with the observation of Fisher *et al.* (2002), who identified the linkage region 10q as a male-specific locus. Similar to the observation made by Tenesa *et al.* (2006), the observed gender-specificity is a result of divergent allele frequencies in healthy controls and not in the CD patients. Friedrichs *et al.* (2006) argument that this transmission ratio distortion of the Q allele among controls is likely to be the consequence of prenatal processes. The question remains why the observed gender-dependent disparity in Q allele frequencies is offset in male

CD patients. Significant differences were also found in transmissions of R30Q between the *CARD15*^{risk} and *CARD15*^{nonrisk} group with no association seen in the latter and an increased association in the first (Stoll *et al.*, 2004; Vermeire *et al.*, 2005).

The DLG5 protein belongs to the group of membrane-associated guanylate kinases (MAGUKs) and as such contains 4 PDZ domains, one SH3 domain followed by one guanylate kinase (GUK) domain. Additionally, DLG5 contains an N-terminal domain DUF622 of unknown function. All domains are assumed to be involved in protein-protein interactions, supporting the notion that DLG5 serves as a multi-functional adapter and scaffold protein which is involved in the maintenance of epithelial integrity (Friedrichs *et al.*, 2006). In addition, DLG5 has been reported to be involved in maintaining cell shape and polarity (Humbert *et al.*, 2003), and to be located at cell-cell contact sites (Wakabayashi *et al.*, 2003). The R30Q variant is thought to disturb binding to Rab GTPase and thus is likely to have functional implications (Stoll *et al.*, 2004).

1.2.5.6 TNFSF15

TNFS15 was identified as a susceptibility gene for CD by Yamazaki *et al.* in 2005. They conducted a multi-stage genome-wide association study in 94 unrelated Japanese patients and 752 controls using 72,738 gene-centric SNPs. A highly significant association ($p = 1.71 \times 10^{-14}$) of SNPs and haplotypes within the *TNFS15* (tumor necrosis factor superfamily, member 15) gene was replicated in the second stage using an additional 390 CD samples. The most significant SNP (tnfsf15_28) resides in intron 3 of the gene. In the third stage, the association was confirmed in a large family-based and case-control IBD panel from the UK, with p-values ranging between 0.01 and 0.05.

TNFSF15 is located on 9q32 (114,631,166–114,647,962 bp), a linkage locus that was first connected with CD by Cho and colleagues (1998). The protein is a novel TNF-like factor, expressed primarily in endothelial cells (Yue *et al.*, 1999). Upregulation of transcript and protein levels of TNFSF15 have recently been reported for macrophages and $CD4^+/CD8^+$ lymphocytes of the intestinal lamina propria of CD patients (Bamias *et al.*, 2003).

1.2.5.7 HLA/MHC on chromosome 6

Evidence suggests that *CARD15* is only associated with CD (Hugot *et al.*, 2004), however, a stronger association exists between genes of the human leukocyte antigen (HLA, involved in regulating the immune response) region and UC than for CD (Satsangi *et al.*, 1996; van Heel *et al.*, 2004; Yap *et al.*, 2004). A review by Zheng *et al.* (2003), who summarized 18 studies published since 1995, describes DRB1 and DQB1 as the key regions for UC and CD.

Regarding the class II genes, the most consistent positive association in UC is with DRB1*1502 across multiple ethnicities (Ahmad *et al.*, 2003; Yoshitake *et al.*, 1999; Trachtenberg *et al.*, 2000) and the susceptibility to extensive disease conferred by the DRB1*0103–DQB1*0501 haplotype (Ahmad *et al.*, 2003, Roussomoustakaki *et al.*, 1997). However, both these haplotypes are rare (UC frequencies ~2% and ~8%, respectively), implicating importance in subgroups of UC patients only (Ahmad *et al.*, 2003). The most consistent negative class II association in UC is with the DRB1*0401–DQB1*0301 haplotype of approximately the same magnitude (OR ~0,5) as the DRB1*0401 association in PSC (Stokkers *et al.*, 1999). To what extent these class II associations in UC are biologically important in themselves, or only serve as markers for other candidate polymorphisms elsewhere in the HLA-complex is currently not known. For the individual class III and class I genes (of which TNF-α, LTA, HSP-70, MICA and HLA-B are the most studied), there is less consistency. In sum, TNF-α variants (especially at promoter positions -308 and -857) may confer susceptibility to UC (van Heel *et al.*, 2002; Yamamoto-Furusho *et al.*, 2004; O'Callaghan *et al.*, 2003), LTA associations seem to be in LD with TNF-signals (Coss *et al.*, 2000), particular MICA and HLA-B alleles can only be demonstrated in Japanese patients (Ahmad *et al.*, 2002; Nomura *et al.*, 2004) and the role of HSP-70 polymorphisms is still unclear (Ahmad *et al.*, 2002). Thus, the primary associations may be located outside and in LD with loci studied so far, and systematic mapping of the entire HLA-complex in UC is ongoing (Stenzel *et al.*, 2004).

Within the MHC region, the identification of a causative mutation is complicated through the high density of genes and polymorphisms and the extent of LD (Shiina *et al.*, 2004). Studies are further limited due to the great number of possible haplotypes as well as the variation of haplotypes between different populations (Kawasaki *et al.*, 2000). The majority of the reported associations for UC and CD differ (Ahmad *et al.*, 2002; Ahmad *et al.*, 2003), and possibly the region contains more than one IBD susceptibility locus (Yap *et al.*, 2004; Hampe *et al.*, 1999).

1.2.5.8 Other proposed IBD susceptibility genes

Several other susceptibility genes have been previously described, but in contrast to the *CARD15*, 5q31, or HLA locus, none of the preliminary findings was consistently replicated by other groups. A diverse range of reasons can be claimed for this discrepancy among association studies, for example false-positive findings, missing power, population stratification, or genetic heterogeneity. In the light of discordant data-sets, it is difficult to ascertain whether genes impact clinical outcome or not.

Toll-like receptors (TLRs) are transmembrane glycoproteins which recognize conserved products unique to microbial metabolism and signal via a number of downstream molecules, e.g. MyD88, IL-1R-associated kinases, TGF-β, and TNF-receptor associated factor 6 (Takeda *et al.*, 2003). Eleven members of the TLR family have been identified to date (Akira *et al.*, 2004) and two mutations in the TLR4 gene, Asp299Gly and Thr399Ile, have been associated with CD and/or UC (Franchimont *et al.*, 2004; Gazouli *et al.*, 2005; Torok *et al.*, 2004; Brand *et al.*, 2005). Although TLR4 antagonists prevent mice from developing colitis (Fort *et al.*, 2005), no associations with IBD and the two SNPs were seen by other groups (Arnott *et al.*, 2004). Identification of a protective polymorphism in the gene coding for flagellin-sensing TLR5 (Gerwitz *et al.*, 2005) is of great pertinence, as recent studies report synergism between NOD2 and TLR5 signaling (Netea *et al.*, 2005).

The anterior gradient 2 gene (*AGR2*) is located on chromosome 7p21.3, a region implicated as a susceptibility region for IBD in a previous genome-wide linkage scan (Satsangi *et al.*, 1996), with stronger evidenve for linkage in UC patients. According to this fact, plus the existence of a loss-of-function mouse model, which spontaneously develops symptoms of diarrhea and goblet cell dysfunction, *AGR2* was selected as a good UC candidate gene by Zheng and colleagues (2006). Genotyping German and British patient samples yielded a moderate but consistent association between SNP hCV1702494 in the 5' region of *AGR2* and IBD.

It is evident that genetic association studies alone will not decipher the etiology of inflammatory bowel diseases. Therefore, investigators have successfully employed a broad range of molecular techniques in parallel, including e.g. gene expression studies with microarrays (Heller *et al.*, 1997; Dieckgraefe *et al.*, 2000; Lawrance *et al.*, 2001; Langmann *et al.*, 2004; Okahara *et al.*, 2005; Costello *et al.*, 2005), protein-protein interaction studies by the yeast two-hybrid system (Barnich *et al.*, 2005), or proteome expression profiling using serum of patients (Sauer *et al.*, 2005; Din *et al.*, 2005). Exploiting and combining the results gathered on the genomic, transcriptional, and protein level will lead to a further understanding of complex diseases.

1.2.6 Animal models

Investigations of intestinal inflammation in animals have shown the association between genetic, bacterial, barrier function, and immunological contributions to pathogenesis of CD (Blumberg *et al.*, 1999; Wirtz *et al.*, 2000). However, there has been no *de novo* identification of orthologous genes that contribute to IBD predisposition in humans. Although no animal is an exact replica of CD, owing to the polygenic and complex etiology, the spontaneous disease in SAMP1/Yit mice (Kosiwicz *et al.*, 2001) seems to be closest. Spontaneous intestinal disease that occurs in C3H/HeJBir mice is partly due to their TLR4 deficiency (Brandwein *et al.*, 1997) and, thus, susceptible to bacterial invasion. It is common to all IBD animal models that the expression of a disease depends on colonization with bacterial flora, irrespective of the underlying genetic defect. Under sterile conditions, susceptible animals do not develop IBD-like symptoms. Chronic intestinal inflammation of the bowel that mimics some characteristics of human CD can be chemically induced in mice by *ad libitum* administration of e.g. 2,4,6-trinitrobenzene sulfonic acid (TNBS; Neurath *et al.*, 1996) or dextrane sodium sulfate (DSS; Vowinkel *et al.*, 2004). All exogenous IBD-causing chemicals have in common that they destroy the intestinal epithelial lining leading to massive bacterial invasion and inflammation.

One good mouse model that adds to the key role of TNF-α and its therapeutical implication is that the increased TNF-α biosynthesis, due to deletion of 3' regulatory elements from the TNF transcript, results in a CD-like phenotype (Kontoyiannis *et al.*, 1999). Moreover, $TNF^{-/-}$ mice show marked reduction in chemically induced intestinal inflammation (Neurath *et al.*, 1997).

In addition to innumerable studies on rodents, few investigations focus on higher mammals, though mostly suffering from limited numbers of animals. The cotton-top tamarin (CTT; *Saguinus oedipus*) is an endangered New World primate that develops a highly prevalent and spontaneous idiopathic colitis resembling human UC if kept in captivity (Chalifoux *et al.*, 1985). A study by Mansfield and others (2001) found that enteropathogenic *Escherichia coli* (EPEC) caused acute colitis in CTTs, which was associated with UC. Saunders and colleagues (1999) describe a novel *Heliobacter* species that is associated with UC in CTTs. Colitis-affected animals entered remission when returned to their natural habitat (Wood *et al.*, 2000). Besides cotton-top tamarins, siamang gibbons tend to develop colitis (Stout *et al.*, 1969) when kept in captivity and colitis has been as well reported for chimpanzees (Fremming *et al.*, 1955). Further evidence that stress increases disease activity is provided by studies of Gue *et al.* (1997), who showed that stress may exacerbate experimental colitis in rats, provoked through intracolonic 2,4,6-trinitrobenzenesulfonic acid instillation. Chronic colitis has also

been observed in domesticated animals, such as in dogs (*Canis lupus* f. familiaris) or cats (*Felis silvestris* f. catus) [Cave *et al.*, 2003; Feinstein *et al.*, 1992].

In conclusion, animal models of inflammatory bowel diseases (IBD) have been useful in the identification of those immune responses involved in IBD pathogenesis and in defining the important roles of environmental influences, such as the bacterial flora, but models have proven to be inappropriate for unravelling the predisposing genetic factors.

More details concerning IBD mouse models can be found in the reviews of Byrne *et al.* (2006) and Blumberg *et al.* (1999).

1.3 Aims of this study

The etiology of inflammatory bowel diseases (IBD) is mostly unkown and complex, as both genetic and environmental factors are involved. This complexity presents a major challenge to the efforts to identify the underlying mechanisms behind IBD. As only a limited number of IBD susceptibility genes have been described so far, and as more are thought to exist, the approach of this thesis aimed to identify, map, and study further existing gene variants that contribute to the disease, in other words to connect phenotype with genotype. Two unbiased genome-wide association studies of direct and indirect nature were carried out to find such additional susceptibility variants in a large sample collection of German CD patients. To this end, a cSNP-based scan with approximately 20,000 non-synonymous SNPs using SNPlex™ technology plus an LD-based scan with over 100,000 genome-wide distributed SNPs was performed. Rather than correcting for multiple testing (e.g. Bonferroni correction), different levels of replication were employed. Once an associated mutation was determined, further studies were carried out to investigate its potential effect. In addition, full mutation detection and subsequent fine mapping was carried out for a detected susceptibility region. A summary of the experimental approach is shown in the following flow chart:

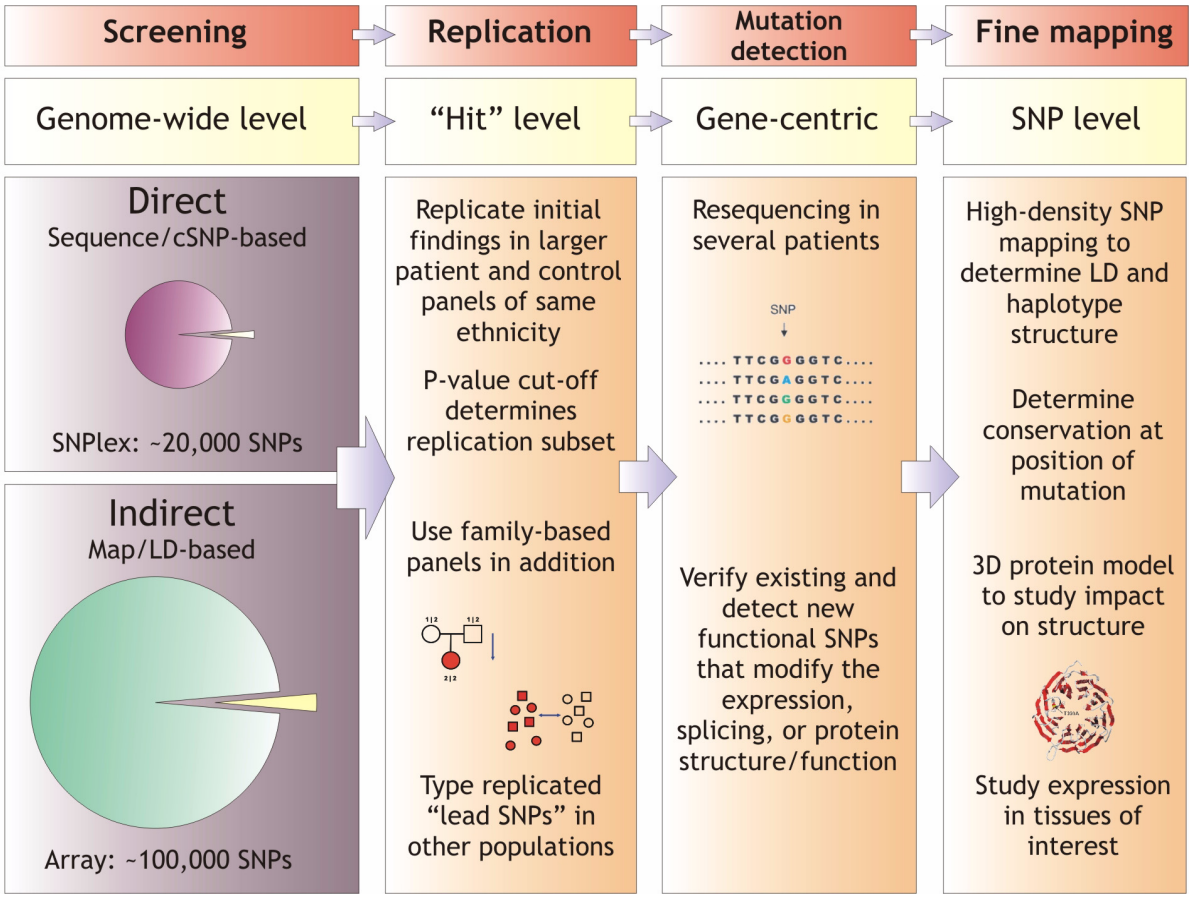

Fig. 1-8 Experimental workflow of this thesis. Two unbiased association screenings generate a great wealth of significant results. Due to massive multiple testing, which leads to a high rate of false positive associations, further genotyping is necessary to "filter" the initial results. Sample sizes are significantly larger in the replication studies, yielding a higher power to detect an association. Finally, causative SNPs are found by resequencing the regions of interest in many individuals.

2 Methods

While this project was part of a collaborative effort, the following techniques were established during this PhD thesis by me:

- Automation of SNPlex™, which included the setup of robotic scripts and liquid classes. LIMS Integration of SNPlex™ was done in collaboration with the bioinformatic group.
- Automation of whole-genome amplification (WGA) and its applicability for the in-house genotyping platform (TaqMan® and SNPlex™). Several quality checks were performed to check for allelic imbalance and other possible pitfalls.
- Development of statistical analysis tools for genome-wide data sets.

2.1 Laboratory information management system (LIMS)

An in-house available database was used to systematically store and retrieve information. Microsoft's SQL Server 7 was the database management system (DBMS) of choice and the name of the used database was 'ibdbase'. Various in-house Visual Basic client applications were used to facilitate lab processes and the analysis of genotyping data (Hampe *et al.*, 2001; Teuber *et al.*, 2005) plus to further enhance quality control. In the database, the following information is integrated: pedigree and phenotype information, sample and microtiter plate information, marker information, genotypes.

Data protection of private information such as the patient's name or address was ensured by using anonymized identifiers. For the POPGEN individuals, an even more sophisticated anonymization was used as described in Krawczak *et al.* (2006).

Since the database has to meet constantly changing demands, it has an open design. Throughout the years, the LIMS has grown, thus to meet for example the requirements of SNPlex™. More information is available at http://www.ikmb.uni-kiel.de/research.html under "Bioinformatics". An older database scheme can be found here: http://www.ikmb.uni-kiel.de/research/bioinformatics/db_schema.pdf.

Genotype data was exported in the standard Linkage Pedigree format (pre-file) as required by the used analysis programs. The file did not have a header line (i.e. the first line should be for the first individual, not the names of the columns):

```
3 128      91     2      1 2    3 3    0 0    4 2
a b c      de     f    ----------g-------------
```

a. pedigree name: A unique alphanumeric identifier for this individual's family. Unrelated individuals should not share a pedigree name.

b. individual ID: An alphanumeric identifier for this individual. Should be unique within his family (see above).
c. father's ID: Identifier corresponding to father's individual ID or "0" if unknown father. Note that if a father ID is specified, the father must also appear in the file.
d. mother's ID: Identifier corresponding to mother's individual ID or "0" if unknown mother Note that if a mother ID is specified, the mother must also appear in the file.
e. sex: Individual's gender (1 = male, 2 = female).
f. affectation status: Affectation status to be used for association tests (0 = unknown, 1 = unaffected, 2 = affected).
g. marker genotypes: Each marker is represented by two columns (one for each allele, separated by a space) and coded either “ACGT” or “1–4” where: 1=A, 2=C, 3=G, T=4. A “0” in any of the marker genotype position (as in the the genotypes for the third marker above) indicates missing data.

2.2 Recruitment

An overview of the patient and control samples used in this study is provided in table 2-1. The German patient panels A, B, D, E, and F were collected at the Charité University Hospital (Berlin, Germany), the Department of General Internal Medicine at the Christian-Albrechts-University (Kiel, Germany) with the support of the German Crohn and Colitis Foundation (DCCV).

Clinical, radiological and endoscopic (type of lesions, distribution) examinations were required to unequivocally confirm the diagnosis of CD (Lennard-Jones *et al.*, 1989; Truelove *et al.*, 1976). Histological findings also had to be confirmative or compatible with this diagnosis. In case of uncertainty, the diagnosis of indeterminate colitis was assigned and the patient excluded from the study. The sample has been used in several previous studies within the collaborative group (Croucher *et al.*, 2003; Curran *et al.*, 1998; Hampe *et al.*, 1999; Hampe *et al.*, 1998); the respective publications provide a more extensive account of the phenotyping techniques employed.

German control individuals were obtained from the German population biobank PopGen as described (Krawczak *et al.*, 2006; Lamina *et al.*, 2005). Popgen targets the population of northern Schleswig-Holstein (1.1 million people) that is enclosesed by the Danish border (North), the North Sea and Elbe river (West), the Baltic Sea (East), and the Kiel Canal (South). The latter can only be crossed by a limited number of ferries and bridges. There is no historical, demographic or genetic evidence suggesting that etiological factors relevant in the present context differ substantially between this and other regions of Germany (Krawczak *et al.*, 2006).

About half of the control population consisted of healthy, unrelated blood donors, that were collected at the Institute of Transfusion Medicine in the clinic of Kiel. Health status of all blood donors was categorized by questioning and standard diagnostic blood parameters.

Patients for panel E were mostly and controls (Popgen) completely from Northern Germany. Furthermore, patients with familial CD, early age of onset and a clear diagnosis were preferentially selected. This was done to provide a "genetically enriched" sample for the initial screening, as it is more likely that genetic factors play a major role in the etiology of such "extreme" phenotypes.

The UK patients were collected as described previously by the collaborating centre (Onnie *et al.*, 2006). UK Population controls were obtained from the 1958 British Birth Cohort (http://www.b58cgene.sgul.ac.uk). Recruitment protocols were approved by ethics committees at all participating centers prior to commencement of the study and participants gave written, informed consent (9.9 on page 197) besides a detailed questionnaire (9.8 on page 191). The 382 French-Canadian parent-parent-child (PPC) trios were collected through the collaborator Genizon Biosciences. Patients were older than 18 years, had 4 French-Canadian grandparents, and diagnosis were obtained by either colonoscopy, radiological examination with barium, abdominal surgical operation, or exploratory biopsies. Patients with UC or any other form of bowel disease were excluded.

Study	Panel	Patients	Controls	Trios
cSNP	Crohn's disease (Germany) - A	735	368	–
	Crohn's disease (Germany) - B	498	1032	380
100k	Crohn's disease (Germany) - C	393	399	
	Crohn's disease (Germany) - D	567	1082	375
Common	Ulcerative colitis (Germany) - E	788	1032	439
	Crohn's disease (UK) - F	509	656	–
	Ulcerative colitis (UK) - G	442	521	–
	Crohn's disease (French-Canadian) - H	–	–	382

Table 2-1 IBD patient and control samples used for association analysis. The patient samples are organized in 'panels' that correspond to successive steps of the study. Index cases from trios were also used in the case-control analyses so that, for example, a total of 878 cases (498 + 380) were available for the case-control comparison in panel B. The controls from CD panel B and D were also used for the analysis of UC (E). Panels A and B overlap with C and D, which is valid, as these are used in two independent studies.

2.3 Sample preparation

After recruitment of individuals, genomic DNA was isolated from the leukocytes of donated blood samples. DNA samples were quality checked on an agarose gel and used for whole genome amplification. Amplified products were arrayed on 96 well microtiterplates (MTP), thus increasing the throughput for downstream processes. The latter was even increased by merging 4 x 96 well MTPs into a 384 well MTP.

2.3.1 DNA extraction from blood

Genomic DNA (gDNA) was extracted from EDTA whole blood samples, using the Invisorb® Blood Giga Kit for DNA isolation. Only a few modifications were made to the manufacturer's protocol. Blood samples were stored at -80°C and thawed in a cold water bath with the lid sticking right out of the water.

Lysis of erythrocytes, while leukocytes stayed intact, was performed by incubating 9 ml of blood for 10 min with 30 ml of cold buffer 1 at room temperature. Afterwards, the suspension was centrifuged for 3 min at 3,000 rpm and the supernatant was carefully discarded. This step was repeated with 20 ml buffer 1 until the leucocyte containing pellet was free of haem. A DNA contamination with haem causes problems during downstream experiments, as haem inhibits PCR reactions (Heath *et al.*, 1999). The pellet was resuspended in 3 ml of buffer 2 and 50 µl of Proteinase K and incubated for 2 hours in a 60°C water bath under continuous shaking (95 turns/min) to increase the lysis efficiency. This step leads to the lysis of the leukocytes and their nuclei and therefore to a release of DNA into the suspension. To seperate the DNA from cell and protein fragments, 1.8 ml of buffer 3 were added. Vigorous mixing and a 5 min incubation on ice were subsequently carried through. Then, the mix was centrifuged for 15 min at 5,000 rpm. Afterwards, the cleared supernatant was transferred into a 15 ml centrifuge tube. For the precipitation of the DNA, 10 ml of 96% ethanol (according to Bearden *et al.*, 1974; Shapiro *et al.*, 1981; Wilcockson *et al.*, 1975) were added and the tube carefully inverted several times. If precipitation did not take place, the tube was incubated for 2 h at -20°C. The precipitated DNA was obtained by centrifugation for 3 min at 5,000 rpm and was then collected with a pipette tip and transferred into 2 ml reaction tubes containing 1 ml of 70% ethanol. The DNA pellet was rinsed by vortexing and subsequently centrifuged for 2 min at 13,000 rpm. Finally, the ethanol was removed with a pipette and samples were dried for 10 min at room temperature.

All samples were quantified by measuring the concentration with PicoGreen® as described in 2.4 on page 29. The purified gDNA was resuspended in 500 µl of 1x TE buffer and stored at +4°C for short periods or at -20°C for long periods. Average yields were 200 ng/µl, which corresponds to an amount of 100 µg DNA. If no pellet was visible or DNA concentration was be-

low 100 ng/µl, the procedure was repeated. This was possible as every individual donated two to three blood samples. If no DNA was left, individuals were contacted for an additional blood sample. In addition to a label on the lid, each 2 ml reaction tube received a barcode for tracking it in the LIMS.

2.3.2 Plate design

For genotyping, 92 DNA samples were arranged in a 96 well format according to a pre-defined plate layout. Individual of the same pedigree were kept on the same plate. Four wells were used for internal controls and quality control, such as three empty wells (no template controls [NTCs]) and one positive control, the so called CEPH (Fondation Jean Dausset - Cenre d'Etude du Polymorphisme Humain, Paris, France) control. Negative controls were used to reveal potential contaminations. As four 96 well plates were merged into a single 384 well plate, there were four positions with CEPH cell-line DNA in the final plate layout. Genotype concordance was checked for every assay among these four wells holding the same DNA. A low genotype concordance indicated an assay problem or a contamination problem. Each MT plate was labeled with a unique plate name for database storage to allow unmistakable identification. SNPlex™ plates received the prefix "X", e.g. XG01.

	1	2	3	4	5	6	7	8	9	10	11	12
A	DNA01	DNA09	DNA17	DNA23	DNA31	DNA39	DNA47	DNA55	DNA63	DNA71	DNA78	DNA86
B	DNA02	DNA10	DNA18	DNA24	DNA32	DNA40	DNA48	DNA56	DNA64	DNA72	DNA79	DNA87
C	DNA03	DNA11	CEPH	DNA25	DNA33	DNA41	DNA49	DNA57	DNA65	DNA73	DNA80	DNA88
D	DNA04	DNA12	EMPTY	DNA26	DNA34	DNA42	DNA50	DNA58	DNA66	EMPTY	DNA81	DNA89
E	DNA05	DNA13	DNA19	DNA27	DNA35	DNA43	DNA51	DNA59	DNA67	DNA74	DNA82	DNA90
F	DNA06	DNA14	DNA20	DNA28	DNA36	DNA44	DNA52	DNA60	DNA68	DNA75	DNA83	DNA91
G	DNA07	DNA15	DNA21	DNA29	DNA37	DNA45	DNA53	DNA61	DNA69	DNA76	DNA84	DNA92
H	DNA08	DNA16	DNA22	DNA30	DNA38	DNA46	DNA54	DNA62	DNA70	DNA77	DNA85	EMPTY

Fig. 2-1 New ICMB plate layout. Wells D3 and D10 were used as negative controls for TaqMan® genotyping and for allelic ladder in case of SNPlex™.

2.4 Measurement of DNA concentration

The most commonly used technique for measuring nucleic acid concentration is the determination of absorbance at 260 nm (A260). The major disadvantages of the absorbance method are the large relative contribution of nucleotides and single-stranded nucleic acids to the signal, the interference caused by contaminants commonly found in nucleic acid preparations, the inability to distinguish between DNA and RNA and the relative insensitivity of the assay (an A260 of 0.1 corresponds to a 5 µg/mL dsDNA solution).

Concentrations of the genomic DNA were measured using the PicoGreen® method (Ahn *et al.*, 1996; Rengarajan *et al.*, 2002). The PicoGreen® reagent is a proprietary, unsymmetrical cyanine dye. Free dye is essentially nonfluorescent and exhibits >1000-fold fluorescence enhancement upon binding to dsDNA (with excitation and emission maxima of ~500 nm and ~520 nm, respectively). The assay displays a linear correlation between dsDNA concentration and fluorescence and has a detection range extending from 25 pg/mL to 1 µg/mL dsDNA using a single dye concentration. The assay is highly selective for dsDNA over RNA, single-stranded DNA (ssDNA) and oligonucleotides. Furthermore, there is essentially no base selectivity and assay results are not compromised by proteins, nucleotides and other common sample contaminants (Singer *et al.*, 1997; see also fig. 2-2).

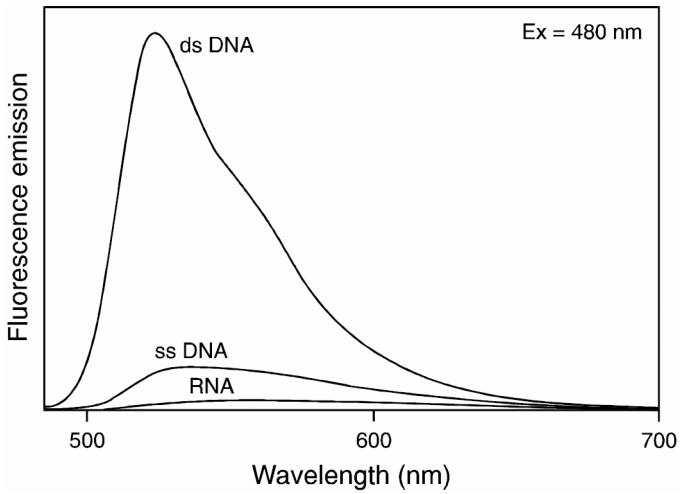

Fig. 2-2 Fluorescence enhancement of PicoGreen® reagent upon binding dsDNA, ssDNA and RNA. Samples containing 500 ng/mL calf thymus DNA, M13 ssDNA or E. coli ribosomal RNA were added to cuvettes containing PicoGreen® reagent in TE. Samples were excited at 480 nm and the fluorescence emission spectra were collected using a spectrofluorometer. Emission spectra for samples containing dye and nucleic acids, as well as for dye alone (baseline), are shown. *Illustration taken from the manufacturer's protocol.*

Pipetting, dilution, and normalization steps were fully automated using a TECAN pipetting robot (Genesis RSP 150). With the in-house software SampleTool, 96 samples can be measured in parallel. 32 samples were arranged in duplicates in a 96 well optical Sarstedt plate, as shown in fig. 2-3. Thus, the worktable of the robot has capacity for three of such plates.

	1	2	3	4	5	6	7	8	9	10	11	12
A	1.0x TE	1.0x TE	EMPTY	EMPTY	DNA01	DNA09	DNA17	DNA25	DNA01	DNA09	DNA17	DNA25
B	1.0x TE	1.0x TE	EMPTY	EMPTY	DNA02	DNA10	DNA18	DNA26	DNA02	DNA10	DNA18	DNA26
C	1 ng/ml	1 ng/ml	EMPTY	EMPTY	DNA03	DNA11	DNA19	DNA27	DNA03	DNA11	DNA19	DNA27
D	10 ng/ml	10 ng/ml	EMPTY	EMPTY	DNA04	DNA12	DNA20	DNA28	DNA04	DNA12	DNA20	DNA28
E	100 ng/ml	100 ng/ml	EMPTY	EMPTY	DNA05	DNA13	DNA21	DNA29	DNA05	DNA13	DNA21	DNA29
F	500 ng/ml	500 ng/ml	EMPTY	EMPTY	DNA06	DNA14	DNA22	DNA30	DNA06	DNA14	DNA22	DNA30
G	1 µg/ml	1 µg/ml	EMPTY	EMPTY	DNA07	DNA15	DNA23	DNA31	DNA07	DNA15	DNA23	DNA31
H	EMPTY	EMPTY	EMPTY	EMPTY	DNA08	DNA16	DNA24	DNA32	DNA08	DNA16	DNA24	DNA32

TE-buffer
Standards
EMPTY-controls
Samples in duplicates

Fig. 2-3 Plate setup for PicoGreen® concentration measurements of 32 samples.

After a 5 min incubation in the dark, plates were read in a TECAN Spectrafluor Plus fluorometer. Concentrations and dilution factors were calculated using the average of the two measurements for each DNA and a standard curve. Normalized DNA concentrations were verified by a second measurement. If single tubes were measured without SampleTool, the Gemini scripts `WL_Fluoro_V2.gem` and `WL_dDNA_Eppi_V2.gem` were used.

2.5 Whole genome amplification (WGA)

Since the yield of DNA from individual patient samples is limited and as larger experiments require larger amounts of DNA (Lasken *et al.*, 2003), whole-genome amplification was established and used for genotyping plate production. The multiple displacement amplification (MDA, see fig. 2-4) method, which relies on isothermal amplification using the DNA polymerase of the bacteriophage ϕ29 from *Bacillus subtilis*, is a recently developed technique for high performance WGA (Lovmar *et al.*, 2006). MDA was first described by Dean *et al.* (2002).

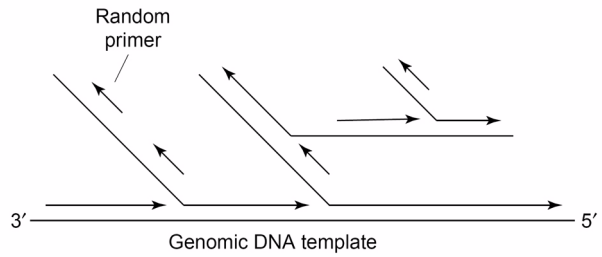

Fig. 2-4 Multiple displacement amplification reaction. DNA synthesis is primed by random hexamers. Exponential amplification occurs by a 'hyperbranching' mechanism. Unlike PCR, which requires thermal cycling to repeatedly melt template and anneal primers, the ϕ29 DNA polymerase acts at 30°C to concurrently extend primers as it displaces downstream DNA products.

The yield of a MDA reaction is less dependent on the amount of input DNA, but because the reaction is self-limited, the yield will depend on the reaction conditions and amount of reagents and hence on the reaction volume (Dean *et al.*, 2002). Consequently, varying DNA concentrations in the initial sample will plateau during MDA, which is a potential benefit for MDA in large-scale genotyping applications because it unifies and increases the DNA concentrations of the samples (Lovmar *et al.*, 2006). Therefore, besides a 100-fold amplification of the DNA, a resource-consuming normalization step with PicoGreen® becomes pointless after the WGA. In contrast, if genomic DNA was used for genotyping, a measurement was necessary (see 2.4 on page 29).

1 µl of stock DNA was used as an input for the reaction, with concentrations ranging from 50 ng/µl to as much as 300 ng/µl. More input DNA was used than recommended by the manufacturer (10 ng recommended) as it was shown that higher amounts of input DNA significantly increase genotyping performance and concordance rates (Bergen *et al.*, 2005).

The kits GenomiPhi v1 and v2 were used for amplification of DNA. All steps were carried out according to the kit's protocol. Pipetting robots were used to set up the WGA reactions (Scripts `Pipet_WGA_nach_SNPlex.gem` and `_Verduennung_nach_WGA_Andre.gem`). Version 1 is an overnight reaction while version 2 yields the same amount after two hours reaction time. Furthermore, there is no random amplification in empty wells thus no artefacts are generated. Generated fragments range between 10 and 100 kb.

When the WGA was finished, the 20 µl (~5 µg) reaction volume was diluted in the following way:

1. 1:5 with 1x TE-buffer, final volume of 100 µl (~50 ng/µl)
2. The 100 µl were split (2x 50 µl) to two fresh 96 well MT plates
3. One plate was used for SNPlex™ and the other for TaqMan® plate production.
4. In case of SNPlex™, the WGA-DNA was fragmented for 5 minutes according to the SNPlex™ protocol and then diluted 1:2 with 1x TE-buffer to a final volume of 100 µl (~25 ng/µl).
5. For TaqMan®, the 50 µl WGA-DNA were further diluted 1:80 to a final volume of 4 ml (~0.63 ng/µl).
6. Four 96 deepwell microtiter plates were then merged to one 384 deepwell MTP using the script `Merge_Deepwells.gem` (if necessary, merging was undone with the script `Unmerge_Deepwells.gem`). This was accomplished with a 96-needle multi pipetting device (Te-MO, TECAN) on a TECAN pipetting robot.

Aliqouts of 5 µl were dispensed via a 384-channel Robbins Scientific Hydra microdispenser to fresh 384 MT PCR plates. For TaqMan®, the "copied" plates were dried down at 60°C for one hour and subsequently sealed. In case of SNPlex™, the plates were left to dry overnight in a closed cupboard. Dried plates were sealed and ready-to-use for genotyping.

TaqMan® PCR plates were kept for two years while SNPlex™ PCR plates were usable for six months (based on empirical experiences). Each plate received a unique barcode label for database tracking.

Various studies have shown that there almost no differences (i.e. no allelic imbalance) exist between genotyping results of genomic compared to WGA DNA, thus making MDA a reliable provider of nearly unlimited amounts of DNA. A good overview of various studies that examined the genotype concordance of WGA-DNA and genomic DNA is given in Lovmar *et al.* (2006). Own results of such a comparison are listed in section 3.5.5.1 on page 115.

2.6 Agarose gel electrophoresis

PCR fragments, depending on the expected size, were separated in 1–2% agarose gels (Takahashi *et al.*, 1969). The smaller the expected product, the higher the concentration of agarose should be for a better seperation and resolution (table 2-2). In case of genomic DNA, as necessary for the quality control of gDNA, an agarose concentration of 0.8% was used.

To visualize DNA-bands ethidiumbromide was added to the gels (1.5 µl EtBr per 100 ml gel solution; 10 mg/ml). 0.5x TAE (Tris-acetate-EDTA) buffer or 0.5x TBE (Tris-borate-EDTA) buffer was used as a running buffer. Pouring gels was done by boiling the buffer/agarose mixture in a microwave, letting it cool down to approx. 60°C, adding EtBr under the hood, pouring it into the casting device, and leaving it there for approx. 30 min until polymerization had finished. Large gels were made using 250 or 300 ml, medium ones using 100 ml and small ones with 50 ml buffer. 2x DNA-loading buffer (0.25% bromphenol blue + 0.25% xylene cyanol FF + 30% glycerol in water) was added to the samples (5 µl for 10 µl sample). In 1% agarose gels, bromophenol blue migrates with 300 bp linear double-stranded DNA fragment, whereas xylene cyanol FF migrates at approximately the same rate as linear double-stranded DNA 4 kb length. These relationships are not significantly affected by the concentration (0.5 to 1.4%) of agarose in the gel.

Fragments were separated in horizontal gel chambers electrophoretically at 110 V for 70 min (for large TBE gels up to 150 V possible) until complete band seperation. The size of DNA fragments was estimated under UV-illumination with the Bio-Rad Gel Doc XR gel documentation system according to the comigrating DNA-size standards such as 100 base pair ladder. A single clear band of the expected size and only a low amount of primer-dimers indicated optimal PCR conditions.

Agarose in gel	Efficient range of separation
0.3%	60.0 kb – 5.0 kb
0.6%	20.0 kb – 1.0 kb
0.7%	10.0 kb – 0.8 kb
0.9%	7.0 kb – 0.5 kb
1.2%	6.0 kb – 0.4 kb
1.5%	4.0 kb – 0.2 kb
2.0%	3.0 kb – 0.1 kb

Table 2-2 DNA seperation in agarose.

2.7 Mutation detection

For the follow-up of candidate gene studies, knowledge of all variations in a limited amount of sequence (e.g. a gene) is needed. In the present study, all exons plus splice sites of *ATG16L1* and *NELL1* were resequenced in 47 CD patients. This methodology of SNP verification and finding is termed mutation detection.

2.7.1 Primer design

Primers for exonic sequences were designed in introns, with the amplicon covering the exon itself, splice sites, and at least 50 bp of flanking intronic sequence. For longer exons or difficult GC-rich regions, primers generating overlapping amplicons were designed. Sequencing was performed for both orientations (forward and reverse), to circumvent sequencing artefacts. The webinterface Primer3 (Rozen *et al.*, 2000) was used for the design. Modifications to the default settings were as follows:

- Mispriming Library (repeat library): HUMAN
  Only a single specific binding site was allowed for each primer, therefore repeating elements in the DNA were excluded by choosing this option.
- Product Size Ranges: 100–850
  Amplicon lengths need to be less than 850 bp, which is the maximum read of the 3730xl sequencer.
- Primer Size: min = 18; opt = 24; max = 30
- Primer Tm: min = 55; opt = 65; max = 75
  High annealing temperatures facilitate specific hybridization.
- Primer GC%: min = 40; opt = 50; max = 60
- CG Clamp: 1
  A “C” or “G” residue at the 3’ end supports binding and elongation of the polymerase.
- Max Poly-X: 4
  Only four repetitive mononucleotides were allowed to avoid mispriming and frameshifts.

Primer sequences were checked for potential hairpin and primer dimer formations. Finally, an *in silico* PCR was made at the UCSC webpage (Kent *et al.*, 2002) to check whether a single amplicon is generated.

All primers were ordered through Eurogentec (Seraing, Belgien) on a 40 nmol scale. The quality was ensured by mass spectrometry by the manufacturer. Primer stocks were diluted to a concentration of 100 µM with DDW, e.g. for 40 nmol of primers, 400 µl water were added.

2.7.2 Polymerase chain reaction (PCR)

The polymerase chain reaction (PCR) is based on the three repeating steps:

- denaturation of double-stranded template by high temperature
- annealing of two oligonucleotides (primers) to template
- elongation by a thermostable DNA polymerase and dNTPs

The transcript is exponentially amplified by cycling through the above steps (Saiki *et al.*, 1985; Mullis *et al.*, 1987). This is accomplished by switching between the specific temperatures for each step. Therefore programmable thermocyclers are used. Reactions terminate if either the template has reached a critical concentration, resources are used up (dNTPs, primers) or if the enzymatic activitiy of the polymerase has vanished. A linear amplification was done by using only one primer. According to Mülhardt (2002), the mother of all PCR programs is the following:

Temperature	Time	Cycle(s)
94°C	5 min	1
94°C	30 s	30
55°C	30 s	
72°C	90 s	
72°C	5 min	1
4°C	∞	1

Table 2-3 Standard PCR protocol.

A typical protocol for a PCR master mix is given in table 2-4, page 36.

2.7.3 DNA sequencing

Sequencing of DNA, which basicly means to determine the order of nucleotides in a given DNA fragment, was performed by means of a modified protocol of the chain terminating technique that was developed by Frederick Sanger *et al.* in 1977. Rather than using labelled primers, labelled terminators, the so-called di-deoxynuleotides (ddNTPs) were used. The major advantage of this approach is the complete sequencing set can be performed in a single reaction, rather than the four needed with the labeled-primer approach. This is accomplished by labelling each of the dideoxynucleotide chain-terminators with a separate fluorescent dye, which fluoresces at a different wavelength.

Primer optimization and PCR

To determine the optimal melting temperature (T_m) of the primers and the perfect PCR conditions, a gradient (12°C across 12 positions) PCR was carried out using the following reaction mix:

Reagent	volume [µl]	final conc.
GeneAmp® 10x PCR buffer II	2.50	1x
$MgCl_2$ [25 mM]	2.00	2 mM
dNTPs [10 mM each]	0.50	200 µM each
Forward Primer [10 µM]	0.10	0.04 µM
Reverse Primer [10 µM]	0.10	0.04 µM
Ampli Taq Gold® [5 U/µl]	0.15	0.03 U/µl
DNA [5 ng/µl]	1.00	0.75 U
Water	18.65	
Total	**25.00**	

Table 2-4 PCR mix for a single reaction.

Specificity was enhanced by using a touchdown thermoprofile (Don *et al.*, 1991).

Temperature	Time	Cycle(s)	Comment
95°C	5 min	1	AmpliTaq Gold® is activated
95°C	30 s	16	
64°C	30 s		recommended annealing temperature of manufacturer + 8°C td=-0.5°C/cycle
72°C	1 min		elongation time depends on length of template: 1000 nt/min
95°C	30 s	15	
56°C	30 s		recommended annealing temperature of manufacturer
72°C	1 min		elongation time depends on length of template: 1000 nt/min
72°C	10 min	1	filling up the ends, especially needed for TA-cloning
4°C	∞	1	

Table 2-5 Primer optimization PCR program.

If no or many bands were seen on the gel, the $MgCl_2$ concentration was either in- or decreased and the PCR repeated. In GC-rich regions, the Qiagen Taq DNA Polymerase Kit was used and steps were carried out according to the protocol. The kit included Q-solution, which facilitates amplification of difficult templates by modifying the melting behavior of nucleic acids.

The most appropriate conditions were used as found out by the optimization, i.e. instead of a gradient, the optimum T_m was used. All reactions were carried out in 96 well plates with 5 ng of liquid or dried DNA. Forward and reverse reactions were performed together on a 96 well MTP for 47 samples. One well was used as a negative control to check for (cross-)contaminations.

Digest

5 µl of PCR product were tested on an 1.5% agarose gel (in 300 ml TBE, 1% EtBr, 150 V for 50 min). If a sharp band was visible, the PCR product was cleaned using an enzymatic digest. Highly concentrated PCR products were diluted 1:5 with water before the digest. The enzymatic digest was performed in a new plate and to remove superfluous primers. For the digest, 8 µl PCR product were mixed with 0.30 µl shrimp alkaline phosphatase (SAP; 1 U/µl), 0.15 µl exonuclease I (ExoI; 10 U/µl), and 1.55 µl water and incubated at 37°C for 15 min. SAP digests single-stranded DNA molecules as described by Berthold *et al.* (1976). Remaining dNTPs were also destroyed as a dephosphorylating SAP was used (Sauer *et al.*, 2000). Thus, a possibly resulting dNTP/ddNTP imbalance was avoided that could disturb the sequencing reaction. The reaction was stopped by a subsequent heat step of 72°C for 15 min. This is an essentiel step as an active SAP could negatively effect the sequencing reaction.

Sequencing reaction

Sequencing reactions were carried out according to the kit's protocol. 2 µl of the digested PCR product were used for the sequencing reaction. Furthermore, 1.0 µl primer [3.2 µM] (either forward or reverse), 1.0 µl BigDye™ Terminator Ready Reaction Mix version 1.1 from the kit, 0.5 µl 5x sample buffer and 5.5 µl water were added to the reaction (total of 10 µl). The following cycle protocol for the sequencing reaction was used:

Temperature	Time	Cycle(s)	Comment
95°C	30 s	25	
$T_m$opt	15 s		optimized annealing temperature
60°C	4 min		chain termination reaction
4°C	∞	1	

Table 2-6 Sequencing thermoprofile. As the BigDye Reaction Mix has its temperature optimum at 60°C, the upper limit for the optimized annealing temperature was 60°C to avoid premature termination stops (Wen *et al.*, 2001).

To lower the background fluorescence noise, unincorporated nucleotides and primers were removed using a Sephadex spin column plate. The latter was prepared by adding Sephadex powder to a fresh MAHVN 4550 plate with the aid of a multi-screen column loader. Sephadex is cross-linked dextran composed of polyglucose (Mort *et al.*, 1998). 300 µl of water were add-

ed to each column, the lid was closed and after soaking for at least 2 hours at room temperature, as a result of which a three-dimensional network and pores (20–80 µm) are formed, the plate was centrifuged for 5 min at 2100 rpm. This was necessary to dry the column. Another 150 µl of water were added to the columns and the plate was immediately centrifuged at 2100 rpm for 5 min. The sequencing product was diluted to a final volume of 30 µl with 10 µl PCR water and pipetted into the centre of the spin column. The column plate was fitted into a MicroAmp® Optical 96-well reaction plate and centrifuged for 5 min at 2100 rpm. The flow through contained the purified sequencing product and the plate was sealed with an aluminum adhesive cover. The sequence detection was performed with an automated, high-throughput 96-capillary fluorescence detection system, the 3730xl DNA Analyzer.

2.7.3.1 Sequence analysis

Traces were manually inspected with Sequencher 4.2 and 4.5 and variation discovery was carried out with novoSNP (Weckx *et al.*, 2005) that allows automated detection of sequence variations from sequence trace files in a fast and reliable manner. A relatively low cut-off score of 10 was used, to eliminate the risk of false-negative variations, as recommended by the developers. Besides biallelic single nucleotide polymorphisms, novoSNP is capable of detecting insertion-deletion polymorphisms (INDELs). The latter have gained attention just recently in context of complex diseases (McGovern *et al.*, 2005; McCarroll *et al.*, 2006; Hind *et al.*, 2006; Conrad *et al.*, 2006).

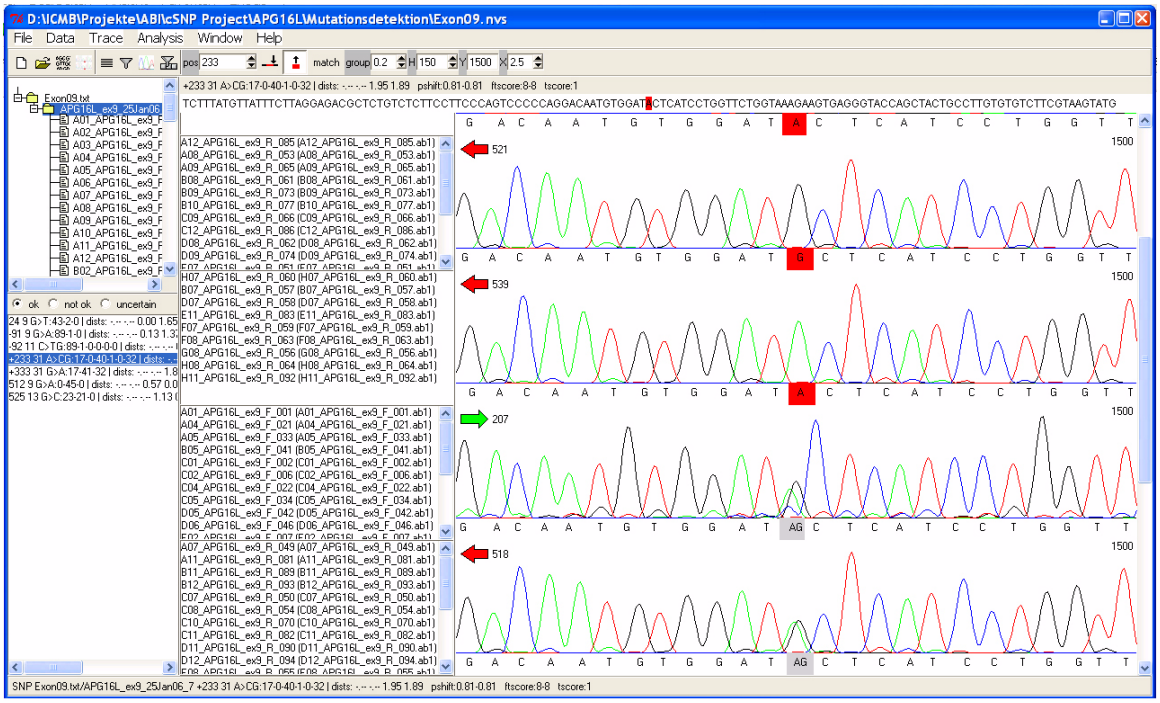

Fig. 2-5 novoSNP's graphical user interface. The main frame is centered on the "A/G" SNP rs2241880 as seen in Exon nine of *ATG16L1*: table 3-8, page 92.

2.8 Genotyping

The key method of this study was genotyping, which refers to the process of determining the genotype of an individual with a biological assay. Depending on the requirements of the experiment, one of the following three genotyping methods were used: TaqMan®, SNPlex™, or Affymetrix SNP chips.

2.8.1 TaqMan®

For genotyping only a small number of SNPs or for follow-up of SNPs that did not work with SNPlex™, the robust genotyping method TaqMan® was chosen. TaqMan® is a single-tube PCR assay (De La Vega *et al.*, 2005; Holland *et al.*, 1991; Livak *et al.*, 1995; Livak *et al.*, 1999; McGuigan *et al.*, 2002) that exploits the 5' exonuclease activity of DNA polymerase.

The assay includes two locus-specific PCR primers that flank the SNP of interest, and two allele-specific oligonucleotide TaqMan® probes. These probes have a fluorescent reporter dye at the 5' end, and a non-fluorescent quencher (NFQ) with a minor groove binder (MGB; N-methylpyrrolecarboxamide) at the 3' end (Afonina *et al.*, 1997).

An intact probe emits little fluorescence when excited, because the close physical proximity of the 5' fluorophore to the 3' quencher causes the fluorescent (or better: Förster) resonance energy transfer (FRET) effect to quench the fluorescence emitted by the fluorophore (Livak *et al.*, 1995). A fluorescent signal is generated when the intact probe, which is hybridized to the target allele, is cleaved by the 5' exonuclease activity of Taq DNA polymerase during each cycle of the PCR reaction. The PCR primers amplify a specific locus on the genomic DNA template, and each fluorescent dye-labeled hybridization probe reports the presence of its associated allele in the DNA sample (see fig 2-6, page 40). In each PCR cycle, cleavage of one or both allele-specific probes produces an exponentially increasing fluorescent signal by freeing the 5' fluorophore from the 3' quencher. The use of two probes, one specific to each allele of the SNP and labeled with two fluorophores, allows detection of both alleles in a single tube. TaqMan® probes were labelled with the fluorescent dyes FAM™ (6-carboxyfluorescein) or VIC® (proprietary dye from Applied Biosystems) and with the quencher TAMRA™ (6-carboxytetramethylrhodamine, succinimidyl ester). Old non-MGB assays were labeled with TET (5'-Tetrachloro-Fluorescein) instead of VIC®. The passive reference dye ROX (6-carboxy-X-rhodamine, succinimidyl ester) was included in every well for normalization.

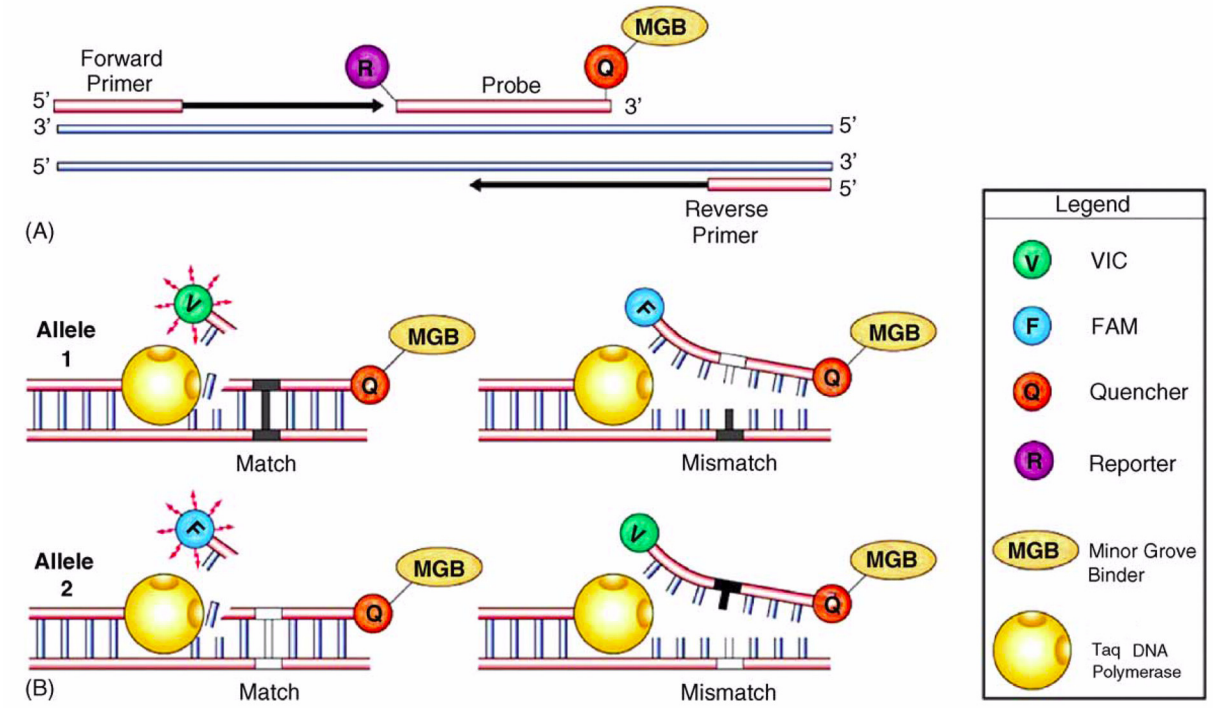

Fig. 2-6 TaqMan® - a fluorogenic 5' nuclease assay. Probe binding and primer extension in a TaqMan® SNP Genotyping Assay. (B) Allelic discrimination is achieved by the selective annealing of matching probe and template sequences, which generates an allele-specific (fluorescent dye-specific) signal.

Fluorogenic probes with an MGB produce enhanced allelic discrimination, because the MGB stabilizes the double-stranded probe template structure, thereby increasing the probe T_m without increasing probe length (Kutyavin *et al.*, 2000). This provides enhanced mismatch discrimination between these shorter probes, resulting in improved allele specificity. These probes also increase the signal-to-noise ratio of an assay, because the reduced distance between the 5' fluorophore and the 3' quencher provides more efficient quenching of an intact probe. Furthermore, a non-fluorescent quencher replaces the TAMRA™ quencher, reducing the background fluorescence and improving the spectral discrimination for MGB assays.

Most of the genotyping assays were Assays-on-Demand (AoD), a pre-designed and validated assay format offered by the manufacturer. If pre-designed assays were not availabe, Assays-by-Design (AbD) were ordered, i.e. the manufacturer made the design according to a user-defined sequence. Both types of assays needed no further optimization.

For the self-designed assays, the program Primer Express 2.0 was used with the default settings (Dieffenbach *et al.*, 1993; Lowe *et al.*, 1990). Probes were manually selected according to the following criteria:

- Polymorphism in middle or last part of probe (minimum of two bases before the 3' end of the probe)
- No guanine (G) on the 5' end
- Melting temperature (T_m) must be 7°C higher than the highest primer T_m (65–67°C)
- Difference in probe T_ms must be smaller than 1°C
- Length of 13–20 bases
- Less than four contiguous G's; no G on the 5' end; less Gs than Cs (cytosines)
- No repetitive elements in the target region (http:\\www.repeatmasker.org)
- Only one SNP per probe is allowed

Primers were chosen that had a GC-content of 30–80% and that generated an amplicon between 50 and 150 bp in length. Less than 2°C T_m difference were allowed between the two primers. Further primer design requirements included:

- Length of 9–40 bases
- Maximum of two Gs or Cs inside the last five bases of the 3' end
- 3' end of primer as close to the probe as possible without overlapping
- Presence of SNPs at primer binding sites is not allowed, absence of repetitive elements is desirable

In the primer test panel of the software, primers were tested for dimerization and hairpin loops (secondary structure formation). Not more than three self-annealing or loop-bonds in a row were accepted, primer dimerisation for up to four bonds was allowed. T_ms were calculated using an oligonucleotide concentration of 50 nM and a salt concentration of 50 mM.

An annealing temperature optimization was carried through for self-designed TaqMan®-MGB assays. The experiment was performed with half 96 well MTPs (47x 2.5 ng control sample DNAs) for different annealing/elongation (AE) temperatures. In the first round, 60°C or 62°C were chosen. If the assay failed to discriminate, 58°C and 64°C were tested. Further optimization steps included variable amounts of $MgCl_2$ or longer AE times.

The final reaction mix was set up using the following the protocol:

Reagent	Assay-by-Design	Assay-on-Demand	self-designed Assay
	volume [µl]	volume [µl]	volume [µl]
TaqMan® PCR master mix	2.500	2.500	2.500
FAM™ probe [10 µM]	–	–	0.100
VIC® probe [10 µM]	–	–	0.100
Forward Primer [20 µM]	–	–	0.225
Reverse Primer [20 µM]	–	–	0.225
Ready-to-use assay mix	0.063	0.250	–
$H_2O_{dest.}$	2.437	2.250	1.850
Total	**5.000**	**5.000**	**5.000**

Table 2-7 TaqMan® pipetting scheme.

5 µl of the reaction mix were added to the 384 well plates with the dried genomic DNA by a TECAN Genesis RSP 150 multipipetting robot. This process was executed and tracked with the in-house software Pipettor which is part of the integrated LIMS (Hampe *et al.*, 2001). For all three assay types, the same two-step PCR protocol was used (table 2-8, page 42).

Temperature	Time	Cycle(s)	Function
95°C	10 min	1	activation of Ampli Taq Gold®
95°C	15 s	45	denaturation step
60°C	1 min		annealing, elongation, nucleolytic cleavage of hybridised probes
4°C	∞	1	storage

Table 2-8 Universal TaqMan® PCR protocol. If a mastermix with AmpErase UNG was used (prevents contamination), a 2 min step at 50°C was added in the beginning.

The endpoint fluorescence measurement was carried out with the ABI Prism® 7900 HT Sequence Detection System. If 96 well MTPs were used, they were read in the ABI Prism® 7700 Sequence Detection System. Allele calling for each plate was done manually to ensure data quality. Markers passing a call rate of 95%, and not deviating from Hardy-Weinberg equilibrium, were used in downstream analyses.

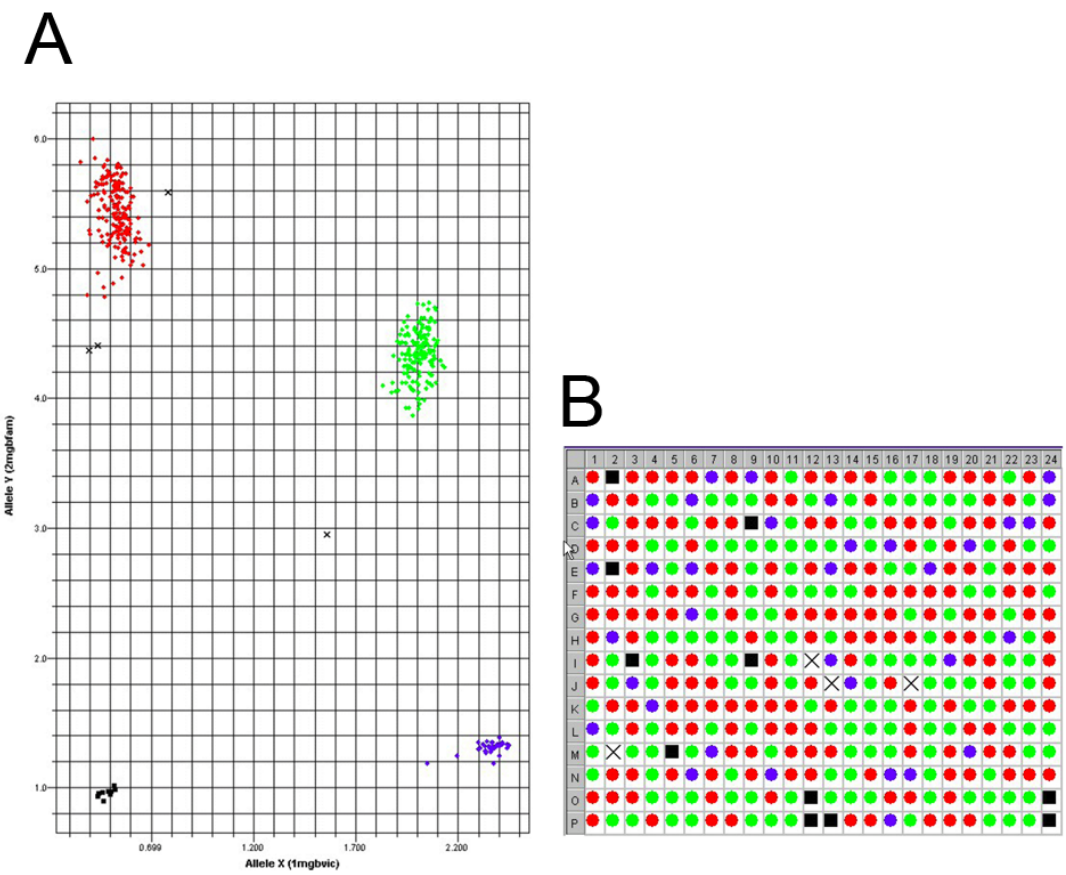

Fig. 2-7 TaqMan® cartesian cluster plot of 384 samples including controls. (A) Normalized VIC®-fluorescence is plotted along the x-axis, while FAM™ is plotted along the y-axis. All three clouds seperate nicely and were manually called with the SDS software, i.e. manual allele assignment. (B) Assignment of genotypes to the positions on the plate. Homoyzygotes for allele 1 are shown in red, heterozygotes in green (both dyes are measured), and homozygotes for allele 2 in blue. The black squares close to the origin are the negative controls, which control for potential contamination, and black crosses are undefined genotypes that were excluded from subsequent analyses.

Given the level of automation, a throughput of 22,000 genotypes per day can be achieved at the Kiel SNP genotyping platform. This corresponds to fifty-seven 384 well MTPs or 20,976 different DNA samples. At the end of the process, genotyping data is imported into the SQL database 'ibdbase'.

2.8.2 SNPlex™

As hierarchical genome-wide scans demand a follow-up of hundreds and more significant SNPs, a high-throughput genotyping platform using SNPlex™ was established which is now also used throughout Germany's National Genome Research Network (NGFN). The current extent of automation yields a throughput of 280,000 genotypes per day, which corresponds to fifteen 384 MTPs that are run with 48plex pools. In contrast to the simple and robust TaqMan® method, multiple pipetting steps are needed for SNPlex™ and the entire process stretches over three days. To avoid contaminations, as universal primers are used to amplify the purified ligation product, pre-PCR steps are physically seperated from post-PCR reactions, i.e. they are carried out in two distinct laboratories. The principle of the oligonucleotide ligation reaction (OLA), which is the allele-discriminating step, is based upon a method developed around the 1990's: the ligation chain reaction (LCR; Barany *et al.*, 1991; Landegren *et al.*, 1988; Weiss *et al.*, 1991). In contrast to LCR, SNPlex™ uses a normal PCR step to exponentially amplify the ligation products. Furthermore, oligonucleotides are designed and synthesized for only one DNA strand by Applied Biosystems. Hence, SNPlex™ takes advantage of ligation's specificity and PCR's increase in sensitivity.

Assays for the SNPlex™ Genotyping System are designed by Applied Biosystem's automated high-throughput pipeline. The pipeline combines SNP-specific assays into compatible multiplex pools. These steps include (according to De La Vega *et al.*, 2005):

- Screening the SNP context sequences against the target genome to avoid designing assays for SNPs in repetitive or duplicated genomic regions thatwould lead to lowspecificity (this step can be omitted for organisms without an assembled genome at hand).
- Selection and design of the SNPspecific ligation probes by applying assay and probemanufacturing rules to select the more suitable strand and probe sequence.
- Assignment of ZipCode sequences to each ASO probe of an assay.
- Separating the assays into compatible multiplex pools that are screened for probe/probe interactions, spurious ligation templates, and unintended probe combinations that may have a significant genomic target.

Each assay includes three SNP-specific ligation probes: Two of the probes are allele-specific oligos (ASOs). These are designed specifically for the detection of polymorphisms by having the discriminating nucleotide on the 3' end. Each ASO probe sequence also contains one of 96 unique ZipCode™ sequences for ZipChute™ probe binding. In a multiplex reaction, the universal ZipCode sequences on each ASO are unique. Therefore, in a 48-plex reaction, there are 96 ASOs (two for each SNP), and 96 different ZipCode sequences. 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. Each LSO also contains a partial universal PCR-primer binding site. In a 48-plex reaction, there are 48 LSOs. All 144 probes for a 48-plex reaction are shipped together as an ASO/LSO probe pool. It is this pool that confers genotyping specificity to the SNPlex™ System assay. All other reagents are universal and not SNP specific.

The assay workflow is outlined in fig. 2-9 and in brief the procedure is as follows:

Day 1 - OLA laboratory

SNPlex™ is based on the oligonucleotide ligation/PCR assay (OLA/PCR). After an initial kinase step to phosphorylate linkers and ligation probes, the activated oligonucleotides are combined with fragmented whole-genome amplified DNA (100–150 ng per well, i.e. 2–3 ng per assay) to perform genotyping in separate reactions. Only 384 well MTPs are used throughout the process. During the "OLA reaction", which is the allele-discriminating step, the genotype information is encoded by highly specific ligation of the ASO probes to the LSO probes using genomic DNA as the target. ASO and LSO linkers connect to the corresponding ASO and LSO probes.

Day 2 - OLA laboratory

After the ligation reaction, unligated and incompletely ligated oligonucleotides, as well as the genomic DNA templates, are removed by an enzymatic digestion using exonuclease I and λ-exonuclease ("purification step"). This reduces the background noise of the signal. Following dilution of the digested material, an aliquot is subjected to a PCR reaction with two universal primers, one of which is biotinylated ("PCR setup"). The plates with the PCR mix are then put into the thermocyclers of the PCR laboratory.

Day 3 - PCR laboratory and capillary electrophoresis (CE)

After PCR, biotinylated amplicons are incubated with streptavidin-coated microtiter plates. Exponential amplification of the ligation products yields higher signals and a better discrimination. Upon removal of the non-biotinylated amplicon strands, a mixture of 102 pre-optimized, universal ZipChute™ probes is added to each well for hybridization and to decode the genotypic information. Of these, 96 ZipChute™ probes correspond to all 96 possible alleles of the 48 addressable SNPs in the multiplex. The six remaining ZipChutes™ are needed for internal controls, such as the positive and the negative hybridization control (PHC/NHC). ZipChute™ probes are fluorescently labeled oligonucleotides, with each probe having a unique size (so-called mobiliy modifiers). The ZipChute™ probes are eluted after stringent washing and detected by electrophoretic separation on Applied Biosystems 3730xl DNA Analyzers. An allelic ladder containing all available ZipChute™ probes is also analyzed to correct run-to-run sizing variations.

The study manager of the GeneMapper® software was used for automated allele calling of all plates. Auto-calls were manually inspected for faulty genotype assignments before the data was exported from GeneMapper® and then imported into the in-house database 'ibdbase'. Additional software tools, e.g. the OLAtool, and database tables were created to meet the requirements of SNPlex™.

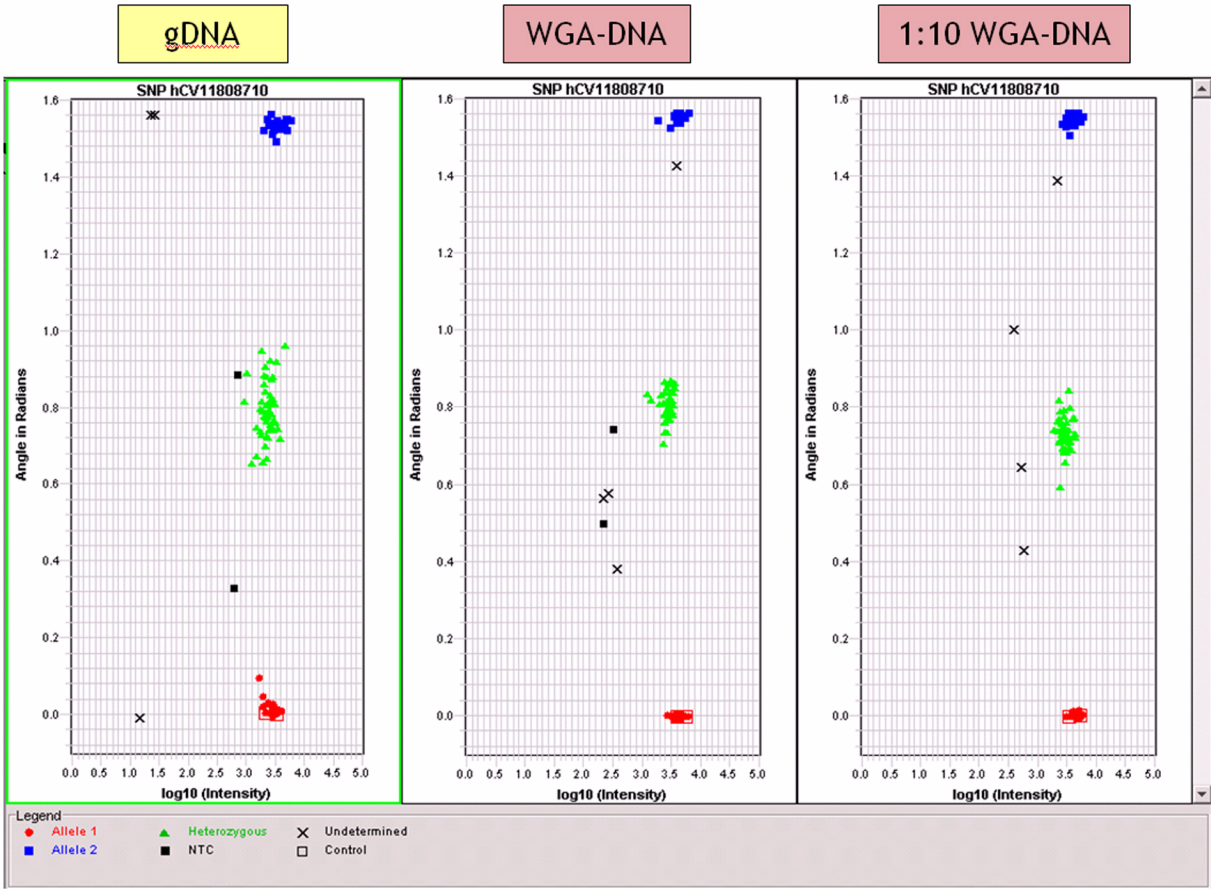

Fig. 2-8 SNPlex™ polar cluster plots. The same SNP is shown for three different types of DNA (genomic, amplified, and diluted amplified DNA). There are no apparent quality differences concerning clustering. The legend below the plots shows the meaning of each symbol. Along the x-axis, the intensity is plotted on a log-scale. Data points above 3.0 are considered to have a good signal intensity and quality. The angle in radians (angle of the vector from the origin in the cartesian plot) is plotted along the y-axis.

All necessary master mixes were prepared manually, while all other pipetting steps were carried out on four different TECAN multipipetting robots.

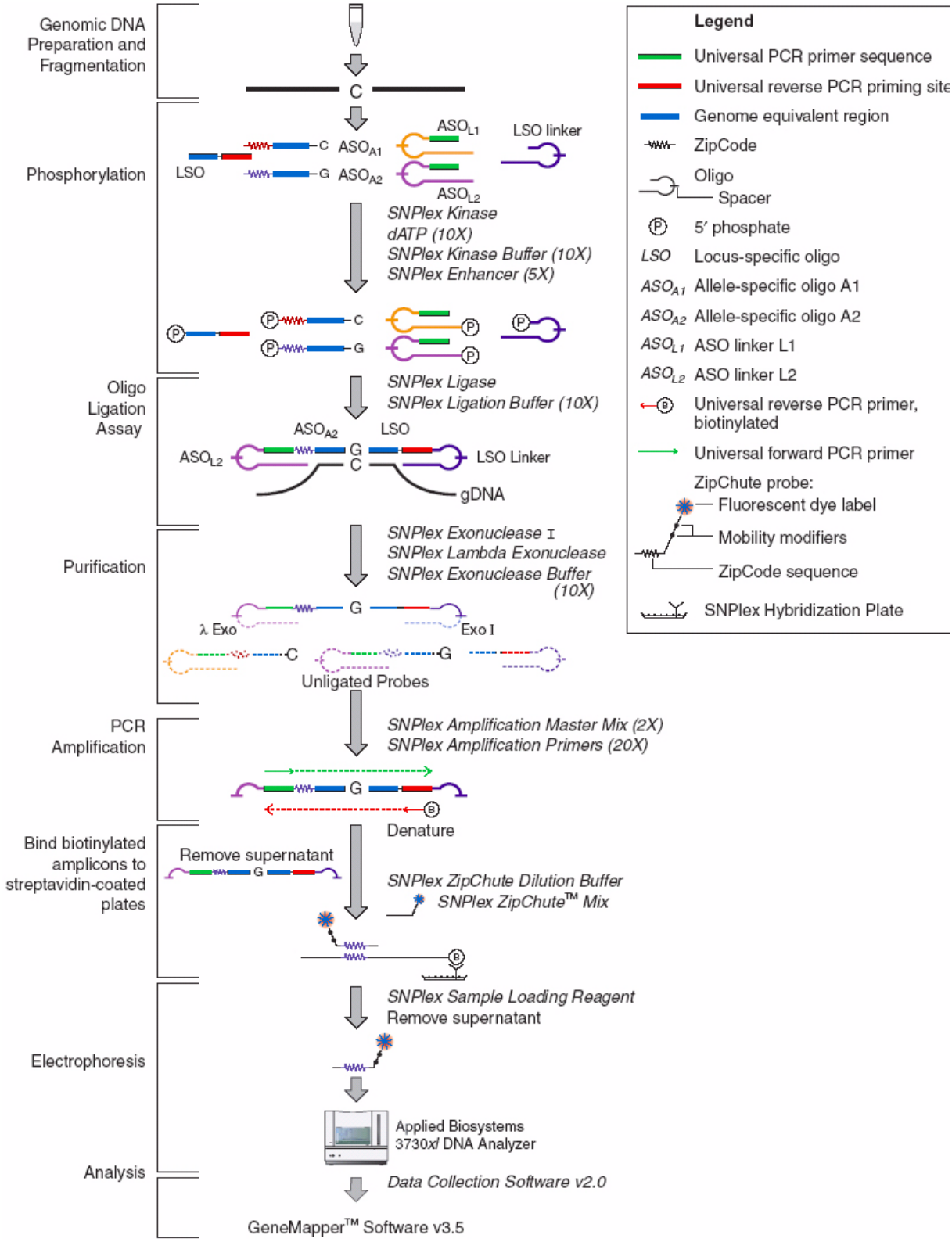

Fig. 2-9 SNPlex™ workflow and chemistry. The key step is the oligo ligation assay, which is the allele-discrimnating step. For a description see text above and for more details see the protocol from Applied Biosystems.

2.8.3 SNP selection

For the first follow-up round of the genome-wide screenings, lead SNPs were selected that were below a certain p-value threshold. Further steps, which included the fine mapping of *ATG16L1* and *NELL1*, required a more thorough SNP selection. Only validated HapMap SNPs were chosen by means of Haploview that passed the following quality criteria: call rate >95%, p-value for HWE >0.01, less than 3 Mendel errors, minor allele frequency >1%. Tagging SNPs were chosen as described in 2.9.5.1 on page 69.

If more SNPs were required to fill in gaps or if additional cSNPs were needed, validated SNPs from the NCBI dbSNP database were selected according to the descriptions of Fredman *et al.* (2006). In their publication, Fredman *et al.* state that the best validation criteria is that the dbSNP entry has more than two submitter handles ("rationalized ssID count" >2). Therefore, the cSNP selection of the present study is biased towards records with many submitters.

Finally, if there were still gaps, SNPs from the Celera Genome Project were chosen with the help of Applied Biosystem's SNPbrowser™ software (De La Vega *et al.*, 2005).

2.8.4 Design of coding SNPlex™ pools by Applied Biosystems

Full detail regarding the construction of the panel of 19,779 nsSNPs is provided in 9.2 on page 176. In brief, SNPs from the dbSNP database build 117 (Sherry *et al.*, 2001) were combined with variants from the Applera exon resequencing project (Adams *et al.*, 2002) together with the SNPs discovered by shotgun sequencing of the human genome by Celera Genomics (Venter *et al.*, 2001). Variants that mapped uniquely to the human genome assembly (Celera R27) were then further filtered based on their measured or expected (i.e. "double-hit" SNPs) heterozygosity in populations of European and African descent. Putative functional SNPs were then defined as non-synonymous variants that altered the amino-acid sequence of an annotated NCBI RefSeq, Celera or ENSEMBL transcript. For the resulting 28,709 nsSNPs, context sequence and allele information was submitted to the assay design pipeline of the SNPlex™ Genotyping System v. 1.0 (Tobler *et al.*, 2005). Context sequence was masked for adjacent double-hit SNPs to avoid probes overlapping with other common SNPs. Finally a total of 19,779 SNPlex™ assay designs that were manufactured and partitioned in 428 multiplex pools of up to 48 SNPs each (mean: 45 SNPs per assay pool). The full list of SNPs and their annotations are provided in `Supplemental_Table_01.xls` on the DVD.

2.8.5 Affymetrix arrays

Parallel genotyping of thousands of SNPs was first reported by Carrasquillo *et al.* (2002), Fan *et al.* (2000) and Wang *et al.* (1998). The 100k GeneChip® is a further development of the 10k GeneChip (Matsuzaki *et al.*, 2004) by Affymetrix. Both combine reduction in genome complexity with the allele-discriminating specificity of oligonucleotide arrays (Kennedy *et al.*, 2003). Using Platinum Pfx polymerase (Invitrogen) instead of normal Taq polymerase results in the preferential amplification of fragments in the size range between ~250 bp and ~2000 bp. This represents ~300 megabases (Mb) of sequence complexity, compared to the five-fold lower complexity generated with Taq polymerase. This correspondingly increases the number of SNPs in the hybridization targets (Matsuzaki *et al.*, 2004).

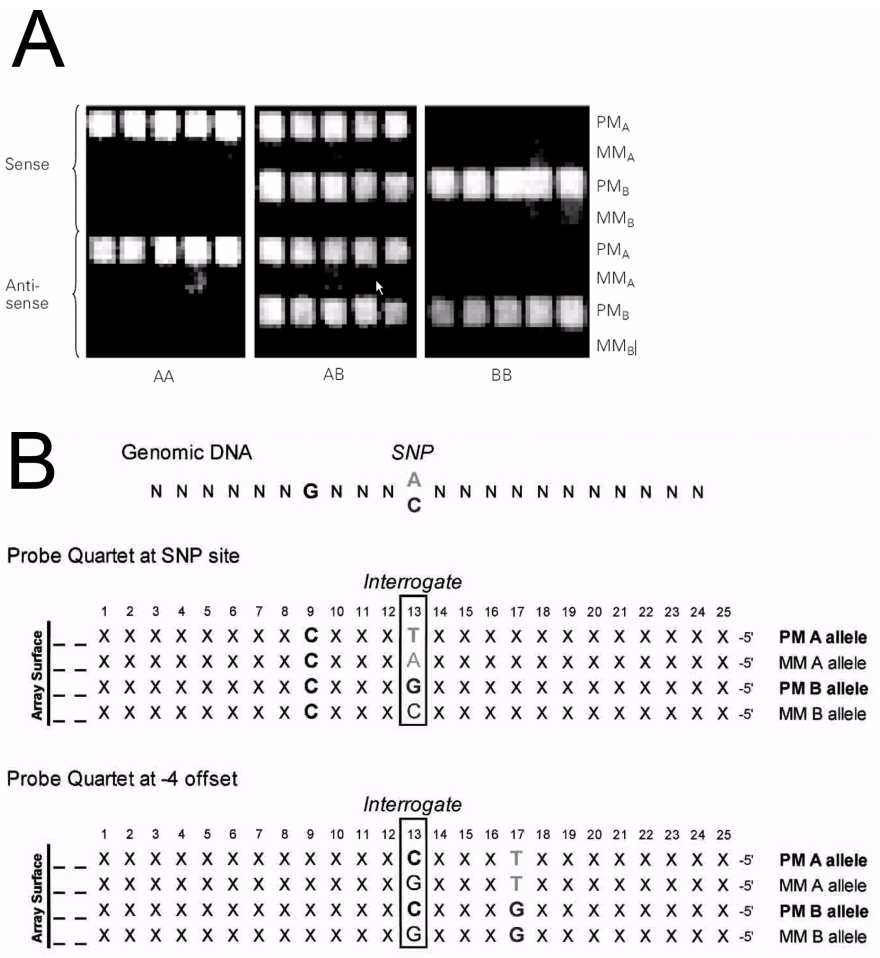

Fig. 2-10 Example of an Affymetrix GeneChip miniblock. (A) SNP miniblock showing hybridization in three individuals, demonstrating the three possible genotypes "AA", "AB", and "BB". Probes are synthesized as perfect-match (PM) 25-mers and as one base mismatches in the center (MM). Probes for both "A" and "B" alleles on both sense and antisense strands are synthesized, for a total of 40 probes per SNP miniblock(B) Sequence prototypes of the oligonucleotide probes. Twenty-five-mer oligonucleotides which are complementary to SNP sites and flanking sequences are synthesized on the surface of the array. The 13th nucleotide is the interrogative position where the probe sequences are either perfectly matched (PM) or mismatched (MM) to one of the two alleles of the SNP. The PM and MM probe pairs provide a basis for signal vs. noise measurements. The two probe pairs corresponding to the two alleles are grouped as probe quartets. Shown are the prototype sequences of the probe quartet at the SNP site, where the probe sequences differ only at the SNP site which is also the interrogative position. To provide data redundancy, four additional probe quartets are offset from the SNP site by one to four nucleotides in either direction. Also shown are prototype sequences for the probe quartet offset by4. In this offset probe quartet, the SNP site has shifted to position 17 of the 25-mer. The probe sequences in this quartet are different at 4 (PM vs. MM) and at the SNP site (allele "A" vs. "B"). Each SNP is represented by five probe quartets (one at the SNP site and four offset) in both orientations, for a total of 40 oligonucleotide probes.

Each array in the Mapping 100K Set includes more than 2.5 million features, the latter consisting of more than one million copies of a 25 bp oligonucleotide probe of a defined sequence, synthesized in parallel by proven photolithographic manufacturing. Each SNP is interrogated by 10 probe quartets see (fig 2-10, page 49, A) where each probe quartet is comprised of a Perfect Match and a Mismatch probe for each allele. In total, there are 40 different 25 bp oligonucleotides per SNP (see fig 2-10, page 49, B).

Affymetrix arrays were hybridized in collaboration with Peter Nürnberg and Christian Becker (Cologne Center for Genomics).

In total, 116,161 SNPs were present on the Affymetrix 100k GeneChip® that actually consists of two 50k chips. Average intermarker distance was 23.6 kb. Genotyping of 393 cases and 399 controls was done on the Affymetrix GeneChip® Human Mapping 50K XbaI and HindIII Arrays. About 250 ng of genomic DNA were digested with the two restriction enzymes XbaI and HindIII and processed according to the Affymetrix protocol. In brief, the following steps were carried out:

Genome complexity reduction

Sample DNAs should not be highly degraded nor contain PCR inhibitors, such as high concentrations of heme or chelating agents. For each individual assayed, 250 ng of genomic DNA are digested separately with 10 U of XbaI or HindIII (New England BioLabs) in volumes of 20 µl for 2 hours at 37°C. Following heat inactivation at 70°C for 20 minutes, 0.25 µM of XbaI adaptor (5'-ATT ATG AGC ACG ACA GAC GCC TGA TCT-3' and 5'phosphate –CTA GAG ATC AGG CGT CTG TCG TGC TCA TAA- 3')(Affymetrix), or HindIII adaptor (5'-ATT ATG AGC ACG ACA GAC GCC TGA TCT-3' and 5'phosphate –AGC TAG ATC AGG CGT CTG TCG TGC TCA TAA-3') (Affymetrix) are ligated to the digested DNAs with T4 DNA Ligase (New England BioLabs) in 25 µl for 2 hours at 16°C. The ligations are stopped by heating to 70°C for 20 minutes, and then diluted 4-fold with water. For each ligation reaction, two to three PCRs are run in order to generate >40 µg of PCR products. Each PCR contains 10 µl of the diluted ligation reactions (25 ng of starting DNA) in 100 µl volumes containing 1.0 µM of primer (5'-ATT ATG AGC ACG ACA GAC GCC TGA TCT-3'), 0.30 mM dNTPs, 1.0 mM $MgSO_4$, 5 U Platinum® Pfx Polymerase (Invitrogen), PCR Enhancer (Invitrogen) and Pfx Amplification Buffer (Invitrogen). 30 cycles of PCRs are run with the following cycling program: 94°C denaturation for 15 seconds, 60°C annealing for 30 seconds, and 68°C extension for 60 seconds. As a check, 3 µl of PCR products are visualized on 2% TBE agarose gels to confirm the size range of amplicons.

The PCR products are purified over MinElute 96 UF PCR Purification plates (Qiagen), and recovered in 40 µl of EB buffer (Qiagen). PCR yields are measured by absorbance readings at

260 nm, and adjusted to a concentration of 40 µg per 45 µl. To allow efficient hybridization to 25-mer oligonucleotide probes, the PCR products are fragmented to <100 bp with DNAse I. 0.20 U of DNAse I (Affymetrix) is added to 40 µg of purified PCR amplicons in a 55 µl volume containing Fragmentation Buffer (Affymetrix) for 35 minutes at 37°C, followed by heat inactivation at 95°C for 15 minutes. Fragmentation products are visualized on 4% TBE agarose gels.

The 3' ends of the fragmented amplicons are biotinlyated by adding 214 µM of a proprietary DNA labeling reagent (Affymetrix) using Terminal Deoxynucleotidyl Transferase (Affymetrix) in 70 µl volumes for 2 hours at 37°C, followed by heat inactivation at 95°C for 15 minutes.

Allele specific hybridization to oligonucleotide arrays

The fragmented and biotinylated PCR amplicons are combined with 11.5 µg/mL human Cot-1 (Invitrogen) and 115 µg/mL herring sperm (Promega) DNAs. The DNAs are added to a hybridization solution containing 2.69 M tetramethylamonium chloride (TMACl), 5.77 mM EDTA, 56 mM MES, 5% DMSO, 2.5x Denhardt's solution, and 0.0115% Tween-20 in a final volume of 260 µl. The hybridization solution was heated to 95°C for 10 minutes then placed on ice. After warming to 48°C for 2 minutes, 200 µl of the hybridization solution is injected into cartridges housing the oligonucleotide arrays (Affymetrix GeneChip® 100K Mapping Set: 50K Array Xba 240 and 50K Array Hind 240). Hybridizations are carried out at 48°C for 16 to 18 hours in a rotisserie rotating at 60 rpm. Following the overnight hybridization, the arrays are washed with 6x SSPE and 0.01% Tween-20 at 25°C, then more stringently washed with 0.6x SSPE and 0.01% Tween-20 at 45°C. Hybridization signals are generated in a three step signal amplification process: 10 µg/mL streptavidin R-phycoerythrin (SAPE) conjugate (Molecular Probes) is added to the biotinylated targets hybridized to the oligonucleotide probes, and washed with 6x SSPE and 0.01% Tween-20 at 25°C; followed by the addition of 5µg/mL biotinylated goat anti-streptavidin (Vector) to increase the effective number of biotin molecules on the target; and finally SAPE is added once again and washed extensively with 6x SSPE and 0.01% Tween-20 at 30°C. The SAPE and antibody were added to arrays in 6x SSPE, 1x Denhardt's solution and 0.01% Tween-20 at 25°C for 10 minutes each. Following the final wash, the arrays are kept in Holding buffer (100 mM MES, 1 M [Na^+], 0.01% Tween-20). The washing and staining procedures are run on Affymetrix fluidics stations. Arrays are scanned using GCS3000 scanners with AutoLoaders (Affymetrix).

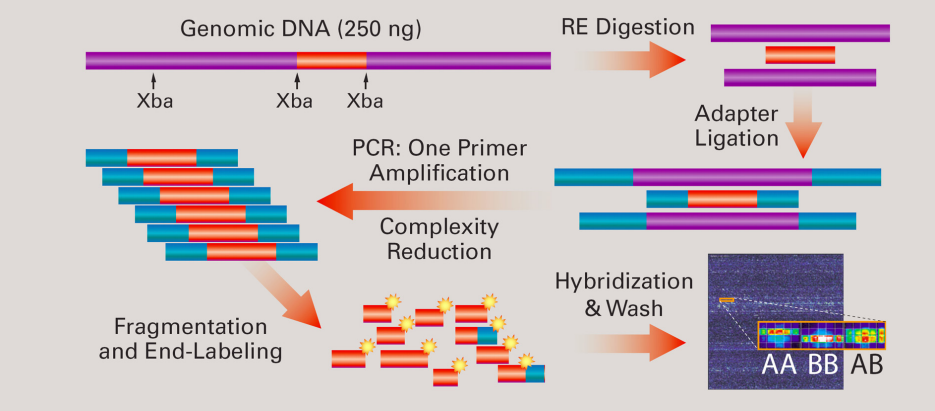

Fig. 2-11 GeneChip® mapping assay overview. For simplification, only the process for one restriction enzyme is shown.

Genotypes were called by the GeneChip® DNA Analysis Software (GDAS v2.0, Affymetrix) using the implemented dynamic model (DM; Di *et al.*, 2005) algorithm and a confidence score cutoff of 0.25. The Dynamic Model Mapping algorithm is a likelihood model based algorithm using Wilcoxon's signed rank test. It provides a genotype call for each SNP, along with quality information for the call (for more details see GDAS User's guide).

To ensure that no samples were confused, the 31 identical SNPs placed on both chips were checked whether they yielded the same genotype for the same individual.

We verified genders by counting heterozygous SNPs on the X chromosome. Experimental details concerning the genotyping of the 100K SNP set are provided in Matsuzaki *et al.* (2004).

Genotype data of all 50k SNPs was exported for 96 chips (corresponding to 96 different DNAs/individuals) into a single file per chip. These large and inscrutable files were converted with the `Converter.jar` of GENOMIZER (2.9.3 on page 63) into the computational efficient "chromosome data files".

2.8.6 Quality control

A genotyping error occurs when the observed genotype of an individual does not correspond to the true genotype (Bonin *et al.*, 2004). Erroneous genotypes markedly decrease the power for detecting associations (Goring *et al.*, 2000; Abecasis *et al.*, 2001; Terwilliger *et al.*, 1990; Gordon *et al.*, 1999). Several levels of checks were applied to the data as suggested by Pompanon *et al.*, (2005) to ensure reliability of the results.

Human errors

Human subjectivity during manual scoring (TaqMan® and SNPlex™) represented a main problem that can hardly be avoided. Obviously, the risk of human scoring error strongly depends on the quality of the data. Other sources of human errors were minimized by a maximum of automization. This included the use of barcodes and scanners for plates and samples.

Low quantity or quality of DNA

Only a few or low-quality target DNA molecules favour allelic dropouts and false alleles (Taberlet *et al.*, 1996). Furthermore, the risk of contamination is increased. Therefore, each DNA sample was quality checked on a gel (2.6 on page 33), normalized to a specific concentration (2.4 on page 29), and an excess amount was used for genotyping and the whole-genome amplification (2.5 on page 31).

Call rate and Hardy-Weinberg equilibrium

All assays were checked for a high call rate (cases and controls) and for HWE (only controls) as described in 2.9 on page 55 and by Hosking *et al.* (2004) and Xu *et al.* (2002). HW disequilbrium can also arise from selection, inbreeding, or population admixture (Morton *et al.*, 1995).

Checking for Mendelian inheritance

In family-based studies, genotypes were checked for Mendelian errors, i.e. the genotype of a child is incompatible with the genotypic constitution of the parents, e.g. the child has the genotype “AA”, but the parents have the genotype “AA” and “BB” (Douglas *et al.*, 2002; Ewen *et al.*, 2000). For most of the Mendelian errors, non-paternity was the cause, as always the same pedigrees were affected. Since the assumed rate of non-paternity is <10% in Germany (according to the Max Planck Institute in Munich), assays that produced more than 10% Mendelian Error, i.e. more than one error per hundred trios, were excluded. This was done because Mendelian errors can also indicate a genotyping or contamination problem (Gordon *et al.*, 2004; Lange *et al.*, 2001). If the number of Mendelian errors was below this threshold or if the error-causing genotype(s) had ambiguous positions in the cluster plot, the genotypes of all pedigree members were set to “0”.

Positive/negative Controls

Another quality control for the assays and the genotyping process was the inclusion of positive and negative controls as described in 2.3.2 on page 28. In addition, each genotyping method had its own internal controls . For a description, see the corresponding chapters (TaqMan®: 2.8.1 on page 39; SNPlex™: 2.8.2 on page 44; Affymetrix Arrays: 2.8.5 on page 49).

Technical replication

Due to the paricular biochemical nature of each genotyping method, a technical replication with a different method was performed before trusting new significant associations (see e.g. page 104).

2.9 Association analyses

Association studies can be family- or population-based resulting in two different analysis methods, the transmission disequilibrium test (TDT) for family-based studies (Spielman *et al.*, 1993) and the case-control analysis for population-based studies. Genetic analyses were performed using the diagnostic disease categories IBD, CD, and UC as described in 1.1.1 on page 2. All tested SNPs had to pass several generally employed quality tests, such as call rate, being polymorphic, and being in Hardy-Weinberg equilibrium. The latter is a fundamental principle in population genetics stating that the genotype frequencies and gene frequencies of a large, randomly mating population remain constant provided immigration, mutation, and selection do not take place (*American Heritage Dictionary*).

All markers were tested for Hardy-Weinberg equilibrium (HWE) in controls using a $\chi 2$ test before inclusion in the association statistics ($p > 0.01$ threshold; variables as in table 2-14, page 62 and with p_1 and p_2 being the allele frequencies):

$$\chi 2 = a + b + c;\ a = \frac{\left(n_{11} - \frac{p_1 \cdot p_1}{z}\right)^2}{\frac{p_1 \cdot p_1}{z}};\ b = \frac{\left(n_{12} - \frac{p_1 \cdot p_2}{z}\right)^2}{\frac{p_1 \cdot p_2}{z}};\ c = \frac{\left(n_{22} - \frac{p_2 \cdot p_2}{z}\right)^2}{\frac{p_2 \cdot p_2}{z}}$$

If SNPs with significant deviations from HWE would be included, unacceptable high type I error (false positives) rates would be the case (Sasieni *et al.*, 1997; Schaid *et al.*, 1999). Only biallelic markers passing a call rate (CR) of >95% in cases and controls were used in downstream analyses. Quality assessments, single locus and transmission disequilibrium tests (TDT) were performed using Haploview (Barrett *et al.*, 2005) and GENOMIZER (Franke *et al.*, 2006; for more details see 2.9.3 on page 63). In families with multiple affected individuals, one trio was randomly extracted for TDT analysis as implemented by Haploview. For case-control analyses, one affected child was randomly extracted per trio using an in-house available Perl script `create_case_controls_random.pl`. Significance assessment of associations with or between single locus genotypes was performed using $\chi 2$ or Fisher's exact test for 2 x 3 contingency tables.

Haplotype frequency estimates among singletons were obtained using an implementation of the EM algorithm in COCAPHASE (Dudbridge *et al.*, 2003). Significance testing of haplotype frequency differences were also performed with COCAPHASE and TDTPHASE. For details see 2.9.5 on page 67.

Standardized coefficients r^2 and D' were computed with Haploview for pair-wise linkage disequilibrium (LD) estimation between markers (2.9.1 on page 56).

2.9.1 Linkage disequilibrium (LD)

The population genetics term linkage disequilibrium is used for the non-random association of alleles at two or more loci (Goldstein *et al.*, 2001), not necessarily on the same chromosome.

Example according to Ott (1999), see also fig. 2-12:

A non-polymorphic gene (wt/wt) is in close proximity to a biallelic marker on the same chromosome. The SNP has the alleles "A" and "B", where "A" is the more frequent allele. At this stage, the haplotypes "wtA" and "wtB" exist in the population. Then a mutation occurs, "wt → mut", and this is statistically more likely to take place on the chromosome with the allele having the higher population frequency, which is in this case allele "A". At this point, the population contains three haplotypes, the two common ones considered above and a new one, "mut/A". As the population grows, recombinations occur between the two loci and slowly a fourth haplotype, "mut/B", emerges. There is then a marked difference between two classes of haplotypes, those carrying a "wt" allele and those carrying a "mut" allele. The former contains marker alleles "A" and "B" in approximately equal frequencies. However, haplotypes carrying a mut allele contain far more "A" than "B" alleles (originally only "A"). This situation is referred to as linkage disequilibrium. If "mut" is a disease allele, there must be a marked difference of marker allele frequencies between affected and unaffected individuals. The extent of LD varies largely, there are "hot" and "cold" spots of recombination in the genome.

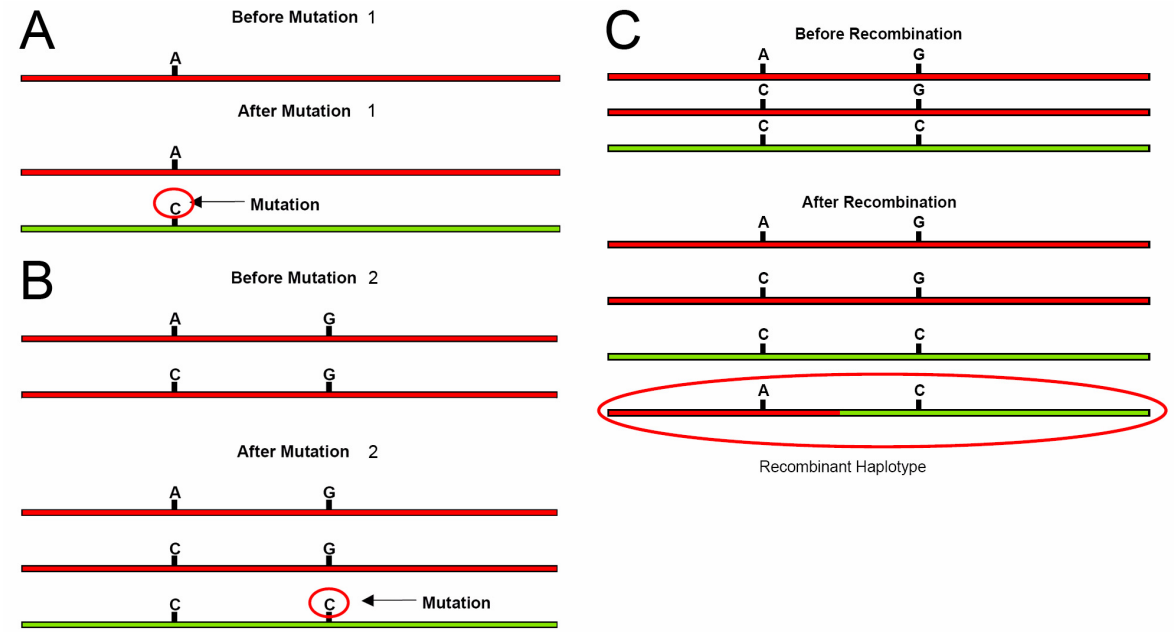

Fig. 2-12 History of two neighbouring alleles. (A) Alleles that exist today arose through ancient mutations. (B) One allele arose first and then the other. The "C–C" haplotype will either die out or "drift" to an appreciable frequency in the population, with or without selective pressure. (C) Recombination creates new arrangements for ancestral alleles. *Illustration from* http://www.sph.umich.edu/csg/abecasis/class/

Assuming two loci “A” and “B” with alleles “A”, “a” and “B”, “b” respectively, the allele frequencies are p_A, p_a, p_B, and p_b with $p_A + p_a = 1$ and $p_B + p_b = 1$. The following haplotypic configurations are possible:

Haplotype	Independence	Observed
A_____B	$p_A \cdot p_B$	p_{AB}
A_____b	$p_A \cdot p_b$	p_{Ab}
a_____B	$p_a \cdot p_B$	p_{aB}
a_____b	$p_a \cdot p_b$	p_{ab}

Table 2-9 Haplotypic configurations. “p” denotes the frequency of the corresponding genotype.

In the presence of linkage equilibrium, the expected haplotype frequencies are equal to the product of the allele frequencies. The deviation from the exptected frequency is determined by calculating the disequilibrium parameter D:

$$p_{AB} = p_A \cdot p_B + D \rightarrow D = p_{AB} - (p_A \cdot p_B).$$

The equilibrium state is characterized by D = 0 and is called gametic phase equilibrium or linkage disequilibrium, and values >0 are referred to as positive disequilibrium or association. For all four possible haplotypes, deviations are the same in case of biallelic loci. To differentiate the LD between the locus pairs, the measure D is normalized. The most commonly used normalizations are D’ and r^2:

$$D' = D/D_{max} \text{ for } D > 0 \text{ and } D' = D/D_{min} \text{ for } D < 0$$

$$\text{with } D_{max} = \min(p_A \cdot p_b, p_a \cdot p_B) \text{ and } D_{min} = \max(-p_A \cdot p_B, -p_a \cdot p_b) \text{ [Hartl } \textit{et al.}\text{, 1988]}$$

$$r^2 = \Delta^2 = D^2/(p_A \cdot p_B \cdot p_a \cdot p_b)$$

A D’ value of one means that there is no evidence for recombination between the two markers. If allele frequencies are similar, high D’ values mean that the markers are good surrogates for each other.

As recombination occurs between two loci, an existing LD between them gradually breaks down (decay of disequilibrium). If one has determined a value of D’ between disease and a marker locus and is willing to assume that the marker has not changed in past generations, it is of interest to estimate the number of generations t from the given value of D’:

$$t = \log(D')/\log(1-r), \text{ where r is the recombination fraction between the two loci.}$$

In the present study, pairwise LD between SNPs was computed using Lewontin’s standardized deviation coefficient D’ (Lewontin *et al.*, 1963 and Lewontin *et al.* 1988) and r^2 (Hill *et al.*, 1968). The algorithm, as implemented in Haploview (Barret *et al.*, 2005), compares estimated two-marker haplotypes with expected haplotypes in each population.

2.9.2 Case-control single-point analyses

The single-marker case-control statistic served as the main analysis method, especially for the initial genome-wide screenings. Summarizing the method, allele and genotype frequencies are compared between unrelated healthy individuals (controls) and unrelated patients (cases). In contrast to this study design, a family-based approach was used, the TDT, which is described in 2.9.4 on page 65. The main advantage of the case-control approach is the less difficult and faster recruitment process and the higher statistical efficiency (TDT: only 1/6 on the null hypothesis; Morton *et al.*, 1998) and power. A major disadvantage of the case-control design is its susceptibility to population substructure. Therefore, matching cases and controls needs careful consideration to avoid stratification. Three distinct methods were used for calculating case-control statistics, which are described in the following sections.

2.9.2.1 Genotype-based case-control comparison (CCG)

Genotype counts were obtained by cluster plot analyses of TaqMan® (2.8.1 on page 39) and SNPlex™ (2.8.2 on page 44) data. Observed genotypes frequencies were compared to the expected frequencies, under the null hypothesis (H_0) that no differences exist between cases and controls.

Genotype	Cases	Controls	Sum
11	n_{11}	m_{11}	w
12	n_{12}	m_{12}	x
22	n_{22}	m_{22}	y
Sum	n	m	z

Table 2-10 Two-by-three contingeny table for genotype-based analyses.

If homozygotes for one allele are missing, then the 2 x 3 contingency table is narrowed down to a 2 x 2 table and the χ^2 value is calculated according to 2.9.2.2 on page 59.

The genotypic Pearson's χ^2 test is calculated according to Ott *et al.* (1985) and table 2-10:

1. Expected values are computed:

$$a = \frac{n \cdot w}{z}; \quad b = \frac{m \cdot w}{z}; \quad c = \frac{n \cdot x}{z}; \quad d = \frac{m \cdot x}{z}; \quad e = \frac{n \cdot y}{z}; \quad f = \frac{m \cdot y}{z}$$

2. Differences to the observed values are calculated:

$$g = \frac{(n_{11} - a)^2}{a}; \quad h = \frac{(m_{11} - b)^2}{b}; \quad i = \frac{(n_{12} - c)^2}{c}; \quad j = \frac{(m_{12} - d)^2}{d}; \quad k = \frac{(n_{22} - e)^2}{e}; \quad l = \frac{(m_{22} - f)^2}{f}$$

3. Chi square is finally computed:

$$\chi^2 = g + h + i + j + k + l$$

A chi square value greater than 3.84 is considered significant at the $\alpha = 0.05$ level and with one degree of freedom.

GENOMIZER was used to calculate the genotypic case-control p-value. If sex-matching was required, an in-house Perl script (`snp_cc_significance.pl`) was used.

2.9.2.2 Allele-based case-control comparison (CCA)

Besides the genotypic case-control test, the allele-based test was used as the second major test for association. Using the values from table 2-10, the following formulas were applied to construct table 2-11:

$$n_1 = 2 \cdot n_{11} + n_{12};\ m_1 = 2 \cdot m_{11} + m_{12};\ n_2 = 2 \cdot n_{22} + n_{12};\ m_2 = 2 \cdot m_{22} + m_{12}$$

Allele	Cases	Controls	Sum
1	n_1	m_1	x
2	n_2	m_2	y
Sum	n	m	z

Table 2-11 Two-by-two contingeny table for allele-based analyses.

The allelic Pearson's χ^2 test is calculated according to Ott *et al.* (1985) and table 2-11:

1. Expected values are computed:

$$a = \frac{n \cdot x}{z}; \qquad b = \frac{m \cdot x}{z}; \qquad c = \frac{n \cdot y}{z}; \qquad d = \frac{m \cdot y}{z}$$

2. Differences to the observed values are calculated:

$$e = \frac{(n_1 - a)^2}{a}; \qquad f = \frac{(m_1 - b)^2}{b}; \qquad g = \frac{(n_2 - c)^2}{c}; \qquad h = \frac{(m_2 - d)^2}{d}$$

3. Chi square is finally computed:

$$\chi^2 = e + f + g + h$$

A chi square value greater than 3.84 is considered significant at the $\alpha = 0.05$ level and with one degree of freedom. The general rule to calculate the degrees of freedom is to multiply the degrees of freedom for the rows and columns [i.e., d.f. = (rows-1) · (columns-1)] .

GENOMIZER (Franke *et al.*, 2006) and Haploview (Barret *et al.*, 2005) were used to calculate the allelic case-control p-value in the present study. If sex-matching was required, an in-house Perl script (`snp_cc_significance.pl`) was used. Permutation testing, as implemented in Haploview, was used to verify significant p-values. This means that the affection status of the individuals is reassigned and the same calculation is repeated, for the current study 10,000 times. After bootstrapping, the output summarizes how many of the permuted data sets had a higher maximum value compared to the non-randomized dataset..

2.9.2.3 Odds ratio (OR)

The odds ratio is often used in retrospective studies as an approximation to the relative risk (for pospective or cohort studies) and is referred to as the approximate relative risk or simply as relative risk (Armitage *et al.*, 1987). The ratio of the odds of an event in the experimental (intervention) group to the odds of an event in the control group is defined as the odds ratio. Odds are the ratio of the number of people in a group with an event to the number without an event. Thus, if a group of 100 people had an event rate of 0.20, 20 people had the event and 80 did not, and the odds would be 20/80 or 0.25. An odds ratio of one indicates no difference between comparison groups. For undesirable outcomes an OR that is less than one indicates that the intervention was effective in reducing the risk of that outcome, i.e. a protective effect:

Relative Risk	Exposure
≤0.3	strong protective effect
0.4 – 0.8	protective effect
0.9 – 1.1	no effect
1.2 – 2.5	risk effect
≥2.6	strong risk effect

Table 2-12 Relative risk and exposure. According to Sachs (2003).

		Disease	
		+ (yes)	- (no)
Factor	+ (Allel 1)	a	b
	- (Allel 2)	c	d
			n

Table 2-13 Typical two by two table for odds ratio calculation. According to Sachs (2003).

Relative risk and odds ratio are computed in the following way:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{a \cdot c + a \cdot d}{a \cdot c + b \cdot c}; \qquad OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

When the incidence rate is small, i.e. a is very small compared to b, which is the case for a rare disease, odds ratios are a good estimate of the relative risks (Gordis, 2004):

$$[(a + b) \approx b \text{ and } (c + d) \approx d], \text{ then } OR \approx RR$$

95% Confidence Intervals (95 CI) were calculated according to the following formula (Bland *et al.*, 2000):

$$95\ \mathrm{CI} = \ln(\mathrm{OR}) \pm \left(1,96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

Table 2-13 gives the descriptive effect for the exposure. Whether the observed effect is significant on the chosen level of significance, is tested with the χ^2 test according to Pearson, Mantel, and Haenszel *et al.* (1959):

$$\chi^2 = \frac{(n-1) \cdot (a \cdot d - b \cdot c)^2}{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}$$

In cases of low frequency SNPs, Fisher's exact test was used to compute the case-control statistics. This was done by means of the webinterface SISA (Daan Uitenbrook, Hilversum, The Netherlands).

Eight different odds ratios (table 2-15) can be calculated under assumptions of different genetic models with the data of the following table:

Genotype	Count
11	n_{11}
12	n_{12}
22	n_{22}
Sum	z

Table 2-14 Genotype counts.

Type of odds ratio	High risk factor	Low risk factor
Allelic #1	$2 \cdot n_{11} + n_{12}$	$2 \cdot n_{22} + n_{12}$
Allelic #2	$2 \cdot n_{22} + n_{12}$	$2 \cdot n_{11} + n_{12}$
Carriership #1	$n_{11} + n_{12}$	n_{22}
Carriership #2	$n_{22} + n_{12}$	n_{11}
Homozygotes #1	n_{11}	n_{22}
Homozygotes #2	n_{22}	n_{11}
Heterozygotes #1	n_{12}	n_{11}
Heterozygotes #2	n_{12}	n_{22}

Table 2-15 Different types of odds ratios. The most commonly used odds ratio is carriership of the minor allele, i.e. homozygous and heterozygous carriers of the rare allele are assumed to have the disease.

2.9.3 GENOMIZER - an analysis tool for genome-wide association studies

At the beginning of this study, no integrated and readily applicable software was available to administer and analyze the large amounts of disease association data, which were going to be generated in the two genome-wide experiments.

Therefore, a public-domain software, namely GENOMIZER (Franke *et al.*, 2006), was developed. GENOMIZER implements the workflow of an association experiment, including data management, single-point and haplotype analysis, ''lead'' definition, and data visualization. The software comes with a complete user manual, and is open-source software licensed under the GNU Lesser General Public License. The use of this software facilitated the handling and interpretation of the genome-wide association data.

The functionality of the package is available both from the command-line and through a graphical user interface. A given genome-wide association experiment is defined in terms of its study population (cases and controls) and the marker set used. Individual genotypes are reorganized in ''chromosome data files'' that represent a compact and memory-efficient intermediate data format. Using these files, the following analyses can be performed on a per-chromosome basis:

1. Assessment of general marker characteristics including call rate, minor allele frequency, and deviation from Hardy-Weinberg equilibrium (HWE). In addition to a χ^2 goodness-of-fit test, a recently proposed exact test of HWE has been implemented (Wigginton *et al.*, 2005).
2. Calculation of single-point allelic and genotypic association statistics (e.g., odds ratios) and significance levels using recessive and dominant models.
3. Calculation of sliding-window haplotype statistics and randomization p-values using the expectation maximization (EM)-implementation provided by the WHAP or COCAPAHSE programs (Dudbridge *et al.*, 2003; Dudbridge *et al.*, 2000). The actual program used can be selected by the user.
4. Assessment of pairwise LD over user-defined ranges of the physical map; significance levels are calculated using an efficient EM-implementation limited to two-marker haplotypes (Morton *et al.*, 2001).
5. Sex-specific analysis of X-chromosomal genotypes in which female data are treated as in the autosomal case. Male genotypes are treated as direct haplotype observations and analyzed using a randomization algorithm similar to WHAP/COCAPHASE.

For more details of the software package see Franke *et al.* (2006) or visit the homepage: http://www.ikmb.uni-kiel.de/genomizer. A sample output plot is shown in fig 9-4, page 182.

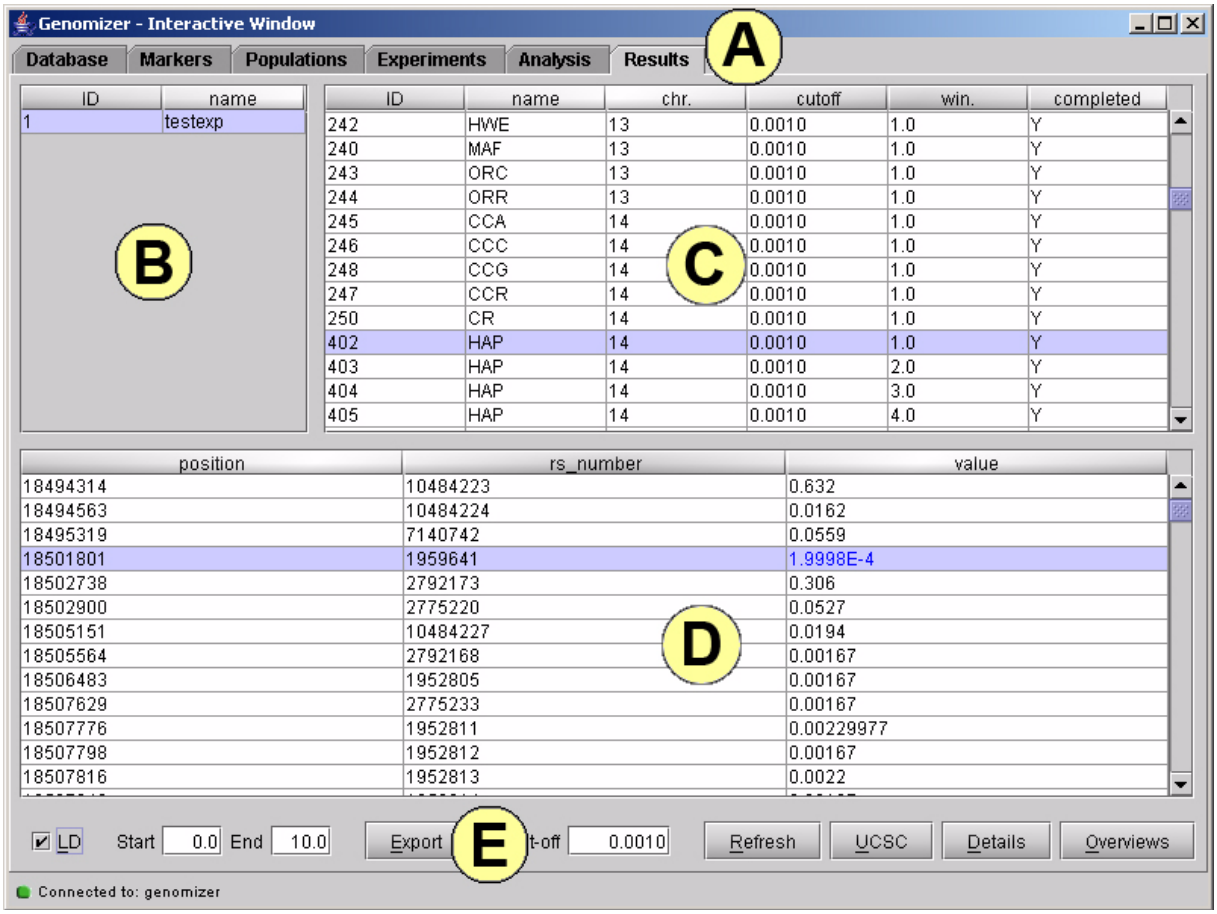

Fig. 2-13 Screenshot of the GENOMIZER graphical user interface (GUI). (A) The main window is based upon a tab interface, which reflects the logical order of a typical association study. (B) GENOMIZER stores multiple experiments based upon the same set of markers in the same database. (C) For each chromosome, various analysis types are performed, each resulting in one unique analysis run. (D) The table shows the analysis results for the selected run. (E) The multifunctional status bar and toolbar for user interaction.

2.9.4 Transmission disequilibrium test (TDT)

The suggestion to use family-based, rather than population-based controls, arises out of concern over the effects of inmeasured population stratification (Spielman *et al.*, 1993), i.e. control individuals may not be well matched to cases, particularly for ethnic origin. Differences in allele frequencies between cases and controls will be interpretetd as evidence for LD even though they may just be the result of differences in ethnic origin. Family-based cases and "internal" controls are able to overcome this problem. Spielman *et al.* (1993 and 1996) coined the term transmission/disequilibrium test using McNemar's test, describing it as "a test for linkage in the presence of association". It is further a meausure of LD in the presence of linkage. The idea that underlies TDTs is that, in the presence of association between a genetic marker and disease susceptibility , the probability of transmission of a marker gene from parents to an affected offspring is significantly increased (over-transmission) or decreased (under-transmission) from the 0.5 value (50% transmission is expected if no linkage and no LD) predicted by Mendelian inheritance.

In the TDT study design, healthy parents with at least one affected offspring (so-called monoplex or nuclear families; Falk and Rubinstein *et al.*, 1987) are used. The number of transmitted alleles from heterozygous parents to affected children is compared to the number of untransmitted alleles (table 2-16). Transmission events from homozygous parents are not informative (fig 2-14, page 65) and these pedigrees are therefore excluded from the TDT test.

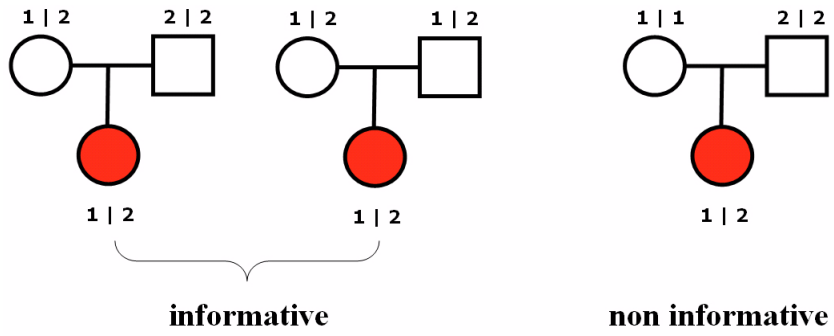

Fig. 2-14 Informative and non-informative trios. Circles denote females and squares denote males, thus mother, father and their diseased child are shown. Only heterozygote parents are informative.

When genotypes for one of the parens were missing, these were excluded as well as Curtis *et al.* (1995) made the observation that the TDT tends to furnish an excess of false-positive results in such cases. However, it is only feasible to trace and genotype both parents of cases for diseases with an early age of onset. In such cases, the use of unaffected sibling controls has been advocated (Spielman *et al.*, 1998).

		Not transmitted	
		1	2
Transmitted	1	n_{11}	n_{12}
	2	n_{21}	n_{22}

Table 2-16 Two-by-two contingency table for the TDT.

As a test statistic, the McNemar test was used (Agresti *et al.*, 1996), which corresponds to an asymptotic chi square distribution under one degree of freedom (dof):

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

In terms of efficiency, the TDT study design is inferior compared to the case-control design. Each trio is scored for 2 transmisssions and two non-transmissions while three individuals in a case-control design give six informative alleles so that the letter provides 1.5 as much information as the former (Morton *et al.*, 1998)

Family-based transmission distortion was analyzed using the transmission disequilibrium test as implemented in Haploview (Barret *et al.*, 2005).

2.9.5 Haplotype analysis

The alleles (at different genes) received by an individual from one parent are called a haplotype (contraction of the phrase "haploid genotype"). In the context of this study, the word haplotype refers to a set of SNPs found to be statistically associated on a single chromatid. Focusing on haplotypes may be more relevant if there is a suspicion that the "additional" disease mutation occured on a specific ancestral haplotype.

In practice, genotypes but not haplotypes are observed. For a two-marker haplotype, eight out of nine genotype combinations can be resolved into pairs of haplotypes. Phase uncertainty exists for the ninth (either 1–1/2–2 or 1–2/2–1), see table 2-17:

	Locus B		
Locus A	**1/1**	**1/2**	**2/2**
1/1	yes	yes	yes
1/2	yes	no	yes
2/2	yes	yes	yes

Table 2-17 Phase uncertainty, a problem of haplotype analyes.

To infer haplotypic phase from unphased genotype data, an algorithm can be formed by a straightforward application of the estimation maximization algorithm (EM; Ott *et al.*, 1977), based on the assumption of Hardy-Weinberg equilibrium. Excoffier *et al.* (1995) were the first to discuss the use of the EM algorithm in this context. In brief, the algorithm works as follows:

Assuming biallelic markers and a two-marker haplotype, only for individuals who are double heterozygotes, the haplotype assignment is ambiguous. It is possible to assign haplotypes to all other individuals because they are homozygous at one of the loci. The EM algorithm is used to assign haplotypes to those people for which haplotype assignment is ambiguous, and it does this by first calculating the frequency of the other haplotype assignments for those individuals whose phase can be resolved. It then makes a guess at the assignment of haplotypes (which is essentially a random guess), and compares this guess to the likelihood based on the observed haplotypes. The goodness of fit between the algorithm's guess and the most likely assignment is then assessed (often through a likelihood function) and the guess adjusted on this basis. This process is iterated until the change in likelihood used to assess the goodness of fit is negligable.

Significance testing of haplotype frequency differences and haplotype estimations were performed with COCAPHASE and TDTPHASE (Dudbridge *et al.*, 2003), making use of the fact that twice the log-likelihood ratio between two nested data models approximately follows a $\chi 2$ distribution with k degrees of freedom, where k is the difference in parameter number be-

tween the two models (Cordell *et al.*, 2002). The logarithm of the likelihood value multiplied by -2 is called the log-likelihood statistic. A log-likelihood statistic is a relative measure of the goodness-of-fit of a specific model and the difference between two log-likelihood statistics produces a test-statistic with an approximate chi-square distribution (Selvin, 2004).

SNPs and haplotypes having a frequency below 1% in the examined population, were excluded from the analysis.

TDTPHASE, COCAPHASE, and Haploview take a standard linkage pedigree file as input. This is a plain text file with lines of the following whitespace-delimited format:

pedigree ID, individual ID, father ID, mother ID, sex, Allele1a, Allele1b, Allele2a, Allele2b...

The pedigree name may be alphanumeric, all other fields must be numeric. Missing data is coded as zero.

COCAPHASE v2.403 was used with the following command:

```
cocaphase input.pre -datafile input.dat -individual -EM -window 2
```

The given example is for a two-marker haplotype. The pre-file contains genotype data for single cases and controls.

TDTPHASE v2.403 was executed from the command prompt as follows:

```
tdtphase input.pre -datafile input.dat -individual -window 2
```

With the above command a two-marker sliding window haplotype analysis will be started. Genotypes for all triplets need to be provided by the input.pre file.

2.9.5.1 Haplotype tagging SNPs

A consequence of the lack of haplotype diversity in regions of strong LD, is that there is considerable redundancy. Most polymorphisms and haplotypes in such a region can be predicted from a smaller subset. Johnson *et al.* (2001) coined the term "haplotype tagging" SNPs, or htSNPs for such a subset. Haplotype tagging SNPs are very valuable for investigating the genetics behind common diseases as genotyping costs are significantly reduced.

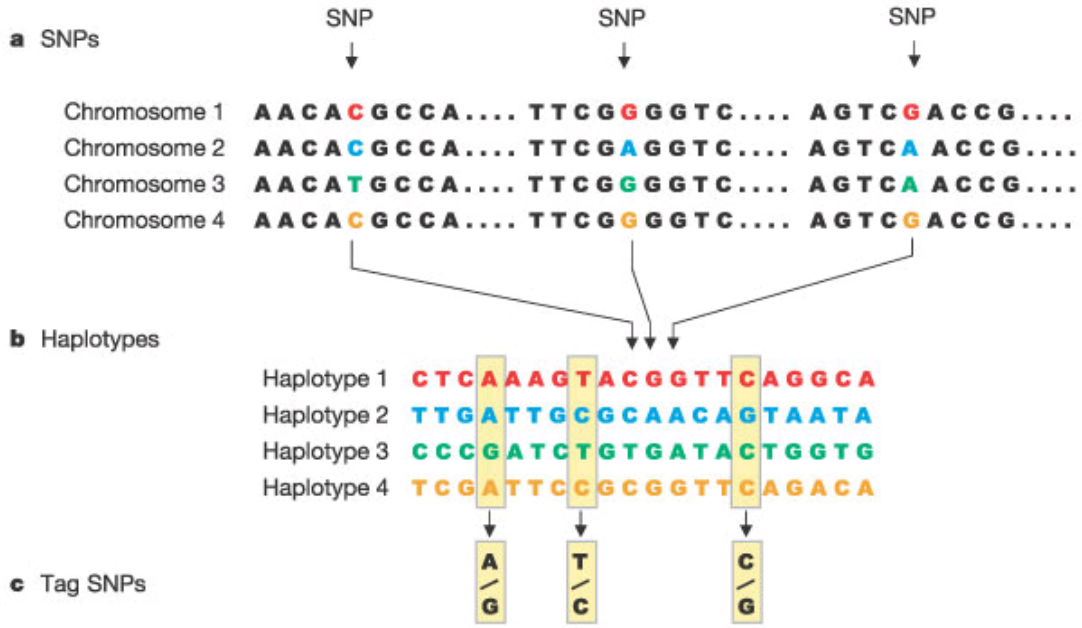

Fig. 2-15 Tagging SNPs. (A) Single nucleotide polymorphisms(SNPs) are identified in DNA samples from multiple indivduals. (B) Adjacent SNPs that are inherited together are compiled into "haplotypes." (c)"Tag" SNPs within haplotypes are identified that uniquely identify those haplotypes. By genotyping the three tag SNPs shown in this figure, researchers can identify which of the four haplotypes shown here are present in each individual. This significantly reduces genotyping costs and efforts. *Illustration from* http://www.hapmap.org.

Haplotype tagging SNPs for candidate genes were selected with Haploview, using the in-built tagger, which is based on Paul de Bakker's algorithm (Bakker *et al.*, 2005). The following default settings were used: "pairwise tagging only" and an r^2 threshold of 0.8. Genotype data was obtained from the HapMap homepage, which currently provides scientists with genotypes for more than three million non-redundant SNPs and four populations (CEU = U.S. residents with northern and western European ancestry, YRI = Yoruba people of Ibadan, Nigeria, CHB = Chinese from Beijing, JPT = Japanese from the Tokyo area; sum of 270 individuals).

2.9.6 Multivariate logistic regression

Logistic regression is part of a category of statistical models called generalized linear modelsand it can be used for genetic data analyses, as the dependent variable is of dichotomous nature (healthy or diseased). It is especially useful to determine the covariates that explain the outcome best. In this study, logistic regression was utilized to determine whether a found mutation is able to predict the disease outcome by itself or whether other surrounding SNPs improve the model. This was done in parallel to a haplotype analysis which gives basicly the same information. Furthermore, logistic regression was used to determine other factors that have an influence on the association, such as sex, age, presence of other known risk factors, et cetera.

Genotype-based logistic regression analysis was performed with R and SPSS 13.0, coding individual SNP genotypes as categorical variables and using forward likelihood ratio inclusion. Analysis of the statistical interactions of risk genotypes, including Breslow-Day tests for odds ratio homogeneity, was done by means of procedure FREQ of the SAS/STAT® software package.

2.9.7 Fisher's exact test

If the number of observations obtained is small, e.g. very rare alleles, Karl Pearson's χ^2 test may produce misleading results. A more appropriate form of analysis (when presented with a 2 x 2 contingency table) is to use Ronald A. Fisher's (1973) exact test (Sachs, 2003). The test is exact because it uses the exact hypergeometric distribution rather than the approximate χ^2 distribution to compute p-values. Fisher's exact test was applied in the present study if the allele count for one of the four cells was less than 5 or if the marginal was very uneven. It was calculated by using the in-house programm `rcexact.exe` (written by Mehta) and by means of the webinterface SISA. Fisher's exact test was not used for regular 2 x 2 tables, as this would have been an over-conservative test.

2.10 Functional experiments

Functional experiments were performed in collaboration with Dr. Philip Rosenstiel, Dr. Andreas Till, Dr. Robert Häsler, and Sven Künzel (all from the Institute of Clinical Molecular Biology, Kiel). In the following sections, only a brief overview of the used methods is given.

2.10.1 Gene expression experiments

To further investigate the pathways of *ATG16L1* and *NELL1*, transcripts were quantified in several stimulated cell lines using fluorogenic real-time PCR.

2.10.1.1 Stimulation of cell lines

Four cell lines from different tissues were used for the subsequent stimulation experiments. All cells were cultured at the cell biology laboratory of the ICMB using DMEM and RPMI culture medium (PAA Laboratories GmbH, Pasching, Austria):

- **HT-29:** Human epithelial colon adenocarcinoma cell line
- **HeLa:** HeLa cells are epithelial-like malignant cells derived from the cervix of Henrietta Lacks (1920-1951).
- **Caco-2:** Human colonic, epithelial-like adenocarcinoma cell line
- **THP-1:** Human acute monocytic leukemia cell line; THP-1 cells have Fc and C3b receptors and lack surface and cytoplasmic immunoglobulins. The cells were described to produce lysozyme and to be phagocytotic. Futhermore, they can be primed to become more macrophage-like and adherent. This induction was carried out by adding PMA (phorbol 12-myristate 13-acetate; Sigma-Aldrich GmbH, Munich, Germany) to the cell culture at a final concentration of 10 nM. Primed THP-1 cells were only used in the stimulation experiment with *Listeria monocytogenes* since adherent leukocytes reflect the *in vivo* response to bacterial invasion (Scorneaux *et al.*, 1996).

The previously mentioned cell lines were incubated/stimulated with the following reagents:

Interferon-γ (IFN-γ)

The cytokine interferon-γ activates the JAK/STAT pathway (Shuai *et al.*, 1992) and is secreted by T-cells and natural killer lymphocyte. 5000 units of IFN-γ were added to the culture.

Lipopolysaccharide (LPS)

LPS is a large molecule that contains both lipid and a carbohydrate. They are a major suprastructure of gram-negative bacteria. LPS contributes greatly to the structural integrity of the bacteria, and protects them from host immune defenses. LPS was acquired from Prof. Zaehringer of the Forschungszentrum Borstel (Germany) and was added to the cell culture at a final concentration of 1 µg/ml.

Flagellin

Flagellin is the principal substituent of the bacterial flagellum, and is present in large amounts on nearly all flagellated bacteria. It is, for example, recognized by Toll-like receptor 5 (TLR-5), which is part of the innate immune system (Gewirtz *et al.*, 2001). Flagelling was added to the medium at a final concentration of 500 ng/ml.

Transforming growth factor β (TGF-β)

TGF-β plays crucial roles in tissue regeneration, cell differentiation, and embryonic development, as it is involved in the TGF signal transduction pathways. A final concentration of 5 ng/ml was used.

Epidermal growth factor (EGF)

The protein EGF is involved in cell growth, proliferatoin, and differentiation. Human EGF is a 6045 Da protein with 53 amino acid residues and three intramolecular disulfide bonds (Carpenter *et al.*, 1984). EGF was added to the medium at a final concentration of 1 µg/ml.

Tumor necrosis factor α (TNFα)

Tumor necrosis factor a is a cytokine with a wide range of pro-inflammatory activities (Beutler *et al.*, 1995) and it is produced primarily by monocytes/macrophages. TNFα has direct cytotoxic effects on the intestinal mucosa in CD and UC but also contributes to the systemic manifestations seen in these diseases. Anti-TNFα antibodies have shown a clear anti-inflammatory effect in patients with Crohn's disease. A final concentration of 10 ng/ml was used.

Listeria monocytogenes

Listeria monocytogenes is a gram-positive bacterium which is motile by means of flagella. It is easy to cultivate, non-spore-forming and ubiquitous. Bacteria were cultured overnight and fresh cultures were inoculated and grown until a density of OD_{650} ~0.2 was reached. Approximately 100 bacterias per cell were added to the medium. When the medium was changed from "-/-" to "+/+" (10% FCS [fetal calf serum] -1% P/S [Penicillin/Streptomycin]) medium after 1 hour of incubation, time was measured.

All cells were exposed to the stimuli for 0, 1, 2, and 8 hours and each experiment was performed in duplicate. Afterwards, the RNA was extracted as described in the next section.

2.10.1.2 RNA extraction from cells

Cells were cultured in 2 ml medium and washed with PBS in 6-well plates (Nunc GmbH, Wiesbaden). Afterwards, the cells were carefully scraped off and centrifuged for 5 min at 2000 rpm. RNA was extracted from the resuspensed pellet (1 ml RLT containing 10 µl/ml β-Mercaptoethanol) by using the RNeasy-Kit of Qiagen (Hilden, Germany). All steps were carried out according to the standard protocol. The protocol included a DNase I digestion step to de-

stroy remaining DNA. All RNAs were checked on a gel for degradation (only 18 S and 28 S bands are visible in intact RNA, smear indicates degradation) before they were used for cDNA synthesis. Concentrations were measured using 1 µl of RNA and a NanoDrop® ND-1000 Spectrophotmeter (NanoDrop Technologies, Wilmington, DE, USA). Quality of RNA was further checked by using an aliquot of RNA in a PCR with *GAPDH* primers. A band of 360 bp indicated the isolation of intact RNA.

2.10.1.3 cDNA synthesis

For each sample, cDNA was reverse-transcribed from 1 µg of total RNA in a total volume of 100 µl using the SMART™ PCR cDNA Synthesis Kit from Clontech.

The success of the first strand cDNA synthesis was then checked by using 2.5 µl cDNA as a template for PCR amplification of *β-Actin*:

95°C for 5 min; 35 cycles: 95°C for 30 sec, 58°C for 30 sec, 72°C for 45 sec; 72°C for 10 min

The amplification of the 518 bp *β-Actin* PCR product, indicated successful first stand synthesis. The final theoretical concentration of cDNA, assuming 100% efficiency of reverse transcription, was 10 ng of cDNA/µl.

2.10.1.4 Plate production

25 µl of cDNAs were diluted 1:5 with water to a final volume of 125 µl and a theoretical concentration of 2 ng/µl before being arrayed to 384-deep well plates. For relative quantitation using a standard curve (according to Livak *et al.*, 2001), several cDNAs were mixed and serially diluted (Undiluted, 1:1.5, 1:5, 1:10, 1:20, 1:40, 1:60, 1:80, 1:100, 1:200). A plate-layout, showing the random distribution of the duplicates, is shown in fig 2-16, page 73.

		Probe												Technical Replicate											
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Housekeeper	A	38.1	36.1	17.1	28.1	49.1	12.1	18.1	15.1	74.1	52.1	30.1	64.1	38.2	36.2	17.2	28.2	49.2	12.2	18.2	15.2	74.2	52.2	30.2	64.2
	B	55.1	79.1	83.1	62.1	9.1	90.1	23.1	71.1	84.1	68.1	87.1	19.1	55.2	79.2	83.2	62.2	9.2	90.2	23.2	71.2	84.2	68.2	87.2	19.2
	C	57.1	69.1	37.1	50.1	42.1	56.1	88.1	59.1	44.1	7.1	27.1	63.1	57.2	69.2	37.2	50.2	42.2	56.2	88.2	59.2	44.2	7.2	27.2	63.2
	D	11.1	39.1	95.1	31.1	35.1	92.1	96.1	5.1	47.1	70.1	40.1	29.1	11.2	39.2	95.2	31.2	35.2	92.2	96.2	5.2	47.2	70.2	40.2	29.2
	E	76.1	58.1	67.1	60.1	4.1	22.1	10.1	3.1	85.1	93.1	75.1	91.1	76.2	58.2	67.2	60.2	4.2	22.2	10.2	3.2	85.2	93.2	75.2	91.2
	F	26.1	32.1	13.1	33.1	2.1	81.1	66.1	34.1	48.1	53.1	21.1	61.1	26.2	32.2	13.2	33.2	2.2	81.2	66.2	34.2	48.2	53.2	21.2	61.2
	G	6.1	16.1	24.1	78.1	77.1	89.1	94.1	65.1	20.1	82.1	8.1	46.1	6.2	16.2	24.2	78.2	77.2	89.2	94.2	65.2	20.2	82.2	8.2	46.2
	H	54.1	1.1	80.1	72.1	25.1	43.1	41.1	86.1	45.1	14.1	73.1	51.1	54.2	1.2	80.2	72.2	25.2	43.2	41.2	86.2	45.2	14.2	73.2	51.2
Target	I	38.3	36.3	17.3	28.3	49.3	12.3	18.3	15.3	74.3	52.3	30.3	64.3	38.4	36.4	17.4	28.4	49.4	12.4	18.4	15.4	74.4	52.4	30.4	64.4
	J	55.3	79.3	83.3	62.3	9.3	90.3	23.3	71.3	84.3	68.3	87.3	19.3	55.4	79.4	83.4	62.4	9.4	90.4	23.4	71.4	84.4	68.4	87.4	19.4
	K	57.3	69.3	37.3	50.3	42.3	56.3	88.3	59.3	44.3	7.3	27.3	63.3	57.4	69.4	37.4	50.4	42.4	56.4	88.4	59.4	44.4	7.4	27.4	63.4
	L	11.3	39.3	95.3	31.3	35.3	92.3	96.3	5.3	47.3	70.3	40.3	29.3	11.4	39.4	95.4	31.4	35.4	92.4	96.4	5.4	47.4	70.4	40.4	29.4
	M	76.3	58.3	67.3	60.3	4.3	22.3	10.3	3.3	85.3	93.3	75.3	91.3	76.4	58.4	67.4	60.4	4.4	22.4	10.4	3.4	85.4	93.4	75.4	91.4
	N	26.3	32.3	13.3	33.3	2.3	81.3	66.3	34.3	48.3	53.3	21.3	61.3	26.4	32.4	13.4	33.4	2.4	81.4	66.4	34.4	48.4	53.4	21.4	61.4
	O	6.3	16.3	24.3	78.3	77.3	89.3	94.3	65.3	20.3	82.3	8.3	46.3	6.4	16.4	24.4	78.4	77.4	89.4	94.4	65.4	20.4	82.4	8.4	46.4
	P	54.3	1.3	80.3	72.3	25.3	43.3	41.3	86.3	45.3	14.3	73.3	51.3	54.4	1.4	80.4	72.4	25.4	43.4	41.4	86.4	45.4	14.4	73.4	51.4

Fig. 2-16 Plate layout for gene expression experiments. Each sample is arrayed once per quadrant, i.e. in duplicates per tested assay. To circumvent artefacts arising by physical proximity, samples and their replicates are semi-randomly distributed. The precise llocalization of each sample is given in the file `APG16L_Realtime_Analysis.xls` on the DVD.

For the real-time PCR reaction, 5 µl of the cDNA solution were pipetted from the deepwell plates into 384 well PCR plates in quadruplicates: two wells for β-actin (“housekeeper”) and two wells for “target” gene quantitation.

2.10.1.5 Real-time PCR

The three commercially available TaqMan® Gene Expression assays for *ATG16L1* (Hs00250530_m1), *NELL1* (Hs00971083_m1), and β-Actin (Hs99999903_m1) were ordered through Applied Biosystems (Foster City, CA, USA). All assay probes spanned an exon junction and did not detect genomic DNA (as indicated by the "_m" in the name). Probes for real-time PCR were labelled with the fluorescent reporter dye 6-carboxyfluorescein (FAM) and the quencher dye 6-carboxytetramethylrhodamine (TAMRA) at the 5' and 3' ends, respectively.

Reaction mix was prepared according to table 2-18 and dispensed by multipipetting into a 384-well PCR plate containing 5 µl of cDNA. All reagents were kept on ice during the pipetting procedure. Plates were sealed with an aluminum adhesive cover and stored at 4°C in the dark if they were not immediately run.

Reagent	Volume [µl]	Final concentration
cDNA template (2 ng/µl)	5	1 ng/µl
20x Assay-on-Demand	0.5	1x
2x TaqMan® Universal PCR Master Mix	4.5	0.9x
Total	10	

Table 2-18 Real-time PCR master mix. Volumes are given for a single reaction. For a 384-well MTP, 423 reactions were prepared (384x + 10% excess volumen)

Prior to running on an ABI 7900HT Sequence Detection System, plates were briefly centrifuged. The following thermocycling profile was used: 50°C for 2 min (AmpErase UNG activation); 95°C for 10 min (AmpliTaq Gold® activation); and 40 cycles of: 95°C for 15 s; 60°C for 1 min. Upon completion of the run, the data was analysed by the Sequence Detection System 2.1 (SDS) software using the following parameters:

1. The threshold line was set as low as possible such that the line intersects with the amplification curves in the exponential phase, thus reducing any amplification in the non-template controls to a negligible levels
2. The baseline was set separately for housekeeper and target such that the baseline starts at three cycles and ends just before the first amplification curves rise above the threshold.
3. Extreme outliers in the standard curve were eliminated from the calculation of the standard curve.

Upon adjusting all of these parameters, the SDS software automatically calculated the relative quantities, which were then exported as a tab-delimited text file for further data processing with MS Office Excel 2003.

Relative quantities of the target gene were normalized by dividing through the corresponding relative quantities of the housekeeping gene, namely β-acting. Then, the fold-changes were calculated by dividing each normalized value through the correspondinging normalized reference (e.g. 0 hour stimulation sample or healthy control person). Finally, average fold-changes and standard deviations were calculated. Fold-changes below zero indicated a downregulation of the transcript and they were normalized to values greater than 1 by taking the inverse 1/n.

2.10.2 Isolation of primary epithelial cells (IECs)

Epithelial cell preparation was carried out using a standard protocol as described (Rosenstiel *et al.*, 2003; Daig *et al.*, 2000). In brief, mucosal biopsies were placed in 1.5 mM EDTA in Hanks balanced salt solution without calcium and magnesium (HBSS) and tumbled for 10 minutes at 37°C. This supernatant containing debris and mainly villus cells was discarded. The mucosa was incubated again with HBSS/EDTA for 10 minutes at 37°C. The supernatant was collected into a 15 ml tube. The remaining mucosa was shortly vortexed in PBS and this supernatant was also collected. It contained complete crypts, some single cells, and a small amount of debris. To separate IECs (crypts) from contaminating non-epithelial cells, the suspension was allowed to sediment for 15 minutes. The cells (mainly complete crypts) were collected and washed twice with PBS. The number and viability of the cells were determined by trypan blue exclusion. The purity of the epithelial cell preparation was checked by routine hematoxylin-eosin staining, showing more than 90% of epithelial cells.

2.10.3 RT-PCR

Total RNA from primary intestinal epithelial cells was isolated using the RNeasy kit from Qiagen. RNA yield was determined using a NanoDrop® ND-1000 spectrophotometer. 300 ng of total RNA were reverse transcribed using the Advantage RT-for-PCR kit with oligo-dT-primers according to the manufacturer's protocol and Waetzig *et al.* (2005).

Resulting cDNA (5 µl of RT reaction) was amplified using GeneAmp PCR buffer and AmpliTaq DNA Polymerase in a 50 µl reaction volume containing primer pairs (0.2 µM/primer) and 0.2 mM dNTPs. Negative RT and PCR controls included reactions in the absence of RNA sample, reverse transcriptase, and cDNA, respectively.

For investigation of tissue specific expression patterns, a commercial tissue panel was obtained from Clontech (Human Multiple Tissue cDNA Panel I and II). The concentration of each cDNA was approx. 1 ng/µl. Primers used for amplification of *ATG16L1* are listed in table 8-3, page 163 (expected amplicon length: 231 bp). The following conditions were applied: Denaturation for 5 min at 95°C; 25 cycles of 30 s at 95°C, 20 s at 60°C, 45 s at 72°C; final

extension for 10 min at 72°C. The number of cycles was kept in the range of linear amplification, thus to allow a semi-quantitative analysis.

To confirm the use of equal amounts of RNA in each experiment all samples were checked in parallel for β-actin mRNA expression. All amplified DNA fragments were analyzed on 1% agarose gels and subsequently documented by a BioDoc Analyzer as described in 2.6 on page 33.

2.10.4 Western Blot

Biopsies from five healthy controls without any obvious intestinal pathology and five Crohn patients with confirmed ileal and colonic inflammation were lysed and subjected to Western blot analysis as described by Waetzig *et al.* (2005). 10 µg of total protein were separated by SDS polyacrylamide gel electrophoresis and transferred to PVDF membrane by standard techniques. ATG16L was detected using a polyclonal clonal anti-ATG16 antibody (anti-ATG16L, ABGENT, San Diego, CA).

2.10.5 Immunohistochemistry

Paraformaldehyde-fixed paraffin-embedded biopsies from normal controls (n = 5) and from patients with confirmed colonic CD (n = 5), which were obtained in parallel from the same sites as the biopsies used for the expression analysis studies (2.10.3 on page 75), were analysed. Two slides of each biopsy were stained with hematoxylin-eosin for routine histological evaluation. The other slides were subjected to a citrate-based antigen retrieval procedure, permeabilized by incubation with 0.1% Triton X-100 in 0.1 M phosphate-buffered saline (PBS), washed three times in PBS and blocked with 0.75% bovine serum albumin in PBS for 20 minutes. Sections were subsequently incubated with the primary antibody (anti-ATG16L) at a 1:200 dilution in 0.75% BSA for 1 h at room temperature. After washing in PBS, tissue bound antibody was detected using biotinylated goat-anti rabbit followed by HRP-conjugated avidin, both diluted at 1:100 in PBS. Controls were included using irrelevant primary antibodies as well as omitting the primary antibodies using only secondary antibodies and/or HRP-conjugated avidin. Bound antibody was detected by standard chromogen technique (Vector Laboratory) and visualized by an Axiophot microscope. Pictures were captured by a digital camera system (Axiocam, Zeiss).

2.11 Protein modeling

Protein modeling of *ATG16L1* was done in cooperation with the Max-Planck Institute for Bioinformatics (Saarbrücken) in the following way:

Protein sequences were retrieved from the UniProt (Wu *et al.*, 2006) and Ensembl (Birney *et al.*, 2006) databases and protein domain architectures were taken from the Pfam database (Finn *et al.*, 2006). 3D crystal structure coordinates were obtained from the Protein Data Bank (Kouranov *et al.*, 2006) and corresponding domain definitions from the SCOP database (Andreeva *et al.*, 2004). Ensembl identifiers, UniProt accession numbers, and PDB codes are listed in table 2-19 and table 2-20 for fig. 3-6 and fig. 9-5, respectively.

Using the alignment program MUSCLE (Edgar *et al.*, 2004), multiple sequence alignments (fig. 3-6 and fig. 9-5) of ATG16L1 homologs were computed, contained in the Ensembl family ATG16L1 (identifier ENSF00000001431) and the Pfam family ATG16 (accession number PF08614). To analyze the alignments further, the following evolutionarily related WD-repeat proteins with known 3D structures were included: yeast cell division control protein 4 (CDC4; Orlicky *et al.*, 2003), yeast SIR4-interacting protein 2 (SIR2; Cerna *et al.*, 2005), yeast glucose repression regulatory protein 1 (TUP1; Sprague *et al.*, 2000), and human transducin-like enhancer protein 1 (TLE1; Pickles *et al.*, 2002). The alignments were manually refined based on multiple sequence alignments of the WD40 Pfam family (accession number PF00400), the published WD-repeat consensus sequence (Li *et al.* 2001; Smith *et al.* 1999), and structure superpositions of WD-repeat proteins using FATCAT (Ye *et al.*, 2004). The alignment figures were prepared and illustrated using the editors GeneDoc (Nicholas *et al.* 1997), Jalview (Clamp *et al.*, 2004), and SeaView (Galtier *et al.*, 1996).

The secondary structure assignment to PDB structures was obtained from the DSSP database (Kabsch *et al.*, 1983). To predict the secondary structure of ATG16L1 homologs in different species, the prediction servers PSIPRED (McGuffin *et al.*, 2000), YASPIN (Pyo *et al.*, 2005), PROFsec (Rost *et al.*, 2004), and Porter (Pyo *et al.*, 2005) were contacted. Consensus predictions by majority voting (Albrecht *et al.*, 2003) were also formed. All servers consistently predicted β-strands characteristic of eight WD repeats in ATG16L1, and a helical linker may precede the eighth WD-repeat as it is the case for CDC4 (fig 9-6, page 184; Orlicky *et al.*, 2003).

To predict the three-dimensional WD-repeat domain structure of human ATG16L1, we investigated the fold recognition results returned by the BioInfoBank online meta-server (Bujnicki *et al.*, 2001) and compared them to the very similar predictions by the web servers FFAS03 (Jaroszewski *et al.*, 2005) and Arby (von Öhsen *et al.*, 2004). The BioInfoBank server contacts a dozen other structure prediction servers and is coupled to a 3D-Jury system that assesses the quality of the returned results based on a sophisticated scoring scheme (Ginalski *et al.*, 2003). FFAS03 and ARBY also provide statistically derived confidence scores for structure predictions.

In agreement with the WD40 Pfam family classification, all servers predicted at least seven WD repeats at the C-terminus of ATG16L1 starting near residue P311. In addition, the secondary structure predictions for human ATG16L1 and its species homologs as well as the conservation of amino acids characteristic of WD repeats (Li *et al.* 2001; Smith *et al.* 1999) indicated an eighth non-canonical WD repeat in the 40-residue region P271-V310 (fig 9-5, page 183 and fig 9-6, page 184).

Since diverged WD repeats have already been observed with other WD-repeat proteins such as CDC4, SIF2, and coronin-1 (Orlicky *et al.*, 2003; Cerna *et al.*, 2005; Appleton *et al.*, 2006), the eight-bladed β-propeller structure of the WD-repeat domain from the CDC4 subunit of the yeast SCF ubiquitin ligase complex was chosen as structural template for ATG16L1. Because the WD-repeat domain of ATG16L1 is replaced by an actin domain in the ATG16L1 homolog of *Ustilago maydis* (UniProt accession number Q4P303), one may speculate that the WD-repeat domain of human ATG16L1 is involved in actin regulation like coronin proteins with a domain architecture similar to ATG16L1 (Rybakin *et al.*, 2005). To model the 3D protein structure of ATG16L1, pairwise sequence-structure alignment from the manually curated multiple alignment of ATG16L1 and WD-repeat domain homologs (fig 9-5, page 183) were extracted and submitted to the WHAT IF server (Rodriguez *et al.*, 1998). The image of the resulting full-atom protein structure model (fig 3-7, page 112) was illustrated using Yasara and POV-Ray.

Species	Alignment	Ensembl identifier
Homo sapiens	ATG16L1_Ho_sa	ENSP00000287735
Macaca mulatta	ATG16L1_Ma_mu	ENSMMUP00000030747
Pan troglodytes	ATG16L1_Pa_tr	ENSPTRP00000022319
Mus musculus	ATG16L1_Mu_mu	ENSMUSP00000027512
Rattus norwegicus	ATG16L1_Ra_no	ENSRNOP00000024445
Canis familiaris	ATG16L1_Ca_fa	ENSCAFP00000017340
Bos taurus	ATG16L1_Bo_ta	ENSBTAP00000005140
Gallus gallus	ATG16L1_Ga_ga	ENSGALP00000002472
Monodelphis domestica	ATG16L1_Mo_do	ENSMODP00000013062
Danio rerio	ATG16L1_Da_re	ENSDARP00000055987
Fugu rubripes	ATG16L1_Fu_ru	NEWSINFRUP00000138508
Tetraodon nigroviridis	ATG16L1_Te_ni	GSTENP00021121001
Xenopus tropicalis	ATG16L1_Xe_tr	ENSXETP00000030368

Table 2-19 Species and Ensembl identifiers for the homologous ATG16L1 sequences that were used to generate fig 3-6, page 111.

Protein	Species	Alignment	UniProt/Ensembl	PDB
ATG16L1	*Homo sapiens*	ATG16L1-Ho-sa	Q676U5	–
ATG16L1	*Bos taurus*	ATG16L1-Bo-ta	ENSBTAP00000005140	–
ATG16L1	*Canis familiaris*	ATG16L1-Ca-fa	ENSCAFP00000017340	–
ATG16L1	*Gallus gallus*	ATG16L1-Ga-ga	ENSGALP00000002472	–
ATG16L1	*Mus musculus*	ATG16L1-Mu-mu	Q8C0J2	–
ATG16L1	*Rattus norwegicus*	ATG16L1-Ra-no	ENSRNOP00000024445	–
ATG16L1	*Tetraodon nigroviridis*	ATG16L1-Te-ni	Q4SB59	–
CDC4	*Saccharomyces cerevisiae*	CDC4-Sa-ce	P07834	1nex, chain B
SIF2	*Saccharomyces cerevisiae*	SIF2-Sa-ce	P38262	1r5m, chain A
TUP1	*Saccharomyces cerevisiae*	TUP1-Sa-ce	P16649	1erj, chain A
TLE1	*Homo sapiens*	TLE1-Ho-sa	Q04724	1gxr,chain A

Table 2-20 Ensembl/UniProt identifiers for ATG16L1 homologs and related WD-repeat proteins shown in fig 9-5, page 183. PDB codes are given for the WD-repeat domain structures CDC4, SIR2, TUP1, and TLE1.

3 Results

3.1 Genome-wide screenings (GWS)

With the objective to find new suceptibility regions and genes involved in the etiology of inflammatory bowel diseases, two (technically) independent and hypothesis-free genome-wide screenings were carried out. In the first scan, approximately 20,000 protein-changing SNPs were directly typed with SNPlex™. The second approach aimed to identify new IBD loci by typing 116,161 evenly distributed SNPs across the genome, thereby relying on existing linkage disequilibrium with the disease-causing variants. A summary of the used samples/ panels is given in table 2.2, page 25.

3.1.1 Direct approach: cSNP experiment

A total of 19,779 non-synonymous SNPs were genotyped in the samples of panel A (735 CD patients and 368 controls from Northern Germany (table 2-1, page 26), using the SNPlex™ system as described in section 2.8.2 on page 44. The control plate XR01 and the two patient plates XR02, and XR03 were used. Genotyping was successful for 16,360 assays, as defined by a mean fluorescence reading greater 500 units on the ABI 3730xl sequencer. Of these SNPs, some 7,159 occurred at a minor allele frequency greater 1% and were thus included in the subsequent analyses with GENOMIZER. These markers were ranked and prioritized for follow-up on the basis of the p-values obtained in the single-locus allele-based and genotype-based test for disease association. A p-value of 0.01 in the allele-based test was used as a cut-off for inclusion in a replication study, which resulted in 72 putative disease variants that were evaluated in panel B (380 German CD trios, 878 single patiens and 1032 independent controls, see also table 2-1, page 26). Initial results and outcomes of the replication for these 72 polymorphisms are given in table 3-1, page 81. For the replication studies, the following plates were used:

- XN33, XGCON01, XGCON02 (controls)
- XG01, XG02, XG03, XG04 (trios)
- XG05, XG06 (single cases)

When $p < 0.05$ in both the TDT and the case-control comparison was held to indicate formal replication, only three markers, rs2241880 in the *ATG16L1* gene and the previously reported variants, rs1050152 (Leu503Phe) in the *OCTN1* gene (Peltekova *et al.*, 2004), and the known *CARD15* SNP rs2066845 ("SNP12", Hugot *et al.*, 2001), were found to match this criterion. The newly found association of the “G” allele at rs2241880 with CD was significant with $p = 1.6 \times 10^{-05}$ in the allele-based case-control comparison and with $p = 2.7 \times 10^{-05}$ in the TDT. Thus, all subsequent mapping and replication efforts of this GWS were confined to this variant.

Screening (panel A)								Replication (panel B)		
#	gene	celera ID	dbSNP ID	chr.	position	P_{CCA}	P_{CCG}	P_{CCA}	P_{CCG}	P_{TDT}
1	DCP1B	hCV2194128	rs12423058	12	1,934,927	$5.8 \cdot 10^{-14}$	$3.6 \cdot 10^{-13}$	0.92	0.54	0.07
2	TINAG	hCV12027972	rs1058768	6	54,232,983	$1.7 \cdot 10^{-12}$	$7.9 \cdot 10^{-11}$	0.27	0.15	0.21
3	OR8H1	hCV25770775	rs17613241	11	55,839,523	$2.2 \cdot 10^{-09}$	$2.8 \cdot 10^{-09}$	0.27	0.41	0.15
4	TTN	hCV25626488	rs10497517	2	179,646,084	$3.4 \cdot 10^{-07}$	$2.7 \cdot 10^{-06}$	0.17	0.37	0.70
5	OR10A4	hCV15895352	rs2595453	11	6,862,804	0.00005	0.0003	0.18	0.21	0.43
6	hCG1744077	hCV3111449	rs211716	1	75,529,932	0.0001	0.0006	0.05	0.16	0.14
7	hCG1744077	hCV928121	rs211715	1	75,530,066	0.0002	0.001	0.07	0.21	0.19
8	S100Z	hCV8796177	rs1320308	5	76,255,325	0.0003	0.0010	0.05	0.11	0.40
9	IL7R	hCV2025977	rs6897932	5	35,920,076	0.0004	0.0002	0.91	0.99	0.95
10	ATG16L	hCV9095577	rs2241880	2	234,470,182	0.0004	0.002	0.00001	0.00007	0.00001
11	FLJ23577	hCV25770123	–	5	35,715,804	0.0004	0.004	0.21	0.41	0.81
12	U2	hCV25637975	rs6730351	2	223,793,960	0.0007	0.003	0.55	0.81	0.50
13	APBB2	hCV1558531	rs4861358	4	40,931,441	0.0009	0.004	0.04	0.06	0.57
14	SLC17A3	hCV1911085	rs1165165	6	25,970,445	0.0009	0.004	0.58	0.26	0.90
15	hCG1789632	hCV25929364	rs10948733	6	52,867,218	0.0009	0.004	0.05	0.04	0.50
16	NALP13	hCV2092168	rs303997	19	61,116,255	0.001	0.005	0.73	0.86	0.82
17	hCG1812162	hCV25994942	rs10483261	14	20,346,679	0.001	0.005	0.79	0.76	0.08
18	hCG1646471	hCV15965545	rs2291479	3	179,495,857	0.001	0.006	0.49	0.72	0.60
19	HS6ST3	hCV3118872	rs2282135	13	95,187,906	0.001	0.003	0.12	0.25	0.86
20	PKD1L2	hCV8443426	rs1869348	16	80,921,788	0.002	0.003	0.0004	0.002	0.52
21	VGF	hCV25649609	–	7	100,378,082	0.002	0.006	0.27	0.17	0.13
22	TXNDC11	hCV1388401	rs3190321	16	11,740,094	0.002	0.002	0.09	0.04	0.74
23	PLSCR4	hCV25647383	rs3762685	3	147,259,528	0.002	0.005	0.71	0.77	0.30
24	OR5U1	hCV2519378	rs9257694	6	29,382,496	0.002	0.008	0.52	0.77	1.00
25	UBQLN4	hCV16187524	rs2297792	1	153,228,236	0.002	0.006	0.003	0.009	0.65
26	CARD15	hCV11717466	rs2066845	16	50,543,573	0.002	0.008	$8.6 \cdot 10^{-08}$	$7.1 \cdot 10^{-07}$	0.002
27	FUCA1	hCV12023629	rs11549094	1	23,650,437	0.002	0.008	0.48	0.54	0.69
28	hCG1999532	hCV2481084	rs3129096	6	29,291,365	0.002	0.008	0.45	0.67	0.83
29	OR2J2	hCV11194783	rs3116817	6	29,257,553	0.002	0.010	0.57	0.84	0.61
30	FLJ25660	hCV2537241	rs541169	19	40,410,860	0.003	0.01	0.77	0.39	0.32
31	KUB3	hCV25770320	rs3751325	12	56,621,893	0.003	0.001	0.46	0.75	0.25
32	SLC16A4	hCV15961275	rs2271885	1	110,220,442	0.003	0.01	0.31	0.09	0.10
33	U1	hCV2475291	rs2157453	1	170,103,324	0.003	0.0008	0.002	0.006	0.29
34	SLC22A4	hCV3170459	rs1050152	5	131,752,536	0.003	0.003	$2.6 \cdot 10^{-06}$	$1.5 \cdot 10^{-06}$	0.02
35	AQP9	hCV11669234	rs1867380	15	56,192,337	0.003	0.01	0.02	0.03	0.67
36	DHX34	hCV11507064	rs12984558	19	52,548,176	0.003	0.010	0.82	0.87	0.36
37	hCG26636	hCV2942610	rs1864147	16	64,719,751	0.003	0.01	0.06	0.16	0.29
38	DP58	hCV622249	rs32857	5	79,939,445	0.003	0.0008	0.93	0.36	0.15
39	PLSCR4	hCV9539784	rs1061409	3	147,238,670	0.003	0.01	0.93	0.99	0.29
40	hCG2038517	hCV1734658	rs3810071	18	2,508,697	0.003	0.006	0.26	0.24	0.02
41	ST5	hCV1506057	rs3812762	11	8,715,949	0.003	0.008	0.71	0.54	0.87
42	OAS2	hCV8920052	rs15895	12	111,860,241	0.004	0.01	0.60	0.74	0.91
43	FLJ46906	hCV8275411	rs1129180	6	138,998,702	0.004	0.006	0.72	0.91	0.51
44	C14orf125	hCV8601135	rs7157977	14	29,848,248	0.004	0.02	0.29	0.35	0.96
45	AKAP10	hCV926535	rs203462	17	19,974,570	0.004	0.004	0.55	0.79	0.01

Table 3-1 Lead SNPs of cSNP screening. Top 72 CD-associated SNPs, ranked with respect to the p-value obtained in an allele-based case-control comparison (CCA) in panel A. Also included are the p-values for the genotype-based case-control comparison (CCG) and the TDT. Nucleotide positions refer to NCBI build 34. The 17 markers with $p < 0.05$ in either the case-control or the TDT analysis of replication panel B are highlighted in blue. SNPs with a significant result in both panel B tests are marked by grey shading. In addition to rs2241880, only SNPs rs1050152 (Leu503Phe) in the *OCTN1* gene, reported earlier (Peltekova *et al.*, 2004) and the known *CARD15* SNP rs2066845 ("SNP12", Hugot *et al.*, 2001) yielded consistent replication.

Screening (panel A)								Replication (panel B)		
#	gene	celera ID	dbSNP ID	chr.	position	P_{CCA}	P_{CCG}	P_{CCA}	P_{CCG}	P_{TDT}
46	CACNA1E	hCV1432822	rs704326	1	178,999,038	0.004	0.01	0.32	0.61	0.17
47	KNSL7	hCV25924111	rs3804583	3	44,845,239	0.004	0.01	0.22	0.38	1.00
48	THRAP3	hCV25749777	rs6425977	1	36,180,095	0.004	0.01	0.36	0.47	0.01
49	SLC1A4	hCV2681351	rs759458	2	65,219,899	0.004	0.02	0.81	0.82	0.25
50	U15	hCV3215915	rs4774310	15	56,701,220	0.005	0.02	0.23	0.45	0.33
51	*MYO10*	hCV3132500	rs27431	5	16,723,356	0.005	0.01	0.70	0.83	0.55
52	*IFI44L*	hCV11873694	rs3820093	1	78,518,119	0.005	0.007	0.35	0.63	0.92
53	*CAPSL*	hCV8811801	rs1445898	5	35,956,030	0.005	0.01	0.35	0.53	0.75
54	*FLJ31846*	hCV25959811	rs3764147	13	42,255,925	0.005	0.009	0.05	0.11	0.77
55	*hCG1794790*	hCV37420	rs13092702	3	147,440,204	0.005	0.02	0.28	0.43	0.79
56	*FLJ46320*	hCV25973127	rs3829486	16	86,881,788	0.006	0.02	0.05	0.14	0.02
57	*NUDCD1*	hCV15883329	rs2980618	8	110,258,581	0.006	0.02	0.29	0.56	0.33
58	*UBAP2*	hCV8778477	rs1785506	9	34,007,106	0.006	0.02	0.10	0.21	0.96
59	*A2BP1*	hCV2973884	rs2191423	16	6,387,642	0.006	0.01	0.06	0.16	0.03
60	*LRRK2*	hCV3215842	rs3761863	12	39,044,919	0.007	0.02	0.60	0.80	0.06
61	*MYO5A*	hCV25592080	–	15	50,351,450	0.007	0.006	0.14	0.27	0.04
62	*hCG1994124*	hCV2441812	rs2157650	8	17,715,645	0.007	0.0004	0.11	0.24	0.13
63	*hCG2040272*	hCV25988773	rs10427252	2	215,765,119	0.007	0.02	0.09	0.14	0.90
64	*BCAR1*	hCV25764619	–	16	75,048,530	0.008	0.02	0.04	0.04	0.74
65	*C14orf8*	hCV2434490	rs9624	14	19,490,249	0.008	0.010	0.35	0.64	0.21
66	*U10*	hCV12017135	rs1826619	10	31,005,500	0.008	0.02	0.21	0.41	0.37
67	*FLJ23577*	hCV25742805	rs7710284	5	35,738,276	0.008	0.03	0.80	0.53	0.42
68	*NALP8*	hCV8110157	rs306481	19	61,179,415	0.008	0.04	0.97	0.93	0.96
69	*U1*	hCV27080230	rs4534436	1	119,108,418	0.008	0.02	0.21	0.16	0.31
70	*IGHMBP2*	hCV25474530	rs17612126	11	68,481,034	0.009	0.01	0.98	0.92	0.27
71	*USP16*	hCV2870492	rs2274802	21	29,330,540	0.009	0.04	0.52	0.18	0.11
72	*CLEC2D*	hCV25992569	rs3764022	12	9,724,791	0.009	0.007	0.82	0.80	0.14

Table 3-1 Lead SNPs of cSNP screening. Top 72 CD-associated SNPs, ranked with respect to the p-value obtained in an allele-based case-control comparison (CCA) in panel A. Also included are the p-values for the genotype-based case-control comparison (CCG) and the TDT. Nucleotide positions refer to NCBI build 34. The 17 markers with $p < 0.05$ in either the case-control or the TDT analysis of replication panel B are highlighted in blue. SNPs with a significant result in both panel B tests are marked by grey shading. In addition to rs2241880, only SNPs rs1050152 (Leu503Phe) in the *OCTN1* gene, reported earlier (Peltekova *et al.*, 2004) and the known *CARD15* SNP rs2066845 ("SNP12", Hugot *et al.*, 2001) yielded consistent replication.

Combined analysis for the mutation rs2241180 in the *ATG16L1* gene

SNP rs2241880, which replicated besides known *CARD15* and *5q31* polymorphisms, is located in the ATG16 autophagy related 16-like 1 (*S. cerevisiae*) gene. In the combined analysis of all German individuals (panels A and B), the odds ratio was 1.45 (95% CI: 1.21 – 1.74) for heterozygous carriership of "G" and 1.77 (95% CI: 1.43 – 2.18) for homozygosity. The combined p-values for the German samples were $p = 4 \times 10^{-08}$ for the allele-based and $p = 2 \times 10^{-07}$ for the genotype-based test (see also 3.2.1 on page 89).

Genotyping results obtained with the SNPlex™ system were confirmed using a TaqMan® assay (C___9095577_20, 99.8% genotype concordance), thus excluding artefacts due to technological problems.

A summary of all results of the genome-wide scan is given in file `cSNP_Results.xls` on the DVD.

3.1.2 LD-based approach: 100k SNP array

116,161 SNPs were successfully genotyped in 399 controls and 393 cases (Panel C) using the Affymetrix 100k GeneChip® as described in section 2.8.5 on page 49. SNPs of which the position was annotated as "unknown" and SNPs that were located on the Y-chromosome, were not included in the analysis. Thus, a subset of 115,571 SNPs was analyzed, of which 92,631 (80.2%) were of good quality, i.e. they were non-monomorphic, had a call rate above 90%, and a HWE p-value greater than 0.01. Since Affymetrix did include repetitive regions in their chip design, much more SNPs were out of HWE than expected. According to Affymetrix, this inclusion was done for the following reasons:

- to provide users with an uniform genomic coverage
- SNPs in repeats might be used for the analysis of copy number variations (CNVs)
- technical restrictions of the method

With respect to the p-value obtained in a case-control comparison for genotypes (CCG), the top 150 SNPs were chosen for replication in panel D. The same SNPlex™ plates were used for follow-up as for the cSNP experiment (3.1.1 on page 80). Hit SNPs were defined as being "replicated", if the p-value was lower than 0.05 in the family-based test (TDT) and in the case-control comparison. A total of ten SNPs matched this criteria, from which five are located in yet unknown regions. In addition to rs2631372, rs272867 and rs2631370, which all localize to the 5q31 haploype, reported earlier by Peltekova *et al.* (2004), *CARD15* (Hugot *et al.*, 2001) yielded consistent replication (SNPs rs2076756 and rs10521209). SNP rs2076756, which had the lowest p-value ($<10^{-12}$) in this GWS, is located between exon 8 and 9 of *CARD15*, the latter exons coding for the LRR domain. Since both variants are in close proximity with the three important disease-causing variants (SNP 8, 12, and 13), the results of the GWS are likely to be consistent.

dbSNP ID	Position [bp]	Distance to next SNP [kb]	Comment
rs2066844	49,303,427	9,783	SNP 8, Arg702Trp
rs10521209	49,313,210	831	Hit #120
rs2066845	49,314,041	341	SNP 12, Gly908Arg
rs2076756	49,314,382	6,898	Hit #1
rs2066847	49,321,280	–	SNP 13, Leu1007fs

Table 3-2 Distance between known mutations and typed SNPs in the *CARD15* gene region. Results are based on HapMap data and positions refer to NCBI build 35.

Screening (panel C)							Replication (panel D)		
#	locus	dbSNP ID	chr.	position	pCCA	pCCG	pCCA	pCCG	pTDT
1	*CARD15, intron*	rs2076756	16	50,534,914	$1.9 \cdot 10^{-13}$	$2.0 \cdot 10^{-12}$	-	-	-
2	*BCL11B, downstream*	rs200354	14	97,365,362	0.003	$2.2 \cdot 10^{-06}$	0.59	0.39	1.00
3	*CARD8, intron*	rs2009530	19	53,411,500	0.002	$4.9 \cdot 10^{-06}$	0.73	0.83	0.52
4	*FLJ40089, intron*	rs10507063	12	94,769,766	0.08	$1.2 \cdot 10^{-05}$	0.13	0.06	0.70
5	*MRPS22, intron*	rs10513059	3	140,313,327	$2.1 \cdot 10^{-06}$	$1.6 \cdot 10^{-05}$	0.15	0.14	0.93
6	*ZNF452, upstream*	rs10484543	6	28,899,189	$9.4 \cdot 10^{-06}$	$3.6 \cdot 10^{-05}$	0.16	0.11	0.07
7	*BC036877, intron*	rs6659639	1	14,802,853	0.002	$5.2 \cdot 10^{-05}$	0.68	0.86	0.05
8	*ATRNL1, intron*	rs1590734	10	117,023,506	0.003	$5.6 \cdot 10^{-05}$	0.30	0.11	0.84
9	*FBXL13, intron*	rs10487283	7	102,031,744	0.16	$8.6 \cdot 10^{-05}$	0.37	0.09	0.42
10	*HPSE2, intron*	rs483952	10	100,048,586	0.005	$8.7 \cdot 10^{-05}$	0.59	0.71	0.79
11	*C18orf1, intron*	rs1559865	18	13,493,887	0.001	0.0001	0.05	0.15	0.30
12	*PRPF18, intron*	rs2478126	10	13,663,744	$2.7 \cdot 10^{-05}$	0.0001	0.30	0.29	0.83
13	*BC050321, intron*	rs10517460	4	37,864,263	0.01	0.0001	0.60	0.85	0.26
14	*RAD23B, upstream*	rs4978664	9	105,199,766	$1.7 \cdot 10^{-05}$	0.0001	0.57	0.33	0.66
15	*SLC7A11, downstream*	rs7674505	4	134,995,809	$2.6 \cdot 10^{-05}$	0.0001	0.56	0.38	0.13
16	*SLC7A11, downstream*	rs2724261	4	135,030,348	$4.0 \cdot 10^{-05}$	0.0002	0.41	0.15	0.06
17	*CDO1, downstream*	rs249721	5	115,098,473	0.006	0.0003	0.07	0.20	0.19
18	*DZIP1, intron*	rs7324781	13	93,932,754	$4.4 \cdot 10^{-05}$	0.0003	0.46	0.75	0.25
19	*NLGN1, downstream*	rs10490876	3	176,472,702	0.0004	0.0003	0.17	0.32	0.39
20	*LTF, upstream*	rs867619	3	46,498,342	0.04	0.0003	0.27	0.21	0.16
21	*BCL9, intron*	rs10494251	1	144,530,971	$9.8 \cdot 10^{-05}$	0.0004	0.32	0.23	0.59
22	*BC036403, intron*	rs6074780	20	14,701,879	$7.1 \cdot 10^{-05}$	0.0004	0.05	0.08	0.64
23	*QPCT, intron*	rs7582749	2	37,573,872	0.17	0.0004	0.91	0.97	0.80
24	*ZFP37, downstream*	rs747066	9	111,078,184	0.0002	0.0004	0.97	0.08	0.86
25	*SYT5, downstream*	rs3859540	19	60,373,858	$6.2 \cdot 10^{-05}$	0.0004	0.04	0.12	0.27
26	*KHDRBS3, downstream*	rs10505673	8	138,349,577	0.0007	0.0004	0.98	0.69	0.90
27	*GJA5, upstream*	rs1342705	1	144,761,333	0.0003	0.0004	0.12	0.28	0.73
28	*MYST4, intron*	rs7092435	10	76,014,914	0.0002	0.0004	0.79	0.22	0.56
29	*KIAA1912, intron*	rs10490395	2	56,422,800	0.004	0.0005	0.70	0.63	0.76
30	*GPC6, intron*	rs1333261	13	92,604,104	0.0002	0.0005	0.22	0.46	0.11
31	*HS3ST4, downstream*	rs10492804	16	26,420,874	0.002	0.0006	0.72	0.43	0.67
32	*BC035912, downstream*	rs1488927	7	153,794,409	0.002	0.0006	0.86	0.68	0.50
33	*NM_178506, upstream*	rs10514013	18	67,303,699	0.0008	0.0006	0.75	0.72	0.82
34	*CTSO, upstream*	rs10517634	4	157,722,979	0.0006	0.0007	0.44	0.72	0.21
35	*PXMP3, upstream*	rs723510	8	78,550,027	0.0006	0.0007	0.04	0.09	0.07
36	*SULF2, upstream*	rs6122682	20	47,749,305	0.29	0.0007	0.69	0.66	0.80
37	*BMPER, downstream*	rs1419790	7	34,517,746	0.54	0.0007	0.71	0.49	0.25
38	*CHORDC1, upstream*	rs10501720	11	89,950,836	0.0002	0.0008	0.82	0.69	0.43
39	*GRM5, intron*	rs549700	11	88,326,335	0.0009	0.0008	0.82	0.97	0.73
40	*MAP3K7, downstream*	rs4707614	6	91,077,734	0.16	0.0008	0.96	1.00	0.55
41	*BCL9, downstream*	rs10494248	1	144,615,337	0.0009	0.0008	0.67	0.65	0.34

Table 3-3 Lead SNPs of 100k genome-wide scan (GWS). The top 150 SNPs are ranked according to the p-value obtained in genotype-based case-control comparison (CCG) in panel C. Also included are the p-values for the allele-based case-control comparison. Nucleotide positions refer to NCBI build 34. Markers with $p < 0.05$ in either the case-control or the TDT analysis in replication panel D are highlighted in blue. SNPs with a significant result in both panel D tests are additionally marked by grey shading. The five SNPs from the two known disease loci *CARD15* and 5q31 are highlighted in red.

Screening (panel C)							Replication (panel D)		
#	locus	dbSNP ID	chr.	position	pCCA	pCCG	pCCA	pCCG	pTDT
42	*API5, upstream*	rs10501292	11	42,960,696	0.01	0.0008	0.67	0.52	0.01
43	*GJA5, upstream*	rs10494257	1	144,761,646	0.0005	0.0008	0.16	0.35	0.84
44	*CRYBA4, downstream*	rs569626	22	26,071,824	0.002	0.0008	0.13	0.32	0.33
45	*SLC7A11, downstream*	rs10518641	4	134,981,546	0.0003	0.0008	0.38	0.38	0.06
46	*JAG1, upstream*	rs973542	20	10,687,559	0.0003	0.0008	0.57	0.21	0.61
47	*C10orf97, intron*	rs10508494	10	15,825,784	0.004	0.0008	0.28	0.33	0.70
48	*JAG1, upstream*	rs6077888	20	10,687,989	0.0003	0.0008	0.57	0.21	0.61
49	*GNG2, intron*	rs10498440	14	50,403,611	0.0002	0.0008	0.12	0.19	0.50
50	*RAB6B, intron*	rs7644124	3	134,896,774	0.0003	0.0008	0.28	0.18	0.02
51	*PPP2R2B, intron*	rs3096085	5	146,074,928	0.001	0.0008	0.44	0.75	0.64
52	*NDN, upstream*	rs10519449	15	21,597,566	0.22	0.0009	0.79	0.64	0.38
53	*LOC339047, downstream*	rs1994893	16	17,879,139	0.007	0.0009	0.62	0.23	0.33
54	*SLC1A2, intron*	rs3847621	11	35,328,744	0.77	0.0009	0.60	0.75	0.82
55	*DSCAM, intron*	rs2837758	21	40,939,405	0.06	0.0009	0.61	0.83	0.55
56	*KIAA1463, upstream*	rs10506296	12	49,374,124	0.001	0.001	0.10	0.18	0.85
57	*LOC151963, intron*	rs2367129	3	193,916,220	0.0002	0.001	0.96	0.91	0.76
58	*FLJ11175, downstream*	rs10520778	15	93,657,496	0.003	0.001	0.63	0.18	0.34
59	*CYLC2, upstream*	rs4743487	9	99,916,141	0.0002	0.001	0.01	0.05	0.80
60	*C21orf7, downstream*	rs2832238	21	29,469,808	0.0008	0.001	0.83	0.63	1.00
61	*NR3C2, upstream*	rs853727	4	150,412,639	0.0002	0.001	0.68	0.85	0.58
62	*NM_030903, downstream*	rs9257453	6	29,076,919	0.0003	0.001	0.38	0.29	0.03
63	*POU3F1, upstream*	rs10493084	1	38,577,447	0.0001	0.001	0.90	0.38	0.88
64	*SLC7A11, downstream*	rs2660693	4	135,078,432	0.0002	0.001	0.87	0.98	0.48
65	*NEBL, intron*	rs10508629	10	20,881,787	0.001	0.001	0.23	0.34	0.04
66	*DKFZP564O0823, downstream*	rs7681167	4	76,619,617	0.0003	0.001	0.89	0.72	0.35
67	*PDCD6IP, downstream*	rs7428001	3	35,500,582	0.0004	0.001	0.84	0.97	0.01
68	*CLUL1, intron*	rs10502288	18	621,435	0.0003	0.001	0.69	0.66	0.08
69	*MAML2, intron*	rs10501843	11	95,528,273	0.0002	0.001	0.48	0.55	0.73
70	*SLC22A4, downstream*	rs2631372	5	131,779,794	0.26	0.001	–	–	–
71	*FGL1, downstream*	rs396462	8	17,680,469	0.0007	0.001	0.46	0.46	0.03
72	*ITGB6, upstream*	rs2925757	2	161,303,713	0.0002	0.001	0.004	0.02	1.00
73	*CYLC2, upstream*	rs1930551	9	100,720,179	0.002	0.001	0.36	0.49	0.50
74	*SNX7, upstream*	rs770918	1	98,391,860	0.0009	0.001	0.74	0.54	0.44
75	*MYST4, intron*	rs951308	10	75,962,082	0.0007	0.002	0.62	0.56	0.85
76	*BC036403, intron*	rs10485515	20	14,844,853	0.0008	0.002	0.33	0.53	0.85
77	*CSMD1, intron*	rs10503282	8	4,626,114	0.0003	0.002	0.98	0.86	0.84
78	*PAK1, upstream*	rs1793483	11	76,701,775	0.0005	0.002	0.41	0.71	0.87
79	*SCG2, downstream*	rs9283536	2	224,133,045	0.0007	0.002	0.24	0.42	0.005
80	*B3GALT2, upstream*	rs1160521	1	191,026,538	0.003	0.002	0.65	0.57	0.21
81	*SLCO1B3, intron*	rs3764007	12	20,905,430	0.0005	0.002	0.73	0.91	0.34
82	*OR5V1, downstream*	rs10484545	6	29,342,503	0.0005	0.002	0.0003	$2.8 \cdot 10^{-07}$	0.16

Table 3-3 Lead SNPs of 100k genome-wide scan (GWS). The top 150 SNPs are ranked according to the p-value obtained in genotype-based case-control comparison (CCG) in panel C. Also included are the p-values for the allele-based case-control comparison. Nucleotide positions refer to NCBI build 34. Markers with $p < 0.05$ in either the case-control or the TDT analysis in replication panel D are highlighted in blue. SNPs with a significant result in both panel D tests are additionally marked by grey shading. The five SNPs from the two known disease loci *CARD15* and 5q31 are highlighted in red.

Screening (panel C)							Replication (panel D)		
#	locus	dbSNP ID	chr.	position	pCCA	pCCG	pCCA	pCCG	pTDT
83	*LOC253827, intron*	rs10506523	12	64,122,774	0.002	0.002	0.36	0.36	0.65
84	*MAP3K7, upstream*	rs1546210	6	92,572,189	0.0004	0.002	0.86	0.58	0.19
85	*FHIT, downstream*	rs812965	3	59,693,634	0.0005	0.002	0.30	0.59	0.30
86	*PTGER4, upstream*	rs1992662	5	40,439,353	0.0005	0.002	$7.6 \cdot 10^{-05}$	0.0002	0.001
87	*TSN, downstream*	rs10496586	2	123,281,730	0.0004	0.002	0.27	0.52	0.32
88	*SPOCK, intron*	rs1859346	5	136,523,790	0.0007	0.002	0.55	0.69	0.25
89	*SLC22A4, downstream*	rs272867	5	131,757,273	0.15	0.002	–	–	–
90	*CD36, intron*	rs3211830	7	79,891,769	0.0003	0.002	0.34	0.58	0.59
91	*C1D, downstream*	rs2902041	2	68,089,432	0.002	0.002	0.19	0.40	0.20
92	*CDH10, upstream*	rs2928266	5	25,827,114	0.0003	0.002	0.23	0.37	0.39
93	*NELL1, intron*	rs1793004	11	20,663,238	0.0005	0.002	0.03	0.06	0.03
94	*SLC14A2, upstream*	rs10502861	18	41,052,135	0.0005	0.002	0.32	0.22	0.86
95	*PTGER4, upstream*	rs1992660	5	40,460,568	0.0005	0.002	$4.5 \cdot 10^{-05}$	0.0002	0.0005
96	*BC030984, intron*	rs10483112	22	21,051,522	0.0004	0.002	0.03	0.02	0.37
97	*LPHN2, intron*	rs3790889	1	81,835,193	0.43	0.002	0.86	0.99	0.74
98	*PTGER4, upstream*	rs1553575	5	40,548,433	0.0005	0.002	$1.7 \cdot 10^{-06}$	$6.4 \cdot 10^{-06}$	0.03
99	*TMEM16A, intron*	rs3781661	11	69,725,941	0.002	0.002	0.26	0.53	0.21
100	*AK091523, intron*	rs4327488	4	103,117,607	0.0005	0.002	0.09	0.04	0.93
101	*NM_018461, upstream*	rs3922590	10	132,706,022	0.0005	0.002	0.55	0.75	0.19
102	*BCHE, upstream*	rs1440657	3	167,262,663	0.003	0.002	0.88	0.90	0.09
103	*GRM8, downstream*	rs6947579	7	125,087,495	0.0005	0.002	0.006	0.01	1.00
104	*SLC7A11, downstream*	rs1493517	4	134,930,683	0.0005	0.002	0.23	0.48	0.11
105	*GLI3, downstream*	rs10486725	7	41,613,489	0.0007	0.002	0.72	0.26	0.93
106	*BCL9, downstream*	rs1953977	1	144,615,549	0.003	0.002	0.91	0.91	0.34
107	*AK095544, upstream*	rs10489924	1	98,921,897	0.0007	0.002	0.65	0.57	0.03
108	*PPP3R2, intron*	rs4743484	9	99,859,318	0.0008	0.003	0.002	0.007	0.90
109	*UBAP2, intron*	rs1785512	9	34,016,538	0.002	0.003	0.003	0.004	1.00
110	*CHORDC1, upstream*	rs562085	11	89,840,073	0.006	0.003	0.59	0.05	0.23
111	*PDE8A, intron*	rs289386	15	83,311,157	0.0006	0.003	0.87	0.61	0.91
112	*PDE1C, downstream*	rs7795363	7	31,529,121	0.0007	0.003	0.34	0.03	0.71
113	*NELL1, intron*	rs951199	11	20,665,084	0.0008	0.003	0.03	0.06	0.27
114	*SLC7A11, downstream*	rs1559781	4	138,684,038	0.0004	0.003	0.65	0.90	0.65
115	*ADAMTS20, downstream*	rs2134066	12	42,000,804	0.0008	0.003	0.24	0.30	0.07
116	*KCND2, upstream*	rs10500069	7	118,069,959	0.0008	0.003	0.54	0.81	0.90
117	*LOC137886, upstream*	rs10504252	8	59,267,687	0.0007	0.003	0.85	0.97	0.07
118	*PAH, downstream*	rs1463446	12	101,488,216	0.0005	0.003	0.32	0.60	0.74
119	*NPAS3, intron*	rs8005131	14	31,581,144	0.0004	0.003	0.44	0.35	0.57
120	*CARD15, intron*	rs10521209	16	50,533,742	0.001	0.003	–	–	–
121	*GRIA1, intron*	rs778999	5	152,936,197	0.001	0.003	0.02	0.04	0.14
122	*TEK, intron*	rs1413829	9	27,191,968	0.0006	0.003	0.27	0.45	0.43
123	*IMMP2L, intron*	rs1528039	7	110,003,303	0.0009	0.003	0.75	0.12	0.30
124	*SIAT8B, upstream*	rs10520704	15	90,450,368	0.0008	0.003	0.02	0.04	0.08
125	*CHI3L2, downstream*	rs4323713	1	111,121,947	0.0005	0.003	0.60	0.87	0.43

Table 3-3 Lead SNPs of 100k genome-wide scan (GWS). The top 150 SNPs are ranked according to the p-value obtained in genotype-based case-control comparison (CCG) in panel C. Also included are the p-values for the allele-based case-control comparison. Nucleotide positions refer to NCBI build 34. Markers with $p < 0.05$ in either the case-control or the TDT analysis in replication panel D are highlighted in blue. SNPs with a significant result in both panel D tests are additionally marked by grey shading. The five SNPs from the two known disease loci *CARD15* and 5q31 are highlighted in red.

Screening (panel C)							Replication (panel D)		
#	locus	dbSNP ID	chr.	position	pCCA	pCCG	pCCA	pCCG	pTDT
126	*FLJ11588, intron*	rs3118223	1	48,833,718	0.0008	0.003	0.03	0.08	0.47
127	*AY320401, downstream*	rs6673693	1	206,802,809	0.0006	0.003	0.39	0.42	0.73
128	*NOVA1, downstream*	rs9323605	14	23,606,664	0.0008	0.003	0.21	0.38	0.62
129	*A2BP1, upstream*	rs7200548	16	5,890,864	0.001	0.003	0.78	0.96	0.30
130	*CYLC2, upstream*	rs10512301	9	100,701,999	0.001	0.003	0.46	0.60	1.00
131	*MIPEP, intron*	rs1536299	13	22,116,963	0.0006	0.004	0.06	0.13	0.15
132	*ITGB6, intron*	rs10497212	2	161,167,244	0.0008	0.004	0.41	0.46	0.64
133	*GRIA1, intron*	rs6580022	5	152,962,446	0.008	0.004	0.01	0.02	0.31
134	*TUSC3, upstream*	rs10503536	8	15,023,600	0.001	0.004	0.47	0.72	0.90
135	*LARGE, upstream*	rs130461	22	32,734,453	0.001	0.004	0.76	0.94	0.23
136	*DGKB, upstream*	rs10255000	7	14,785,866	0.0009	0.004	0.02	0.09	0.32
137	*SYT1, upstream*	rs3911786	12	77,767,173	0.0005	0.004	0.65	0.62	0.06
138	*SLC22A4, downstream*	rs2631370	5	131,779,992	0.14	0.004	–	–	–
139	*DNAH11, intron*	rs6461599	7	21,544,775	0.0007	0.004	0.90	0.67	0.02
140	*OXTR, upstream*	rs2241330	3	8,832,963	0.001	0.004	0.12	0.17	0.79
141	*VRK1, downstream*	rs8011773	14	95,847,773	0.0005	0.004	0.06	0.08	0.37
142	*KUB3, downstream*	rs951901	12	56,684,632	0.001	0.004	0.41	0.46	0.17
143	*TMEM16F, downstream*	rs10492248	12	44,329,880	0.001	0.004	0.11	0.22	0.23
144	*BC036480, downstream*	rs977669	5	87,946,906	0.004	0.004	0.28	0.28	0.78
145	*RNF111, intron*	rs1446240	15	57,012,547	0.001	0.004	0.12	0.31	0.51
146	*AY134745, downstream*	rs9304344	18	43,010,652	0.001	0.004	0.98	0.05	0.74
147	*PDE4D, intron*	rs1027747	5	59,108,797	0.002	0.004	0.67	0.85	0.45
148	*PPP1R3A, upstream*	rs10500037	7	113,262,290	0.0009	0.005	0.01	0.04	0.04
149	*NEBL, downstream*	rs10508621	10	20,650,445	0.0008	0.005	0.91	0.99	0.03
150	*PTGER4, upstream*	rs7725523	5	40,417,724	0.0007	0.005	0.14	0.32	0.01

Table 3-3 Lead SNPs of 100k genome-wide scan (GWS). The top 150 SNPs are ranked according to the p-value obtained in genotype-based case-control comparison (CCG) in panel C. Also included are the p-values for the allele-based case-control comparison. Nucleotide positions refer to NCBI build 34. Markers with $p < 0.05$ in either the case-control or the TDT analysis in replication panel D are highlighted in blue. SNPs with a significant result in both panel D tests are additionally marked by grey shading. The five SNPs from the two known disease loci *CARD15* and 5q31 are highlighted in red.

The five replicated SNPs, that are not restricted to either the *CARD15* or the 5q31 region, are located in *NELL1*, and in the gene-free regions on chromosome 5p13.1 and 7q31.1. The latter was not subjected to further experiments due to borderline replication ($p = 0.04$). The signal at the *NELL1* locus is supported by 2 SNPs (#93 and #113) among the top 150, from which one is only significant in the case-control analysis (#113, rs951199) and not in the TDT. With three replicated SNPs (#86, #95, #98) and one semi-replicated SNP (#150, only significant in case-control sample) among the top 150, the 5p13.1 region is standing out.

A summary of all results is given in file `100k_Results.xls` on the DVD.

3.1.2.1 Randomization of the 100k dataset

An estimate of the rate of false-positive association findings was obtained from GENOMIZER by generating replicates of the entire study under the null hypothesis of no association. To this end, the affection status of cases and controls was randomly shuffled over the entire study population. The same analyses were performed as for the non-randomized dataset and results compared.

Because neither the position, nor the penetrance of the potential new disease variants underlying association signals is known, the single-point analyses results for CCA and CCG were linked with a logical "OR". A given marker was considered a potential "lead" marker if a nominal p-value < 0.001 was reached in either analysis. If multiple such "signals" were identified adjacent to each other, markers were merged into "associated regions" if the distance was below 50 kilobases.

Using these criteria, a total of 272 markers, corresponding to 245 associated regions were identified at the $p < 0.001$ level. Analysis of the randomized dataset at $p < 0.001$ yielded 112 markers corresponding to 94 regions, suggesting the presence of "real" disease driven association signals. This means in other words that approximately 59% of the hits should correspond to "real" hits.

3.2 Replication of leads

SNPs from the GWSs that replicated in two independent German sample collections (trios and case-controls), were tested in other ethnical panels.

3.2.1 Replication of the *ATG16L1* nsSNP in a UK panel

The association of rs2242880 was replicated in a UK-derived CD sample (table 2-1, page 26, panel F: $n_{cases} = 509$, $n_{controls} = 656$), using an independent TaqMan® assay. The British data yielded p = 0.0004 in the allele-based comparison (OR: 1.35; 95% CI: 1.15 – 1.59) and p = 0.0001 in the genotype-based test (OR for homozygous carriership of "G": 1.71; 95% CI: 1.23 – 2.39). This finding strongly supports a consistent association between rs2241880 and CD. Table 3-4 summarizes the results for all panels.

Panel	MAF_{co}	MAF_{ca}	OR_{HOM}	p_{CCA}	p_{CCG}
A - German screening panel	0.48	0.40	1.92 [95% CI (1.33 – 2.79)]	0.0004	0.002
B - German replication panel	0.47	0.40	1.74 [95% CI (1.34 – 2.25)]	0.00001	0.00007
A + B Combined German	0.47	0.40	1.77 [95% CI (1.43 – 2.18)]	4×10^{-08}	2×10^{-07}
E - German UC panel	0.47	0.46	1.11 [95% CI (0.85 – 1.44)]	0.40	0.59
F - UK replication panel	0.48	0.41	1.71 [95% CI (1.23 – 2.39)]	0.0004	0.0002

Table 3-4 Summary of association results for rs2241880. MAF is the minor allele frequency (A allele) in controls (co) or cases (ca). It is apparent that CD patients have a 7% higher frequency of the mutant "G" allele while the nsSNP is not associated with UC.

The analyses point out that CD patients have a 7% higher frequency of the risk allele "G". There is even a 9% increase of the risk genotype "GG" in the patients panel. Furthermore, no significant increase was seen in the IBD subphenotype UC.

Panel	$f_{AA}co$	$f_{AA}ca$	$f_{AG}co$	$f_{AG}ca$	$f_{GG}co$	$f_{GG}ca$
A - German screening panel	0.22	0.15	0.52	0.49	0.26	0.36
B - German replication panel	0.22	0.16	0.51	0.48	0.27	0.36
A + B combined German	0.22	0.16	0.51	0.48	0.27	0.36
E - German UC panel	0.22	0.22	0.51	0.49	0.27	0.29
F - UK replication panel	0.22	0.19	0.53	0.44	0.25	0.37

Table 3-5 Summary of genotype frequencies for rs2241880. CD patients have a 9% increased frequency of the risk genotype "GG", while there is no difference in the UC panel.

3.2.2 Replication of the *NELL1* and 5p13.1 lead SNPs in independent panels

In analogy to *ATG16L1*, the replicated SNPs of the Affymetrix scan were typed in the same UK CD sample (panel F). To this end, two SNPs of the 5p13.1 and one from the *NELL1* region, all replicating in the case-control and family-based panel, were selected for genotyping. In addition, the *NELL1* SNP rs951199 was genotyped, which was only significant in the case-control panel. TaqMan® was used as an independent technology (assays rs951199, C____392093_10, C__11472026_10, C__11472042_10) and the results are shown in the following table:

Region/gene	# in GWS	dbSNP ID	Celera ID	MAF	HWE	CD p_{CCA}	CD p_{CCG}
NELL1	113	rs951199	hCV8886311	0.26	0.32	0.62	0.63
NELL1	93	rs1793004	hCV392093	0.26	0.51	0.35	0.63
5p13.1	95	rs1992660	hCV11472026	0.40	0.55	0.04	0.09
5p13.1	86	rs1992662	hCV11472042	0.30	0.56	0.001	0.0004

Table 3-6 Replication of *NELL1* and 5p13.1 lead SNPs in a UK CD panel. Minor allele frequencies (MAF) and Hardy-Weinberg p-values (HWE) plus single-point p-values for an allelic (CCA) and a genotypic (CCG) test are shown.

While SNPs rs1992660 and rs1992662 of the 5p13.1 region were replicated in the British population, SNPs rs951199 and rs1793004 failed to show a significant association. Allele and genotype frequencies of the four variants are listed in the following table:

	controls			cases			controls		cases	
dbSNP ID	f_{11}	f_{12}	f_{22}	f_{11}	f_{12}	f_{22}	f_1	f_2	f_1	f_2
rs951199	0.541	0.400	0.059	0.566	0.370	0.064	0.741	0.259	0.751	0.249
rs1793004	0.542	0.396	0.062	0.572	0.374	0.055	0.740	0.260	0.759	0.241
rs1992660	0.368	0.467	0.165	0.415	0.466	0.119	0.601	0.399	0.648	0.352
rs1992662	0.489	0.413	0.098	0.437	0.377	0.185	0.696	0.304	0.626	0.374

Table 3-7 Genotype and allele frequencies of the *NELL1* and 5p13.1 lead SNPs. Frequencies are listed for genotypes and alleles in both controls and cases. Significant differences >5% are highlighted in blue.

Although the single-point statistics for rs1793004 gave no significant results, there are 3.0% more homozygous cases for the common allele. The same effect is seen in the German case sample (panel D), where 4.6% more homozygous cases exist for rs1793004.

SNP rs1793004, located in the *NELL1* gene, was also replicated in a French-Canadian trio sample. The TDT test gave a significant p-value of p = 0.02. More detailed results are given in table 9-4, page 185.

3.3 Mutation detection and fine mapping of candidate genes and regions

Promoters and all exons plus flanking intronic sequences of *NELL1* and *ATG16L1* were resequenced in 47 CD patients (plate MD6) as described in section 2.7 on page 34. This was done to verify the previously found mutations and to find new ones, especially SNPs with a putative functional effect. Since the number of sequenced individuals allowed only the detection of rather common mutations, private mutations, which are usually found only in a single family or a small population, were not accessible by this approach,

Power calculations (according to Glatt *et al.*, 2001)

The probability of detecting a variant with allele frequency f in a sample of n individuals (2n chromosomes) was calculated as $1-(1-f)^{2n}$ (for seeing it once) or $1-(1-f)^{2n}-2nf(1-f)2^{n-1}$ (for seeing it at least twice). Therefore, the probability to detect a variant with f = 0.05 and n = 47 was 99% to detect it once and 95% to detect it twice in our experiment. This high power is not achieved for rare SNPs, e.g. for f = 0.01 the probability is only 1%.

3.3.1 *ATG16L1* fine mapping

For a more comprehensive assessment of the CD risk conferred by changes in the *ATG16L1* gene, a systematic search for mutations was carried out by resequencing all exons, splice sites and the promoter regions of 47 CD patients and by fine mapping the gene region.

3.3.1.1 Resequencing of *ATG16L1*

Sequences were retrieved from the UCSC Golden Path (contig NT_005120) and used for primer design as described in section 2.7.1 on page 34. *ATG16L1* comprises 18 exons which give rise to a coding sequence of 1821 base pair length (607 aa).

Sequencing was successful for all exons and the promoter of *ATG16L1*. Twenty-two SNPs were found in the analysis of the traces, from which 12 were yet unknown variants. Detailed results are listed in table 3-8, page 92. The common variant rs2241880 was detected as well in several traces (37% GG, 43% GA, 20% AA), thus verifying the existence of this biallelic polymorphism. Besides rs2241880 in exon 9, no other non-synonymous SNP was found in *ATG16L1*. Only the synonymous coding SNP rs13011156 was verified in exon 3.

#	SNP ID	position (build 35)	SNP type	sequence
1	ATG16L1_01	233,941,476	Promoter	AGAAACAAGAAAGAAACACGAAAAGCAGC TTAACAATCAAAGACAGGTTTATTTTG-GAGAATAAACCTGAGAGGGGCTTCT-TGTCGATTTTGGTCAGGAG **[C/T]** GTTTTCTCTTACACACTAAGGGTGTTTAA GGGTTTACGAAGCGGTGAGCTTATTG-CAGGTTCGTAATGTTTCTGTATGAGG-GAAAGTTTATTGCAGGGTT
2	ATG16L1_02	233,941,591	Promoter	ACTAAGGGTGTTTAAGGGTTTACGAAGCG GTGAGCTTATTGCAGGTTCGTAAT-GTTTCTGTATGAGGGAAAGTTTATTG-CAGGGTTGGAAAGTCTCTGGC **[C/T]** GGAGGGGAGGCTATCTCTGGGTTGGCAT-GTTTCTGGTTGGAAGTGGGTTTATCTTAG GGTTGGAATGTTTCTGGTTATGCTGA-CATTAGCCATTAGGCTG
3	rs1816753	233,941,609	Promoter	–
4	rs12476635	233,941,685	Promoter	–
5	ATG16L1_03	233,941,758	Promoter	GTTTCTGGTTATGCTGACATTAGCCATT-AGGCTGATGTTTTTGGGTTGGATTTAGGC AGTTTCTTAATCAAGGGGAACTTAAAATG-GTGGCAGTTGTCCA **[A/T]** GATGGCAATGCTCCTGCTGTCAGGGTAG-GCCTTGGGAAGACACCAGGTGTCTTTGTA AACTTAGGTTTCCCCTGTACCAAGCA-CAGTGCTGACTGCATTA
6	ATG16L1_04	233,946,704	intronic	TCTTGCAGGTGATTCTGTAGGTTACTGTG GCTGAGGAATGTATGTTCCTGTATTG-CATCCTTCAATACATGACAAGGTAGGTAC-CTAGCAAGTGTACATA **[C/T]** GTTGTAATAAATAACTTAGTTTCTGACTTT TTTATTTTAATAGATAACAAATTGCTG-GAAAAGTCAGATCTTCATTCAGTGTTG-GCCCAGAAACTACAGG
7	rs13011156	233,953,812	synonymous	–
8	rs12994997	233,955,503	intronic	–
9	rs2289476	233,963,556	intronic	–
10	rs2289475	233,963,593	intronic	–
11	rs2289472	233,964,240	intronic	–
12	rs2241880	233,965,368	non-synonymous T→A	–
13	rs2241879	233,965,468	intronic	–

Table 3-8 Results of mutation detection of *ATG16L1*. 12 hitherto unknown polymorphisms were found and 10 annotated SNPs were verified. The latter includes the nsSNP rs2241880. Primer sequences are listed in table table 8-2, page 162.

#	SNP ID	position (build 35)	SNP type	sequence
14	ATG16L1_05	233,980,474	intronic	GTTCCTGGCCGGATCTCAGGGTGGTCT-GACACTCTGACTCTGGAGGTAAGGATGCT GTGGATACTTTGCCAGCATAGGCAGGGC-CTGGCGCCCTGAGGCT **[C/T]** ACCCTGGCTGTGTTCTCTCTCCTAGCACA CACTCACGGGACACAGTGGGAAAGTGCT-GTCTGCTAAGTTCCTGCTGGACAAT-GCGCGGATTGTCTCAGGA
15	ATG16L1_06	233,980,852	intronic	CTGAGGATAGTGATAGTTTTTCTTGTTTA AAGCTTCATTTAAGTGAGTAACTCTGA-CAAGTCAGTGGTTAGAGGGTGGCAGCAT-TATGTAAACAGCCACG **[T/A]** TGGTGCCTCTGCTTGATTAATGATGTTTG CATTTCTTTCAGGCATAAAGACAGT-GTTTGCAGGATCCAGTTGCAATGATAT-TGTCTGCACAGAGCAATGT
16	ATG16L1_07	233,981,059	intronic	AGTGGACATTTTGACAAGAAAATTCGTTT CTGGGACATTCGGTATGATAC-CCAAGCTCCTGACTGGAGGCACATAA-GAGTCTCCACAGTAATGGTTCTGT **[A/G]** CATGGGTTGTGCTTTTAAGATCTCAGGA-CATGGCAGAAATGAGTTTCGGTACACGTA TAGTGTTAGCTTATTAACACTTGGCTGT-TCTTTCCCATGAAGT
17	ATG16L1_08	233,982,979	intronic	AAAAGTTATTGATCTCCGAACAAATGCTA TCAAGCAGACATTCAGGTAACTGAAGAT-GTGCTGGTTGCATGAAGACCAGAGGC-CCAGCCCTGCTCTCTTA **[A/–]** CGTGCTGCGTGGCCTGAGCNCCCTCAGT GACAGTGCCCTGTTGTTTGTTTCAGTG-CACCTGGGTTCAAGTGCGGCTCTGACTG-GACCAGAGTTGTCTTCA
18	ATG16L1_09	233,982,999	intronic	CAAATGCTATCAAGCAGACATTCAGGTAA CTGAAGATGTGCTGGTTGCATGAAGAC-CAGAGGCCCAGCCCTGCTCTCTTANCGT-GCTGCGTGGCCTGAGC[**T/A]** CCCTCAGTGACAGTGCCCTGTTGTTTGTT TCAGTGCACCTGGGTTCAAGTGCG-GCTCTGACTGGACCAGAGTTGTCTTCAG-GTTAGTAGTGACCCACCCC
19	ATG16L1_10	233,983,767	intronic	ACTTGGAGCCCCAAGAAGCAATGTGATGC CACTGTTCATTAGCCGGACGGGGCT-GAAATACTGGGGAAAAGCAGATTTG-GCTGGAAGGGCCATGACTAGC **[A/G]** CACATTCTGAGGTCTGGGTACACAGGAGT GGTCATCAGTGAAGAACAAACATTCCT-TCTTGGGAAAAAGGCCACAGAGATGCT-GAATTGGGGGTGGGGAA

Table 3-8 Results of mutation detection of *ATG16L1*. 12 hitherto unknown polymorphisms were found and 10 annotated SNPs were verified. The latter includes the nsSNP rs2241880. Primer sequences are listed in table table 8-2, page 162.

#	SNP ID	position (build 35)	SNP type	sequence
20	ATG16L1_11	233,984,036	intronic	GAGGGCTCTCTGTATATCTGGAGTGTGCT CACAGGGAAAGTGGAAAAGGT- TCTTTCAAAGCAGCACAGGTAAGAT- GAACCCTGTCTCAGCCCCCCGTTGC **[G/A]** TTGTGAGCAAGGCCTTTGACTTCATCTCA GGGGTCATCCGGTTTAGACCTCAGTCG- GCGCTGTGAGGGCACTGTCCGCCCACCT- GCTCGGCTGGCTGAGC
21	rs6748547	233,984,766	intronic	–
22	ATG16L1_12	233,985,029	3' UTR	CGCCCTCTGGCTCGCACGTTGTCAGTGT GGACAAAGGATGCAAAGCTGTGCT- GTGGGCACAGTACTGACGGGGCTCT- CAGGGCTGGGAGGACCCCAGTGC **[C/T]** CTCCTCAGAAGAAGCACATGGGCTCCTG- CAGCCCTGTCCTGGCAGGTGATGTGCTG GGTATAGCATGGACCTCCCAGAGAAGCT- CAAGCTATGTGGCACT

Table 3-8 Results of mutation detection of *ATG16L1*. 12 hitherto unknown polymorphisms were found and 10 annotated SNPs were verified. The latter includes the nsSNP rs2241880. Primer sequences are listed in table table 8-2, page 162.

3.3.1.2 *ATG16L1* linkage disequilibrium analysis

The CD risk associated with *ATG16L1* variation was then also analysed at the haplotype level using 28 tagging SNPs selected from the Caucasian HapMap data on the basis of $r^2 > 0.8$ and a minor allele frequency >1% (see also 2.8.3 on page 48). The location of the respective SNPs and their linkage disequilibrium structure is shown in fig. 3-1.

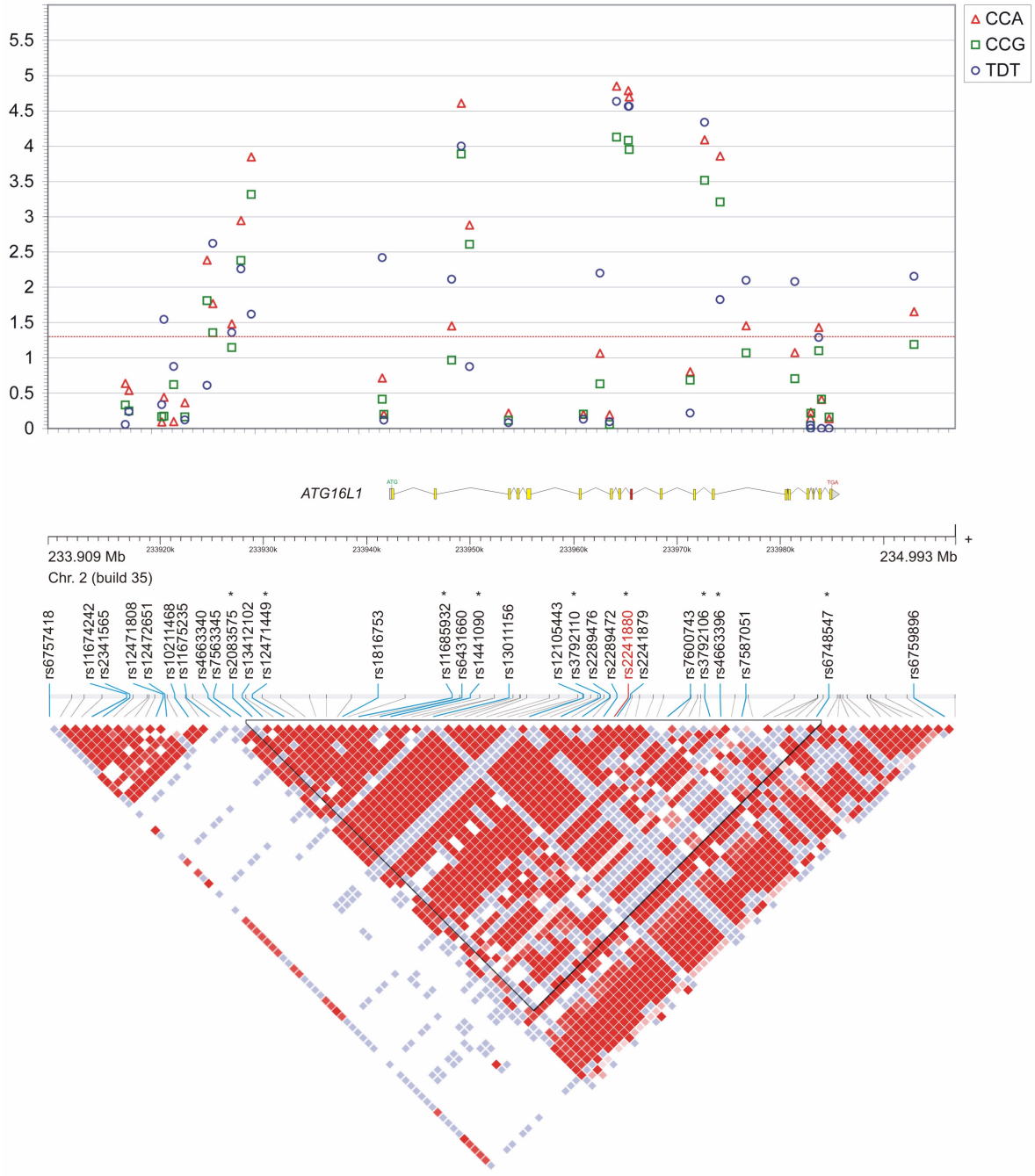

Fig. 3-1 Overview of the physical and genetic structure of the *ATG16L1* gene region. Along the y-axis negative $\log_{10}$ of corresponding p-values are plotted in the upper panel. For detailed results see table 3-9, page 96. Physical positions of the SNPs investigated and a schematic chart of the gene structure are shown in the middle panel. The only coding SNP is marked in red. The coordinates refer to the genome assembly build 35. The lower panel gives an overview of the LD structure of the locus (D') as generated by Haploview (Barrett *et al.*, 2005) from the Caucasian HapMap data. Fields are color-coded according to LD strength (white, D'=0.0; red, D'=1). The SNPs used in the haplotype analysis (table 3-10, page 97) are marked with asterisks.

When the tagging SNPs were genotyped in panel B (results are shown in table 3-9), the intronic SNP rs2289472 was found to have the same minor allele frequency (0.47) as the coding SNP rs2241880, and to yield a slightly more significant disease association ($p = 1.4 \times 10^{-05}$). This variant is located 1082 bases upstream of exon 9 and is not located in any recognizable regulatory motif. Synonymous SNP rs13011156, on the other hand, was not significantly associated with CD.

In a logistic regression model, none of the tagging SNPs significantly improved the model fit in the presence of rs2241880 (all $p > 0.05$). Together with results of a subsequent haplotype analysis (table 3-10, page 97), these findings imply that the CD risk at *ATG16L1* gene variation is indeed mainly due to carriership of susceptibility allele "G" at rs2241880.

#	SNP ID	Position [bp]	MAF_{co}	MAF_{ca}	HWE_{co}	P_{CCG}	P_{CCA}	P_{TDT}
1	rs6757418	233,909,321	0.16	0.16	0.13	0.56	0.95	0.42
2	rs11674242	233,916,795	0.12	0.14	0.59	0.47	0.23	0.88
3	rs2341565	233,917,138	0.15	0.16	0.33	0.57	0.29	0.58
4	rs12471808	233,920,324	0.46	0.46	0.10	0.68	0.81	0.46
5	rs12472651	233,920,522	0.41	0.39	0.24	0.67	0.36	0.03
6	rs10211468	233,921,467	0.43	0.43	0.02	0.24	0.80	0.13
7	rs11675235	233,922,554	0.13	0.14	0.12	0.69	0.43	0.76
8	rs4663340	233,924,691	0.08	0.05	0.66	0.02	0.004	0.25
9	rs7563345	233,925,244	0.34	0.31	0.88	0.04	0.02	0.002
10	rs2083575	233,927,080	0.07	0.05	0.36	0.07	0.03	0.04
11	rs13412102	233,927,971	0.41	0.36	0.96	0.004	0.001	0.006
12	rs12471449	233,928,958	0.14	0.10	0.16	0.0005	0.0001	0.02
13	rs11685932	233,948,322	0.33	0.30	0.94	0.11	0.04	0.008
14	rs6431660	233,949,234	0.47	0.40	0.79	0.0001	0.00002	0.0001
15	rs1441090	233,950,042	0.07	0.05	0.99	0.002	0.001	0.13
16	rs13011156	233,953,812	0.05	0.06	0.99	0.77	0.61	0.83
17	rs12105443	233,961,028	0.01	0.01	0.84	0.63	0.63	0.74
18	rs3792110	233,962,638	0.28	0.25	0.63	0.23	0.09	0.006
19	rs2289476	233,963,556	0.06	0.05	0.89	0.88	0.64	0.80
20	rs2289472	233,964,240	0.47	0.41	0.69	0.00007	0.00001	0.00002
21	rs2241880	233,965,368	0.47	0.40	0.59	0.00008	0.00002	0.00003
22	rs2241879	233,965,468	0.47	0.40	0.82	0.0001	0.00002	0.00003
23	rs7600743	233,971,359	0.06	0.05	0.18	0.21	0.16	0.61
24	rs3792106	233,972,740	0.41	0.35	0.54	0.0003	0.00008	0.00005
25	rs4663396	233,974,251	0.20	0.15	0.74	0.0006	0.0001	0.02
26	rs7587051	233,976,755	0.34	0.31	0.48	0.09	0.04	0.008
27	rs6748547	233,984,766	0.05	0.05	0.36	0.69	0.73	1.00
28	rs6759896	233,992,972	0.41	0.38	0.39	0.06	0.02	0.007

Table 3-9 Fine mapping of the CD association signal at the *ATG16L1* locus. The p-values obtained in panel B in allele-based (CCA) and genotype-based (CCG) association analyses of the tagging and coding SNPs are shown. In addition, the p-value for the TDT test in a family-based sample is shown.The only non-synonymous SNP in *ATG16L1* (rs2241880) is highlighted in blue. MAF: minor allele frequency. Positions are according to NCBI build 35. The region comprises 84 kb and the average SNP density is 3.1 kb.

The association of *ATG16L1* with CD is strongly supported by these results, as 11 out of the 28 typed SNPs are significant in the TDT and the CC analysis.

3.3.1.3 *ATG16L1* haplotype analysis

A nine-marker haplotype analysis was carried out for panel B as described in 2.9.5 on page 67. The results of COCA- and TDTPHASE are listed in table 3-9, page 96.

Obviously, the sole risk haplotype (ACACAGGCG) is fully tagged by rs2241880 allele "G", while all other haplotypes carry allele A. This haplotype pattern strongly suggests that rs2241880 is indeed the major risk variant at the *ATG16L1* locus. Rare haplotypes that had a frequency <1% were excluded from the analysis.

Haplotype	f_{cases}	$f_{con-trols}$	$OR_{case-control}$	p-value COCAPHASE	$f_{transmitted}$	$f_{non-trans-mitted}$	OR_{TDT}	p-value TDTPHASE
ACACA**G**GCG	0.603	0.532	1.34	0.00002	0.535	0.285	2.87	0.0001
ACGCTAACG	0.254	0.283	0.86	0.0502	0.262	0.396	0.54	0.0164
AGATAAATG	0.047	0.069	0.67	0.0045	0.077	0.092	0.82	0.5462
GGACAAATG	0.052	0.069	0.74	0.0420	0.081	0.139	0.55	0.0608
ACGCAAGTA	0.044	0.048	0.91	0.5685	0.035	0.073	0.46	0.0728
ACACAAGTG	<0.01	<0.01	n.d.	n.d.	0.012	0.015	0.80	0.7050

Table 3-10 Results of a haplotype analysis of 9 SNPs at the ATG16L1 locus. SNPs included in the haplotype analysis are marked by asterisks in fig. 4-1, thereby allowing to discern their block assignment. All analyses were carried out using either COCAPHASE or TDTPHASE (Dudbridge et al., 2003). Non-synonymous SNP rs2241880 is highlighted in bold and the risk allele is underlined.

SNPs #20 (2289472) and #22 (rs2241879), which are adjacent to the nsSNP rs2441880, were also strongly associated with CD. This is obvious as linkage disequilibrium among these three variants is very high ($r^2 > 0.80$), even when using the conservative measure r^2:

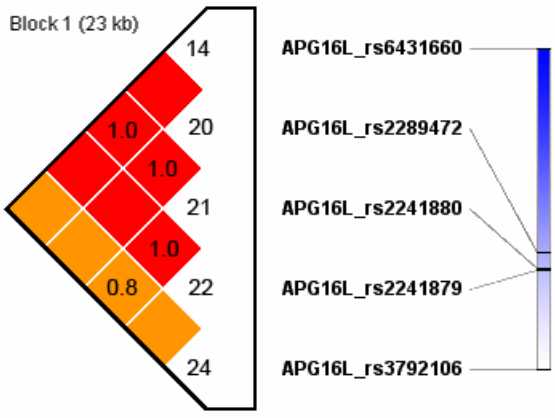

Fig. 3-2 Linkage disequilibrium between significant SNPs of the ATG16L1 gene. Using the measure r^2, pairwise LD between the five SNPs with p-values <0.001 in the case-control and TDT analysis (see table 3-9, page 96) is shown.

3.3.2 *NELL1* fine-mapping

The best and only replicated candidate gene out of the LD-based 100k scan was *NELL1* (nel-like 1). The gene is located on the plus strand of chromosome 11p15.1 and comprises 20 exons that stretch over an immense genomic region of 906 kb. The coding sequence of *NELL1* is 2430 bp long (810 aa). NELL1 contains epidermal growth factor (EGF)-like repeats and the encoded heterotrimeric protein may be involved in cell growth regulation and differentiation.

3.3.2.1 Resequencing of *NELL1*

Using the same 47 CD patient DNAs as described for *ATG16L1* (3.3.1.1 on page 91), mutation detection was carried out for *NELL1*. This was done to detect potentially existing non-synonymous SNPs. The results are listed in the following table:

#	SNP ID	position (build 35)	SNP type	sequence
1	rs1715283	20,646,687	Promoter	–
2	rs3808993	20,646,953	Promoter	–
3	NELL1_01	20,647,024	Promoter	CTGCCCTGCAGAATGAGAAGGTTTGCAAA TAGACTTCCCAAACCCCAACCA-CAGCTCGCTCCGCCTCGAGGAC-CCCTTTTCTGCACCCCCACCTCAGCGC **[C/A]** CTCTTCCTGCACCCACAAAGAGAGTACT-CAGTCATAGGGGTTCAACAGGAGAGAG-GAGACAGAAGGTACAGGCGGTGAGCAGG GACTCAGCCATCATCCC
4	rs1793003	20,655,970	intronic	–
5	rs2280362	20,761,694	intronic	–
6	rs8176785	20,761,862	non-synonymous R→Q	–
7	rs2280363	20,761,911	synonymous	–
8	NELL1_02	20,825,777	non-synonymous R→S	AAACAAATGTTTGTTCCTTTACATACAGC-TATTTTGAACTGGAGAGCAGTGGCCTGAG GGATGAGATTCGGTATCACTACATACA-CAATGGGAAGCCAAG **[G/C]** ACAGAGGCACTTCCTTACCGCATGGCAGA TGGACAATGGCACAAGGTTGCACTGT-CAGTTAGCGCCTCTCATCTCCTGCTC-CATGTCGACTGTAACAGGT

Table 3-11 Results of mutation detection of *NELL1*. Corresponding primer sequences are shown in table table 8-5, page 163.

#	SNP ID	position (build 35)	SNP type	sequence
9	NELL1_03	20,825,826	non-synonymous A→T	TGGCCTGAGGGATGAGATTCGGTATCAC-TACATACACAATGGGAAGCCAAGGACAGA GGCACTTCCTTACCGCATGGCAGATGGA-CAATGGCACAAGGTT **[G/A]** CACTGTCAGTTAGCGCCTCTCATCTCCTG CTCCATGTCGACTGTAACAGG-TATTTCTTTGTCTTTGAGTGTTGCTGAT-TCTGCCCTTGAAATCAAGCAAT
10	rs3740874	20,896,231	intronic	–
11	rs1429785	20,897,351	intronic	–
12	rs2293241	20,905,691	intronic	–
13	rs1880088	20,915,761	Intronic	–
14	rs8176786	20,915,970	non-synonymous R→W	–
15	rs1880087	20,916,005	Intronic	–
16	rs8176791	20,916,041	intronic	–
17	rs1880086	20,916,058	intronic	–
18	rs8176796	20,925,428	intronic	–
19	rs2280584	20,938,470	intronic	–
20	rs3758810	21,091,662	intronic	–
21	rs8176792	21,349,039	synonymous	–
22	rs4151056	21,349,063	synonymous	–
23	rs7119475	21,538,241	intronic	–
24	rs8176789	21,538,381	synonymous	–
25	rs8176790	21,538,573	intronic	–
26	rs4922847	21,548,804	intronic	–
27	rs8176793	21,549,083	intronic	–

Table 3-11 Results of mutation detection of *NELL1*. Corresponding primer sequences are shown in table table 8-5, page 163.

Three new polymorphisms, including two non-synonymous SNPs, were found and 24 annotated SNPs were verified. In total, 4 non-synonymous and 4 synonymous SNPs were detected in coding regions but no SNPs were found in splice sites.

3.3.2.2 Linkage disequilibrium and association statistics for *NELL1*

118 tagging SNPs were selected from the Caucasian HapMap data and genotyped in the German replication panel D. A different set of 93 tagging SNPs was genotyped by a collaborator (Genizon BioSciences) in French-Canadian CD trios (panel H). The results for this experiment are listed in table 9-4, page 185 and they are included in plot fig 3-3, page 103 as well. Genotyping 118 SNPs across the genomic region of 963 kb, gave an average SNP density of 8.2 kb. For the French-Canadian sample, an average of 10.5 kb was achieved. Although the extent of LD varies greatly across the genome, this high density will extract most of the information in this region.

#	SNP ID	position (build 35)	MAF	P_{CCG}	P_{CCA}	P_{TDT}	P_{HAP3}
1	rs7942430	20,594,635	0.69	0.61	0.75	0.53	0.75
2	rs3740872	20,605,132	0.63	0.42	0.55	0.87	0.26
3	rs7122910	20,636,147	0.70	0.32	0.60	0.69	0.42
4	rs1792969	20,646,062	0.89	0.11	0.05	0.87	0.09
5	NELL01	20,647,024	0.99	0.57	0.56	0.03	0.21
6	rs1793005	20,655,334	0.73	0.03	0.06	0.50	0.10
7	rs1793004	20,655,505	0.72	0.03	0.06	0.03	0.22
8	rs1793003	20,655,970	0.73	0.05	0.10	0.50	0.07
9	rs951199	20,657,351	0.73	0.03	0.07	0.33	0.02
10	rs870194	20,663,881	0.79	0.28	0.25	0.05	0.03
11	rs1607616	20,674,469	0.65	0.005	0.01	0.001	0.03
12	rs1792983	20,674,751	0.79	0.24	0.23	0.02	0.15
13	rs11025705	20,688,454	0.86	0.95	0.99	0.20	0.43
14	rs7952174	20,695,492	0.85	0.30	0.54	0.54	0.06
15	rs908940	20,708,141	0.97	0.44	0.15	0.25	0.90
16	rs7116986	20,711,993	0.86	0.93	0.99	0.35	0.83
17	rs1996623	20,715,992	0.85	0.65	0.74	0.27	0.46
18	rs1519727	20,721,594	0.65	0.83	0.49	0.73	0.98
19	rs1554368	20,732,334	0.82	0.82	0.45	0.13	0.17
20	rs17232778	20,735,857	0.92	0.62	0.82	0.28	0.09
21	rs1429796	20,739,579	0.98	0.007	0.006	0.37	0.10
22	rs4922623	20,741,119	0.69	0.97	0.98	0.06	0.51
23	rs327025	20,746,233	0.85	0.62	0.40	0.03	0.57
24	rs16906777	20,748,858	0.66	0.42	0.73	0.19	0.45
25	rs327028	20,748,995	0.92	0.78	0.84	0.41	0.50
26	rs1949523	20,750,221	0.73	0.10	0.08	0.75	0.43
27	***rs8176785***	***20,761,862***	***0.73***	***0.12***	***0.04***	***0.39***	***0.44***
28	rs2280363	20,761,911	0.73	0.12	0.08	0.42	0.30
29	rs7129413	20,769,224	0.87	0.70	0.76	0.48	0.36
30	rs1158547	20,771,723	0.67	0.34	0.37	0.02	0.18
31	rs11601634	20,791,637	0.78	0.35	0.38	0.55	0.18
32	rs1519735	20,791,833	0.65	0.03	0.06	0.05	0.28
33	rs7109624	20,792,580	0.95	0.54	0.11	0.80	0.61

Table 3-12 Results of *NELL1* fine mapping. The p-values obtained for the tagging and coding SNPs that were typed in panel D are shown. SNPs that are significant in either one of the case-control analyses (CCA: allele-based comparison; CCG: genotype-based comparison) or the TDT are highlighted in blue. Lead SNPs from the initial screening are highlighted by a grey shading. Non-synonymous SNPs are marked in red and significant nsSNPs are highlighted in bold italics. MAF: minor allele frequency in unrelated control individuals. p_{HAP3}: p-value obtained by a sliding three-marker window haplotype analysis by means of COCAPHASE.

#	SNP ID	position (build 35)	MAF	P_{CCG}	P_{CCA}	P_{TDT}	P_{HAP3}
34	rs7130897	20,796,450	0.84	0.75	0.88	0.42	0.47
35	rs435001	20,800,510	0.88	0.25	0.50	0.75	0.10
36	rs17298565	20,806,806	0.96	0.14	0.16	0.80	0.44
37	rs1914984	20,807,913	0.92	0.91	0.56	1.00	0.90
38	rs2680989	20,809,773	0.88	0.62	0.49	0.22	0.54
39	rs919473	20,822,486	0.82	0.71	0.44	0.26	0.67
40	rs7114248	20,824,095	0.78	0.79	0.89	0.22	0.99
41	rs1429799	20,824,886	0.79	0.78	0.80	0.30	0.99
42	rs1346690	20,827,819	0.81	0.80	0.72	0.19	0.87
43	rs1367002	20,828,374	0.89	0.45	0.32	0.93	0.76
44	rs11025788	20,829,897	0.79	0.62	0.70	0.24	0.02
45	rs7121400	20,831,014	0.90	0.93	0.53	0.13	0.02
46	rs12293297	20,832,922	0.98	0.0009	0.006	0.87	0.04
47	rs1429794	20,838,013	0.82	0.55	0.69	0.29	0.51
48	rs6483735	20,844,275	0.79	0.10	0.19	0.12	0.26
49	rs10766733	20,844,335	0.76	0.33	0.25	0.07	0.10
50	rs919476	20,845,201	0.93	0.19	0.19	0.04	0.30
51	rs10766735	20,846,648	0.83	0.07	0.16	0.50	0.31
52	rs4923128	20,848,373	0.71	0.18	0.34	0.05	0.81
53	rs7109004	20,849,339	0.71	0.21	0.43	0.07	0.75
54	rs1549717	20,867,190	0.86	0.80	0.96	0.10	0.49
55	rs2082080	20,891,024	0.75	0.85	0.55	0.14	0.42
56	rs1880088	20,915,761	0.75	0.14	0.03	0.49	0.54
57	rs8176786	20,915,970	0.95	0.63	0.27	0.34	0.44
58	rs1880084	20,927,365	0.73	0.71	0.63	0.07	0.04
59	rs10833417	20,945,048	0.74	0.02	0.04	0.49	0.008
60	rs10500884	20,955,364	0.88	0.94	0.99	0.74	0.005
61	rs1400373	20,956,134	0.88	0.98	0.94	0.68	0.06
62	rs10500885	20,971,636	0.90	0.05	0.02	0.87	0.05
63	rs2896623	20,973,973	0.84	0.33	0.19	0.36	0.50
64	rs10500886	20,976,742	0.76	0.11	0.29	0.44	0.63
65	rs11025878	20,998,335	0.76	0.31	0.61	0.80	0.21
66	rs4922728	21,009,790	0.81	0.74	0.63	0.55	0.79
67	rs7937542	21,017,772	0.56	0.78	0.25	0.71	0.81
68	rs10833444	21,029,035	0.76	0.49	0.79	0.80	0.70
69	rs9666195	21,038,467	0.65	0.56	0.78	0.56	0.71
70	rs1543153	21,047,961	0.77	0.82	0.96	0.80	0.68
71	rs7932820	21,056,960	0.58	0.94	0.86	0.96	0.10
72	rs12417046	21,066,312	0.81	0.17	0.17	0.84	0.29
73	rs2403652	21,076,903	0.53	0.80	0.29	0.83	0.54
74	rs7933049	21,087,224	0.79	0.58	0.25	0.95	0.99
75	rs4922753	21,096,666	0.71	0.91	0.98	0.51	0.81
76	rs10766767	21,108,451	0.70	0.91	0.99	0.56	0.95
77	rs6483748	21,137,843	0.70	0.71	0.37	0.44	0.48
78	rs4475918	21,147,706	0.60	0.35	0.39	0.60	0.31
79	rs1453983	21,166,668	0.75	0.95	0.81	0.53	0.22
80	rs1454003	21,175,735	0.83	0.21	0.46	0.32	0.66
81	rs1823843	21,185,111	0.97	0.85	0.36	0.78	0.90

Table 3-12 Results of *NELL1* fine mapping. The p-values obtained for the tagging and coding SNPs that were typed in panel D are shown. SNPs that are significant in either one of the case-control analyses (CCA: allele-based comparison; CCG: genotype-based comparison) or the TDT are highlighted in blue. Lead SNPs from the initial screening are highlighted by a grey shading. Non-synonymous SNPs are marked in red and significant nsSNPs are highlighted in bold italics. MAF: minor allele frequency in unrelated control individuals. p_{HAP3}: p-value obtained by a sliding three-marker window haplotype analysis by means of COCAPHASE.

#	SNP ID	position (build 35)	MAF	P_{CCG}	P_{CCA}	P_{TDT}	P_{HAP3}
82	rs1945327	21,190,821	0.90	0.61	0.83	0.80	0.52
83	rs1453988	21,201,279	0.69	0.47	0.75	0.32	0.62
84	rs1670638	21,207,776	0.68	0.69	0.88	0.34	0.98
85	rs1670640	21,218,026	0.54	0.74	0.15	0.64	0.87
86	rs1454008	21,230,114	0.59	0.53	0.74	0.57	0.93
87	rs1791822	21,240,131	0.54	0.65	0.12	0.87	0.86
88	rs10500901	21,240,719	0.97	0.84	0.98	0.66	0.45
89	rs1453990	21,249,027	0.53	0.57	0.48	0.91	0.21
90	rs716577	21,262,140	0.63	0.46	0.62	0.42	0.22
91	rs6483756	21,267,196	0.62	0.24	0.46	0.41	0.07
92	rs10833498	21,277,881	0.68	0.32	0.60	0.02	0.28
93	rs4335544	21,286,758	0.59	0.83	0.86	0.52	0.25
94	rs1349818	21,292,747	0.51	0.28	0.37	0.74	0.76
95	rs4399327	21,302,941	0.52	0.52	0.09	1.00	0.94
96	rs2187522	21,313,688	0.51	0.91	0.73	0.25	0.61
97	rs11026036	21,323,317	0.64	0.44	0.73	0.87	0.59
98	rs7126959	21,333,727	0.71	0.32	0.59	0.95	0.23
99	rs1945404	21,343,840	0.63	0.81	0.13	1.00	0.24
100	rs4151056	21,349,063	0.95	0.07	0.18	0.41	0.56
101	rs10833520	21,352,936	0.80	0.79	0.47	0.18	0.93
102	rs1945443	21,365,895	0.66	0.58	0.25	0.61	0.80
103	rs4539321	21,375,322	0.54	0.45	0.03	0.66	0.23
104	rs11026072	21,385,901	0.67	0.47	0.58	0.07	0.02
105	rs7943922	21,394,268	0.50	0.09	0.22	0.20	0.68
106	rs11026079	21,406,963	0.75	0.59	0.50	1.00	0.68
107	rs7110569	21,418,064	0.83	0.88	0.97	0.56	0.91
108	rs1945408	21,428,394	0.86	0.88	0.62	0.36	1.00
109	rs7945802	21,438,700	0.57	0.66	0.48	0.83	0.52
110	rs10766821	21,458,271	0.62	0.72	0.91	0.54	0.09
111	rs7116826	21,472,159	0.85	0.44	0.60	0.39	0.86
112	rs10219188	21,479,105	0.53	0.48	0.74	0.87	0.91
113	rs6483774	21,491,019	0.51	0.55	0.81	0.96	0.76
114	rs10766829	21,499,146	0.59	0.45	0.41	0.91	0.49
115	rs7926887	21,523,755	0.61	0.53	0.54	0.96	0.30
116	rs4319515	21,534,943	0.51	0.14	0.34	0.96	0.28
117	rs8176789	21,538,381	0.83	0.86	0.50	0.94	–
118	rs4922850	21,557,214	0.93	0.30	0.10	0.58	–

Table 3-12 Results of *NELL1* fine mapping. The p-values obtained for the tagging and coding SNPs that were typed in panel D are shown. SNPs that are significant in either one of the case-control analyses (CCA: allele-based comparison; CCG: genotype-based comparison) or the TDT are highlighted in blue. Lead SNPs from the initial screening are highlighted by a grey shading. Non-synonymous SNPs are marked in red and significant nsSNPs are highlighted in bold italics. MAF: minor allele frequency in unrelated control individuals. p_{HAP3}: p-value obtained by a sliding three-marker window haplotype analysis by means of COCAPHASE.

Of the 118 typed SNPs, the three SNPs rs1793004, rs1607616, and rs1519735 were significant in the TDT and the CC analysis, while 17 were significant in either analysis (7x TDT/10x CC). Out of the 20 significant SNPs and with a genotypic p-value of 0.005 and a TDT p-value of 0.001, rs1607616 showed the strongest association. A strong peak showed up in the 5' region of *NELL1*, while (fig 3-3, page 103) the neighbouring gene *SLC6A5* can be excluded. Using Pearson's χ^2 (Ott *et al.*, 1985), no significant or borderline association was seen for the two

non-synonymous SNP rs8176785 and rs8176786, respectively. More detailed results for all four nsSNPs are given in table 3-13, page 104.

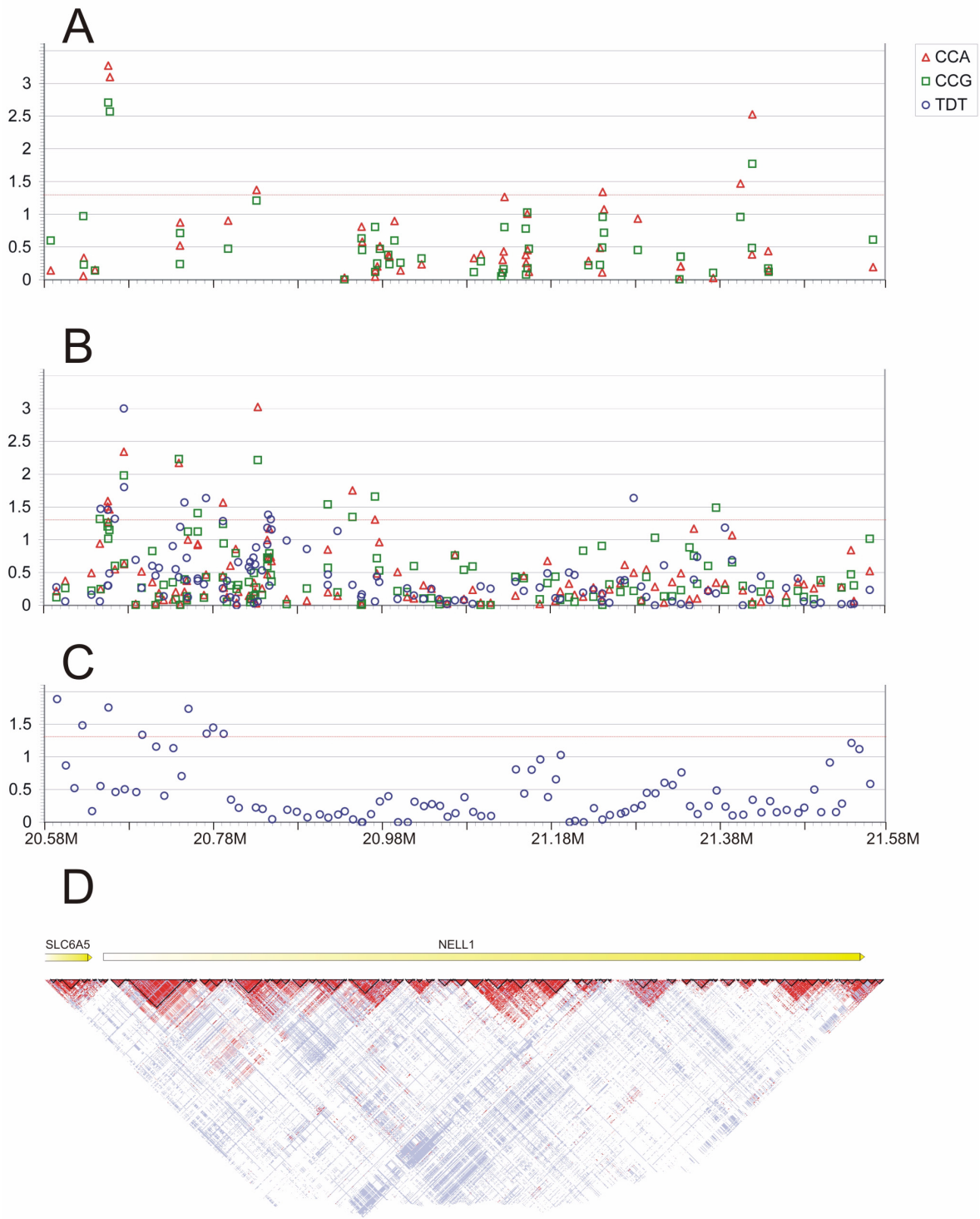

Fig. 3-3 Fine mapping results and LD structure for *NELL1*. The vertical axis shows the negative $\log_{10}$ of corresponding p-values and plotted along the horizontal axis are chromosomal positions according to NCBI build 35. The region shown here comprises 0.96 Mb. (A) SNP density and initial analysis results of the genome-wide scan in panel C; (B) Results of replication and fine mapping in panel D and (C) panel H. The red dotted line depicts the significance threshold of p = 0.05. (D) Linkage disequilibrium across the *NELL1* locus based on HapMap data (1298 SNPs) using the measure D'. LD blocks are shown in red color. The legend next to (A) gives an explanation of the symbols used for the different analyses (CCA = allele-based case-control comparison, CCG = genotype-based case-control comparison, and TDT = transmission disequilibriumg test).

Analyses of the French-Canadian genotype data revealed that the peak, which was observed in German CD patients, co-localizes with the results of the TDT test (fig 3-3, page 103, C). This finding in an independent population strengthens the detected association of *NELL1* and CD.

Since the strongest associated SNP rs1607616, which is located between exon 3 and 4 of *NELL1*, fails to provide a direct explanation for a harmful change on the NELL1 protein level, a closer look was taken at the four non-synonymous SNPs. Results are listed in the following table. The characteristics described for rs1607616 concerns all associated SNPs in non-coding regions of *NELL1* as well.

Two additional nsSNPs are listed in NCBI's dbSNP for *NELL1*: rs11820003 und rs8176788. Both were not chosen for genotyping because the first is annotated as monomorphic in Caucasians and the latter is not validated.

		controls			cases			controls		cases	
dbSNP ID	**Exon**	f_{11}	f_{12}	f_{22}	f_{11}	f_{12}	f_{22}	f_1	f_2	f_1	f_2
rs8176785	4	0.515	0.427	0.058	0.570	0.365	0.065	0.729	0.271	0.752	0.248
NELL1_02	5	1.000	0.000	0.000	0.998	0.002	0.000	1.000	0.000	0.999	0.001
NELL1_03	5	0.991	0.009	0.000	0.992	0.009	0.000	0.996	0.004	0.996	0.004
rs8176786	11	0.907	0.092	0.001	0.905	0.088	0.006	0.953	0.047	0.949	0.051

Table 3-13 *NELL1* non-synonymous SNPs. Frequencies are listed for genotypes and alleles in both controls and cases. Significant differences are highlighted in blue.

Fisher's exact test (as described in 2.9.7 on page 70) was calculated for the low-frequency nsSNPs NELL1_02 and rs8176786, since both contingency tables contained values smaller five. The results were not significant: p = 0.225 for NELL1_02 and p = 0.128 for rs8176786. Interestingly, there are 6% more heterozygotes for rs8176785 in the control population than in the CD sample but fewer homozygotes for both alleles. This finding could not be replicated in a British CD sample (panel F). Besides rs8176785, five percent more homozygous cases for the common allele exist for rs1793004 and rs951199. No significant differences in genotype or allele frequencies are seen for rs8176786. Therefore, other causative SNPs might exist that contribute to the risk associated with this loci. Using the same R script and logistic regression as done for rs2241880, no statistical interaction was observed between rs8176785, rs8176786, and the three *CARD15* mutations "SNP 8", "SNP 12", and "SNP 13".

Finally, SNPlex™ results for the two nsSNPs rs8176785 and rs8176786 were validated with TaqMan®, an independent genotyping method. The genotype concordance was 99.78% for rs8176785 and 99.94% for rs8176785 (n = 3231 and 3247 genotypes compared, respectively).

3.3.3 Fine mapping of the new susceptibility region on 5p13.1

The region on chromosome 5p was identified in the genome-wide scan using Affymetrix arrays. Three SNPs in this region replicated in an independent German sample of single patients and nuclear families. In the inital screening, besides a high and wide peak with a maximum at rs1992662, a small narrow peak at rs2122564 was detected. The same peaks showed up in the replication study, when 28 SNPs were typed across a region of 777.5 kb at an average SNP density of 28.8 kb. For the replication in panel H, a sample of French-Canadian CD trios, 72 tagging SNPs were typed across 1058.1 kb by the collaborator (average density: 14.9 kb). 13 SNPs had a significant p-value (<0.05) in the TDT and the same peaks were seen as in the outbred German population. Detailed results of this analysis can be found at the end of this thesis: 9.6 on page 187.

#	SNP ID	position (build 35)	MAF	pCCG	pCCA	pTDT	pHAP3
1	***rs189814***	***40,292,083***	***0.49***	***0.25***	***0.21***	***0.04***	***0.0002***
2	***rs1445011***	***40,315,959***	***0.28***	***0.05***	***0.08***	***0.01***	***$5.62{\cdot}10^{-05}$***
3	rs1445002	40,355,634	0.14	$3.94{\cdot}10^{-05}$	0.0002	0.04	0.0001
4	***rs443583***	***40,368,840***	***0.21***	***0.06***	***0.16***	***0.02***	***0.13***
5	rs12518245	40,381,478	0.06	0.50	0.70	0.91	0.0001
6	***rs7725523***	***40,407,980***	***0.25***	***0.14***	***0.32***	***0.01***	***$4.14{\cdot}10^{-06}$***
7	rs1992662	40,429,609	0.32	$7.59{\cdot}10^{-05}$	0.0002	0.001	0.0007
8	rs1992660	40,450,824	0.38	$4.53{\cdot}10^{-05}$	0.0002	0.0005	0.0001
9	rs4957300	40,499,496	0.33	0.003	0.006	0.0002	$4.89{\cdot}10^{-05}$
10	***rs4422570***	***40,518,098***	***0.10***	***0.83***	***0.75***	***0.03***	***$2.77{\cdot}10^{-05}$***
11	rs7725052	40,523,027	0.44	$3.24{\cdot}10^{-06}$	$1.95{\cdot}10^{-05}$	0.002	$9.20{\cdot}10^{-06}$
12	***rs1553575***	***40,538,689***	***0.35***	***$1.68{\cdot}10^{-06}$***	***$6.37{\cdot}10^{-06}$***	***0.03***	***$1.05{\cdot}10^{-06}$***
13	***rs7718309***	***40,564,656***	***0.19***	***0.02***	***0.01***	***0.08***	***$7.60{\cdot}10^{-09}$***
14	rs7713972	40,588,231	0.16	0.26	0.15	0.14	$1.80{\cdot}10^{-08}$
15	rs6451525	40,590,026	0.30	$1.64{\cdot}10^{-05}$	$4.49{\cdot}10^{-05}$	0.04	$1.02{\cdot}10^{-07}$
16	***rs6864749***	***40,604,350***	***0.19***	***0.03***	***0.01***	***0.12***	***0.04***
17	rs6451529	40,611,106	0.08	0.66	0.68	0.27	0.02
18	***rs4409138***	***40,639,362***	***0.12***	***0.05***	***0.14***	***0.35***	***0.12***
19	rs924967	40,650,879	0.12	0.06	0.16	0.40	0.23
20	rs6877753	40,677,032	0.11	0.50	0.56	0.52	0.43
21	rs4546432	40,718,307	0.38	0.43	0.20	0.37	0.22
22	rs6451535	40,723,788	0.32	0.98	0.27	0.22	0.13
23	rs10053664	40,821,568	0.43	0.43	0.48	0.46	0.007
24	rs10512747	40,877,498	0.12	0.08	0.23	0.06	0.11
25	***rs16870407***	***40,888,804***	***0.11***	***0.05***	***0.13***	***0.92***	***0.23***
26	rs2122564	41,023,131	0.44	0.69	0.53	0.13	0.03
27	rs896118	41,054,986	0.49	0.99	0.84	0.15	–
28	***rs325832***	***41,069,536***	***0.16***	***0.007***	***0.02***	***0.36***	–

Table 3-14 Association results for the 5p13.1 locus. The p-values obtained in the German patient panel D in allele-based (CCA) and genotype-based (CCG) association analyses of the tagging and coding SNPs are shown. Lead SNPs from the initial screening (table 3-3, page 84) are highlighted by grey shading, coding SNPs in *CARD6* in red. Polymorphisms that are significant in either the TDT or the case-control analysis, are highlighted in bold italics. Those that are significant in both are highlighted in blue. p_{TDT} is the p-value for the TDT and p_{HAP3} is the p-value for a three-marker haplotype analysis performed with COCAPHASE.

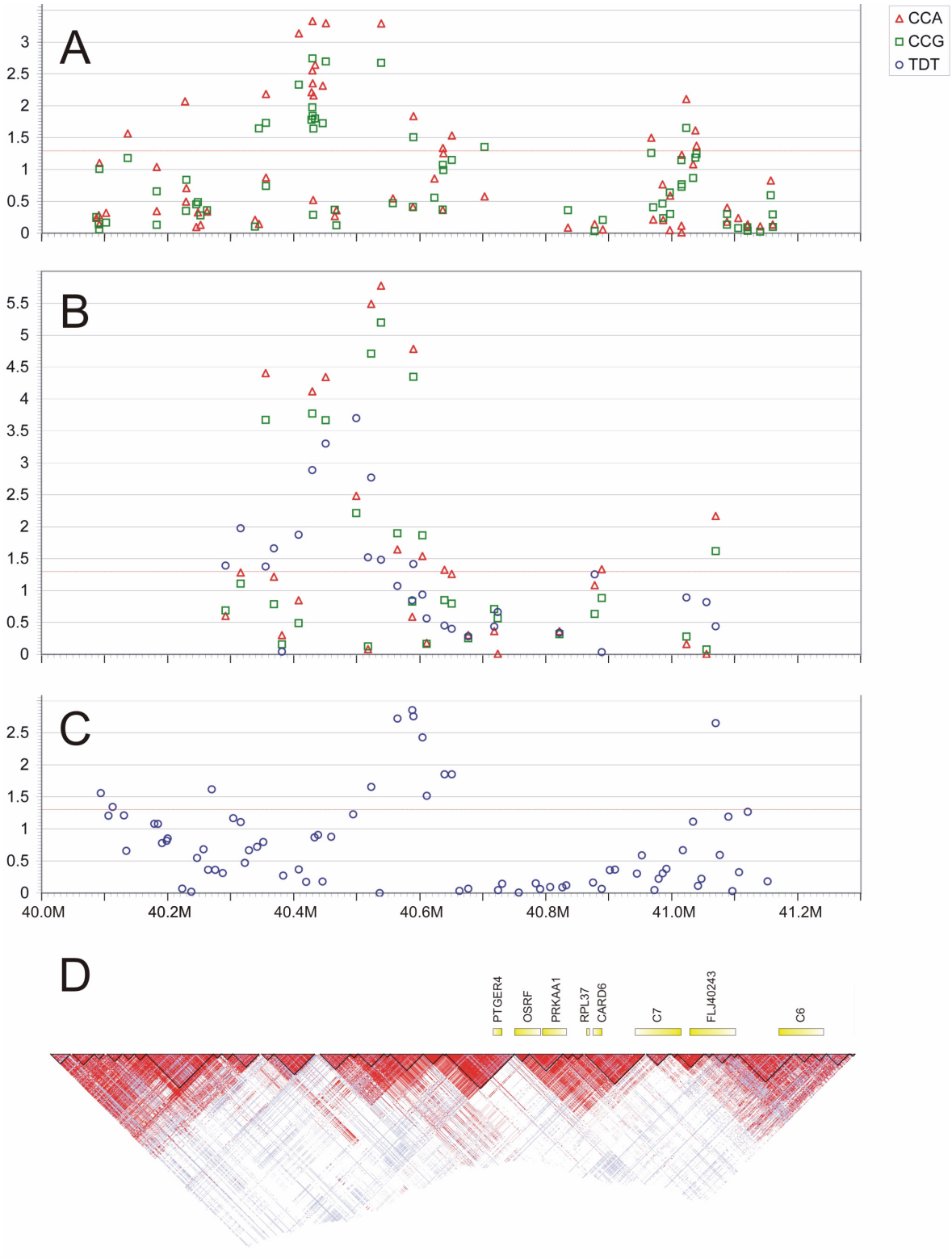

Fig. 3-4 Fine mapping and LD analysis of 5p13.1. The y-axis shows the negative $\log_{10}$ of corresponding p-values and plotted along the x-axis are chromosomal positions according to NCBI build 35.The entire region spans 1.30 Mb. (A) SNP density and initial analysis results of the genome-wide scan in panel C; (B) Results of replication and fine mapping in panel D and (C) panel H. The red dotted line depicts the significance threshold of p = 0.05. (D) Linkage disequilibrium across 5p13.1 based on HapMap data (1476 SNPs) using the measure D'. LD blocks are shown in red color. Known genes are drawn above the LD-plot. The legend next to (A) gives an explanation of the symbols used for the different analyses (CCA = allele-based case-control comparison, CCG = genotype-based case-control comparison, and TDT = transmission disequilibriumg test).

The consistent peak on 5p13.1 localizes to the gene-free region upstream of prostaglandin E receptor 4 (*PTGER4*, see figure above). A closer look at the data of the UCSC genome browser in this region revealed no convincing gene prediction (chr5: 40,315,959–40,604,350, NCBI build 35), although a single mRNA fragment (BC041051, 752 bp, poly-A tail) and four spliced ESTs exist (BG182136, BG188413, BG201263, BG184600). A `tblastx` search for the poly-adenylated mRNA BC041051 at the NCBI homepage showed that it is likely to be a repetitive sequence. More than 100 hits were found on different chromosomes with E values below 2×10^{-73}.

It should be noted here that *PTGER4* contains a bi-directional promoter and a first-exon prediction, thus another upstream-located gene could be transcribed theoretically. No significant association was seen in the LD block comprising the promising candidate genes *PTGER4* and *CARD6*.

The SNPlex™ results for the two lead SNPs rs1992660 and rs1992662 were validated with TaqMan®. The genotype concordance was 99.51% for rs1992660 and 99.36% for rs1992662 (n = 2,638 and 2,650 genotypes compared, respectively).

3.4 Testing of associations with ulcerative colitis

As mentioned above, CD and UC are both inflammatory bowel diseases, and though they represent different disorders, they share some characteristics. Furthermore, many diagnoses change during the course of the disease from one type of IBD to the other. Hence, it is reasonable to test the previously found polymorphisms for association with UC as well. This was carried out in large UC patient collections of German and British descent.

3.4.1 Evaluation of rs2241880 in ulcerative colitis

SNP rs2241880, found to be significantly associated with CD, was also evaluated in a sample of German patients with UC (panel E, table 2-1, page 26). Allele frequencies of allele "G" in cases (0.46) and controls (0.47) were virtually identical, and evidence for association was thus neither obtained from the case-control comparison ($p > 0.4$ in both the allele- and genotype-based test), nor from the the TDT ($p > 0.9$).

3.4.2 Testing of 5p13.1 association with ulcerative colitis

Similarly to rs2241880, the two lead SNPs of 5p13.1 (rs1992662, rs1992660) were tested in a UC sample. The sample (panel E) comprised 439 trios, 1227 single cases, and 1032 healthy normals. No evidence of association was detected between rs1992662 and UC:

$p_{CCA} = 0.2391$, $p_{CCG} = 0.4719$, and $p_{TDT} = 0.5940$, $MAF_{controls} = 32\%$ vs. $MAF_{cases} = 30\%$

SNP rs rs1992660 was not significant either:

$p_{CCA} = 0.1784$, $p_{CCG} = 0.3999$, and $p_{TDT} = 0.4434$, $MAF_{controls} = 37\%$ vs. $MAF_{cases} = 35\%$

With p-values > 0.4 in the allelic and genotypic χ^2 test, both lead SNPs of the 5p13.1 region were not associated with UC in UK patients (panel G). Therefore, polymorphisms in the 5p13.1 regions seem to play a role only in the etiology of CD but not UC.

3.4.3 *NELL1* and its association with ulcerative colitis

Replicated *NELL1* lead SNPs of the 100k scan, rs951199 and rs1793004, were tested for association with UC in the German panel E and the UK panel G. As both SNPs were signficant in the case-control analysis of the German sample, the two non-synonymous SNPs rs8176785 and rs8176786 were tested in these panels as well. Detailed results are given in the following table.

dbSNP ID	Panel	p_{CCA}	p_{CCG}	p_{TDT}	11_{co}	11_{ca}	12_{co}	12_{ca}	22_{co}	22_{ca}
rs1793004	GER (E)	0.002	0.003	0.54	0.514	0.598	0.414	0.343	0.071	0.059
	UK (G)	0.24	0.38	–	0.542	0.496	0.396	0.438	0.062	0.066
rs951199	GER (E)	0.001	0.0008	0.19	0.497	0.601	0.431	0.334	0.072	0.065
	UK (G)	0.58	0.81	–	0.541	0.528	0.400	0.404	0.059	0.068
rs8176785	GER (E)	0.06	0.006	0.68	0.517	0.580	0.425	0.352	0.058	0.068
	UK (G)	0.77	0.95	-	0.594	0.604	0.337	0.329	0.069	0.067
rs8176786	GER (E)	0.35	0.64	0.77	0.906	0.893	0.092	0.105	0.002	0.003
	UK (G)	0.43	0.049	-	0.901	0.876	0.92	0.124	0.008	0.000

Table 3-15 Evidence of association between *NELL1* and UC. Significant differences >5% are highlighted in blue. Single-point p-values are listed for an allelic (p_{CCA}) and genotypic (p_{CCG}) test as well as for the TDT (p_{TDT}) test.

Single-point analyses in the German UC panel show a significant association of the *NELL1* gene with the disease, interestingly, associations are even stronger with UC than with CD (see 3.3.2.2 on page 100). Substantial differences in genotype frequencies are seen for all SNPs in the German and the British population, except for the nsSNP rs8176786. Although the TDT results are insignificant for all four SNPs, which could be a result of transmission ratio distortion, the massive lack or excess of heterozygotes in the case sample provides further evidence for an association of the *NELL1* gene with UC. For example, there are 7% fewer heterozygous cases for rs8176785 in the German panel compared to healthy normals. In the British UC sample, there are 4% more heterozygous cases than in the controls for rs1793004.

Associations with UC and CD lead to the conclusion that *NELL1* is a true IBD susceptibility gene, though further experiments and analyses are necessary to determine the more complicated genetic background. It is possible that other functional variants exist that contribute to the risk at this locus and that are in linkage disequilibrium with the above three associated SNPs. This could also explain the inconsistent results that were obtained for the tested mutations.

3.5 Further genetic and *in silico* analyses

3.5.1 Epistasis between *ATG16L1* and *CARD15*

The *ATG16L1* gene encodes a protein in the autophagosome pathway that processes intracellular bacteria. Since both the ATG16L1 and the NOD2 proteins are involved in the innate defence against bacterial pathogens, the disease-associated variants in the two genes were investigated for a possible statistical interaction with respect to CD risk. To this end, individuals in the German fine mapping and replication sample (panel B) were classified as either homozygous wild-type (dd), heterozygous carrier (dD) or homozygous carrier (DD, which included compound heterozygotes) for the three main causative *CARD15* SNPs rs2066844 (R702W, "SNP 8"), rs2066845 (G908R, "SNP 12") and rs2066847 (1007InsC, "SNP 13"). Appropriateness of this classification is supported by the published haplotype structure of the *CARD15* gene (Croucher *et al.*, 2003). The frequency and odds ratio for individual *CARD15* risk genotypes, stratified by rs2241880 genotype, are shown in table 3-16.

affection status	*ATG16L1*	*CARD15*		
		dd	**dD**	**DD**
control	GG	219	62	2
	AG	435	87	2
	AA	185	35	5
CD	GG	175	92	42
	AG	232	136	57
	AA	73	50	21
odds ratios	GG	2.03 (1.43 – 2.88)	1.04 (0.59 – 1.84)	5.00 (0.76 – 41.05)
	AG	1.35 (0.98 – 1.87)	1.09 (0.64 – 1.88)	6.79 (1.04 – 55.16)
	AA	1.00	1.00	1.00

Table 3-16 Analysis of the statistical interaction between SNP rs2241880 and *CARD15* genotype. See description above.

The odds ratio difference was significant for rs2241880 genotype "GG" (2.03 versus 1.04; Breslow-Day χ^2 = 4.267, 1 d.f., p = 0.039), and not for "AG". Thus, the *ATG16L1* variant rs2241880 is a risk factor even in the absence of *CARD15* mutations. On the background of *CARD15* high-risk genotype "DD", the risk conferred by carriership of the rs2241880 allele "G" appeared to be higher than in the presence of "dd" or "dD", but the confidence intervals of the respective odds ratios were still wide owing to the small number of "DD" controls (table 3-16). Nevertheless, when rs2241880 genotypes "GG" and "AG" were combined, the joint OR of 5.89 (95% CI: 1.23 – 29.21) was found to be statistically significantly larger than unity (Fisher's exact two-sided p = 0.016), thereby confirming that rs2241880 allele "G" is a risk factor on a high-risk *CARD15* genotype background as well.

3.5.2 Location of T300A in ATG16L1 (protein model)

ATG16L1 homologues are present in a wide range of eukaryotes in the same domain architecture, except for yeast ATG16 (fig. 3-5).

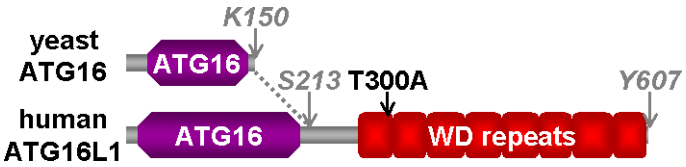

Fig. 3-5 Domain architecture of human ATG16L1 and yeast ATG16. The position of the variant amino acid T300A in the WD repeat domain is marked. The annotated ATG16 Pfam domain consists of coiled-coils. The C-terminal residue K150 of yeast ATG16 corresponds to S213 of human ATG16L1 according to a pairwise sequence alignment.

The threonine residue at position 300, which is substituted to alanine by rs2241880, is conserved across many species including mouse and rat (fig. 3-6), suggesting an important role of this amino acid. Human ATG16L1 is organized into an N-terminal ATG16 domain consisting of coiled-coils and eight C-terminal WD repeats.

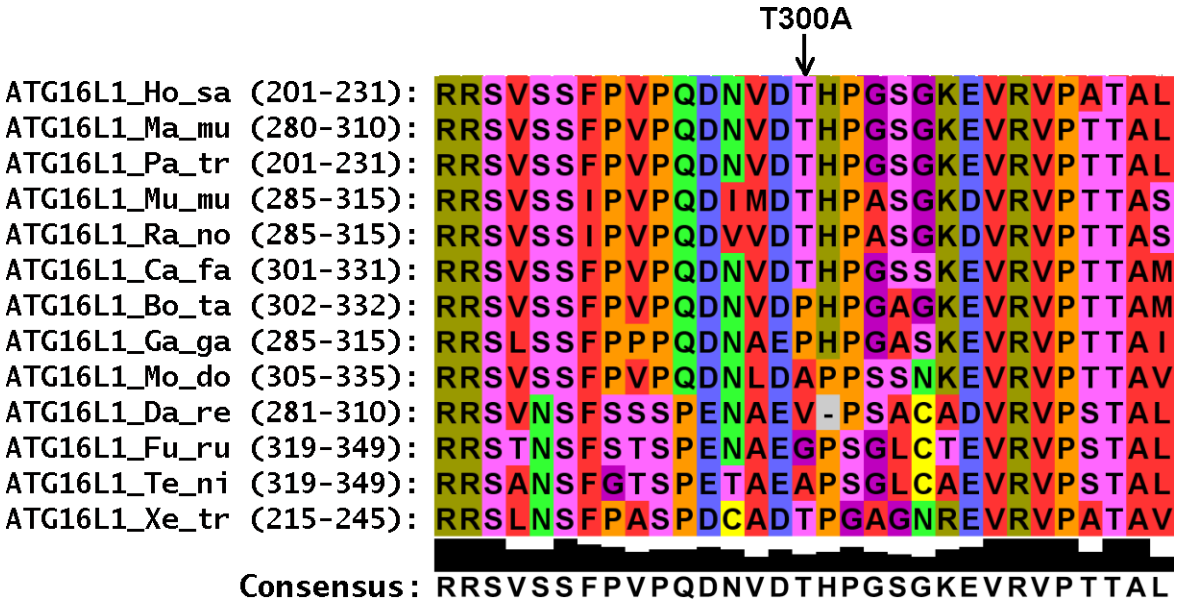

Fig. 3-6 Multiple sequence alignment of conserved region surrounding variant T300A in ATG16L1 homologs. Physicochemically similar amino acids are highlighted with identical colors. *Homo sapiens* (human), *Macaca mulatta* (rhesus monkey), *Pan troglodytes* (chimpanzee), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog), *Bos taurus* (domesticated cow), *Gallus gallus* (domesticated chicken), *Monodelphis domestica* (opossum), *Danio rerio* (zebrafish), *Fugu rubripes* (Japanese pufferfish), *Tetraodon nigroviridis* (Spotted green pufferfish), *Xenopus tropicalis* (pipid frog). Further details of used sequences are given in table 2-19, page 79.

The 3D structure of ATG16L1 was modelled (see 2.11 on page 77) using the eight-bladed β-propeller crystal structure of the evolutionarily related WD-repeat domain in yeast CDC4 (Orlicky *et al.*, 2003). The location of the T300A variant in human ATG16L1 corresponds to T397 of CDC4, where it lies at the N-terminus of the WD-repeat domain in the β3 strand of the first propeller blade (fig. 3-7 and fig 9-5, page 183). Therefore, the Thr to Ala amino acid change might have a detrimental effect on the structural stability of the affected blade and on potential binding sites nearby.

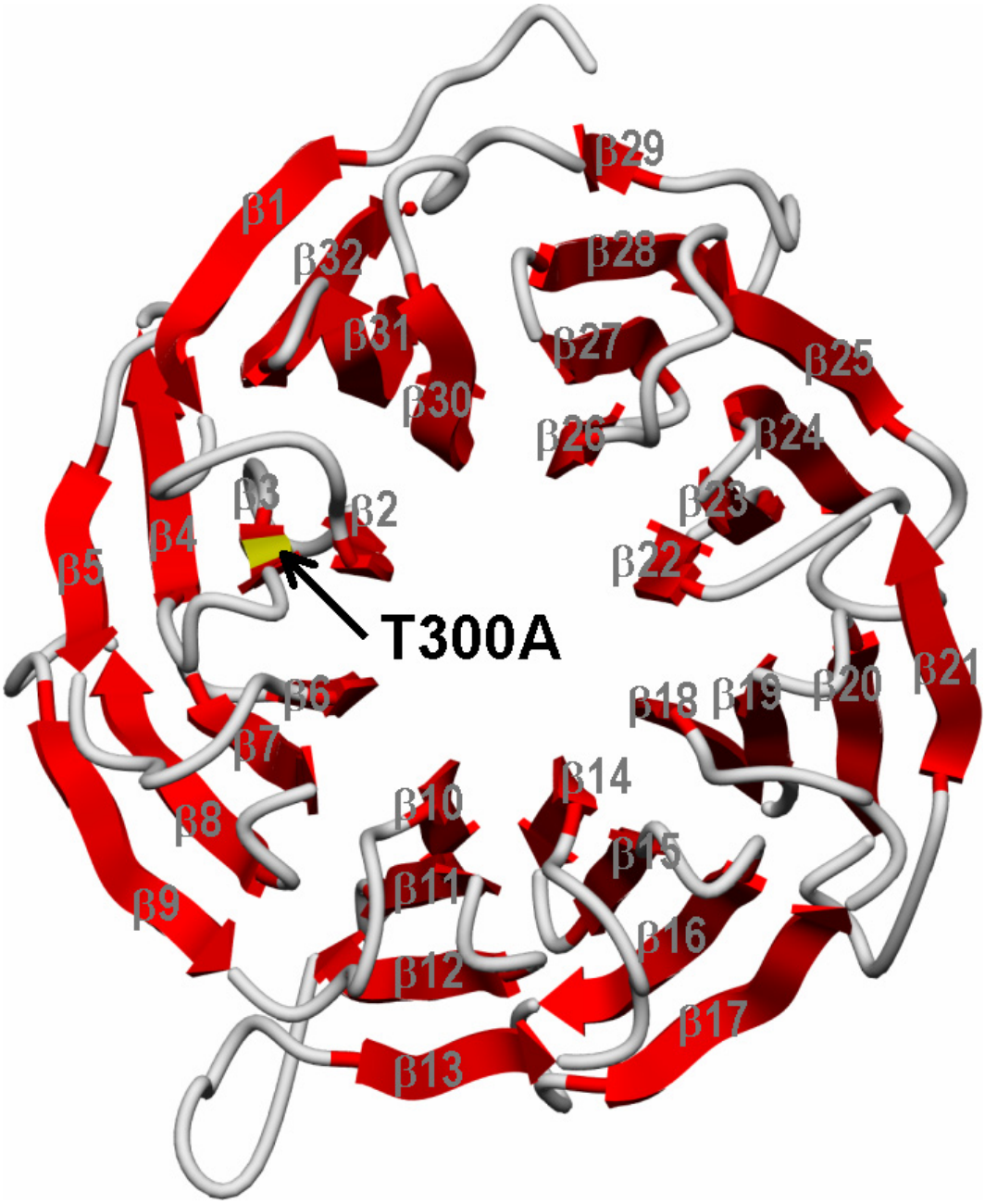

Fig. 3-7 3D structure model of the WD-repeat domain of human ATG16L1. The 32 β-strands forming an 8-bladed β-propeller are numbered as in fig 9-5, page 183. The location of the variant amino acid T300A in strand β3 is marked in yellow.

3.5.3 Determination of the rs2241880 ancestral allele

Since the frequency of the more common "G" allele is significantly higher in CD patients than in healthy normals (53% in controls, 60% in cases), we determined whether "G" is also the actual mutant allele. Judging by the frequency itself, one would expect the rarer "A" allele to be the mutant and "G" to be the ancestral allele. The latter is the sequence that is found in the last common ancestor between humans and chimpanzees (Hacia *et al.*, 1999).

A multi-species comparison on the nucleotide level of exon nine revealed that the conserved "A" variant is the ancestral allele (fig 3-8, page 113). This supports our finding that "G" is the mutant allele that predisposes to CD.

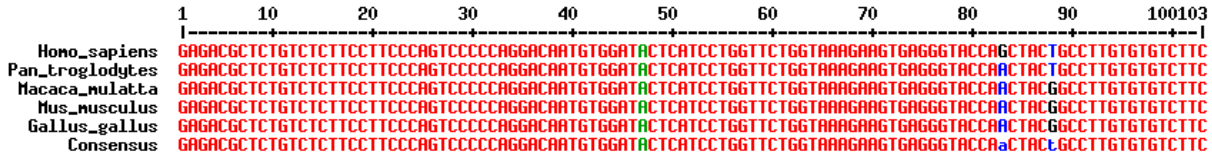

Fig. 3-8 Exon 9 multi-species alignment - Which is the ancestral allele? Exon nine sequences from six different species (human, chimpanzee, rhesus macaque, mouse, chicken) were retrieved from UCSC Golden Path and aligned with MultAlin. The conserved "A" allele of rs2241880 is shown in green color. Other mutations that occured during the evolutionary history are highlighted in blue.

3.5.4 Subphenotype analyses for *ATG16L1* mutation

Previous investigations showed that IBD associated mutations are more prevalent in certain disease subgroups. Hampe *et al.* (2002), for example, found that CD patients with inflammation in the ileum are more often carriers of the *CARD15* risk allele of "SNP 13". Therefore, several subgroups of a retrospective CD sample (panel B) were examined for significant differences in carriership of the risk genotype "GG". Results are summarized in the following table:

Clinical phenotype	n	f_{GG}	f_{GA}	f_{AA}	f_G	f_A
Controls	1032	0.27	0.51	0.22	0.53	0.47
All cases	929	0.36	0.48	0.17	0.60	0.40
Only males	245	0.40	0.40	0.20	0.60	0.40
Only females	684	0.34	0.50	0.15	0.59	0.41
Stenoses	507	0.37	0.48	0.15	0.61	0.39
Fistulae	488	0.36	0.49	0.15	0.61	0.39
Family history	88	0.35	0.43	0.22	0.57	0.43
Ileal	718	0.36	0.47	0.17	0.60	0.40
Colonic	594	0.35	0.49	0.16	0.59	0.41
Extra intestinal manifestations	665	0.35	0.50	0.15	0.60	0.40
Median age of disease onset	929	37 years	37 years	36 years	–	–

Table 3-17 Results of subphenotype analysis for retrospective panel B. The case panel was divided into several subgroups representing distinct clinical presentations. Frequencies (f) are given for each genotypic or allelic subgroup. More than one family member had to be affected by IBD in case of "Family history".

No significant differences were observed concerning allele frequencies among the different clinical subgroups, thus classifying rs2241880 allele "G" as a rather general risk factor for CD. Furthermore, the median age of disease onset is very similar for risk- and non-risk carriers. Although sample numbers are low (n = 245), significantly more men are carriers of the risk genotype "GG" compared to female patients.

3.5.5 Validation of genotyping methods

Several quality control experiments were necessary to generate consistent high-throughput genotyping data. This included the comparison of genomic DNA with WGA-DNA and the comparison between the different used genotyping methods.

3.5.5.1 Performance of genomic DNA and amplified DNA in genotyping

To compare amplified DNA with genomic DNA, 360 samples (plate XKN34) of each type were genotyped with control pool A. Thirty-seven ng of fragmented gDNA, undiluted WGA product (~1 µg), and 1:10 diluted WGA product (~100 ng) were used. This experiment resulted in 17,280 genotypes for each experiment, which were analyzed for concordance by means of MS Excel. Failed assays and null alleles were excluded from the analysis (max. 2,200 gt). Only 3 disconcordant genotypes were found between WGA-DNA and gDNA, which gives a concordance rate of more than 99.98% (15,077/15,080). Genotype clusters were even tighter in the polar plot and low-performing assays worked significantly better with WGA-DNA compared to gDNA. No important difference was seen between diluted and undiluted WGA-DNA.

Another experiment was carried out to validate different WGA kits that were being used throughout these studies: GenomiPhi version 1, version 2, and high yield (HY). Amplification products of v1 and v2 were diluted 1:10, while the HY product was diluted 1:40 with 1x TE-buffer. For this test, WGA was performed for the same 92 DNA samples according to the protocols. Control pool A, containing 48 distinct assays, was used for genotyping. No difference was seen in the concordance analyses for all comparisons.

Given the results of these experiments, WGA-DNA was used in SNPlex™ and TaqMan® genotyping experiments throughout this study. The latter was done to spare DNA resources and to meet the demands of the large-scale studies.

3.5.5.2 Genotype concordance between Affymetrix arrays and SNPlex™

Concordance rates were tested between SNPlex™ and array-based hybridization. To this end, genotypes of 210 SNPs that were typed with the Affymetrix 100k GeneChip®, and which were genotyped in a subset of 140 single cases by means of SNPlex™, were compared. 28,186 comparisons were possible and 222 discrepancies were detected. With 99.21%, the concordance rate was convincingly high among the overlapping sample. Excluding three DNA samples with conspiciously high error rates (52/64/24 errors), yielded only 82 remaining errors, increasing genotype concordance to 99.71%. The result is consistent with the genotype concordance of 99.78%, mentioned by Affymetrix. In this experiment, almost 2 million array-based genotypes were compared with the HapMap Release 8.

Only assays that passed the following quality criteria were used for this comparison:
CR > 95%, $p_{HWE} > 0.05$, MAF >5%. Therefore, this approach is biased towards common and high-quality SNPs as they were used in the actual analysis of the screening.

Except for the lead SNPs of the three susceptibility regions (rs2241880, rs951199, rs1793004, rs1992660, rs1992662), no further investigations were carried out with regard to concordance rates between SNPlex™ and TaqMan®. This was extensively studied by De La Vega and colleagues (2005). By comparing 3,901 genotypes (47 assays, 83 DNAs), they found a genotype concordance of 99.7% between the two systems.

Given the results of these experiments, false positive or false negative findings due to genotyping errors can be neglected.

3.6 Functional experiments for *ATG16L1*

Genetic findings need functional validation to support and elucidate the etiology of corresponding suceptibility genes. The previously found *ATG16L1* gene was further characterized, given the clear genetic replication.

3.6.1 Detection of differentially spliced transcripts

Recently, the existence of multiple splice variants of *ATG16L1* was reported (Zheng *et al.*, 2004) and many such variants are annotated in the Golden Path assembly. In all annotated and reported splice variants, exon 9 (containing susceptibility variant rs2241880) is contained in the same frame orientation, thus consistently leading to a Thr to Ala amino acid substitution. The results of Zheng and colleagues were verified by cloning and subsequent sequence analysis of 233 clones. Inserts were amplified from colon, small intestine spleen, and inflamed bowel tissue by using the ATG16L_ex1-9 primers listed in table 8-4, page 163. 42% of the *ATG16L1* transcriptome is represented by the splice variant α, namely the RefSeq AY398617.

3.6.2 Expression in various tissues

Expression of the *ATG16L1* gene was investigated by RT-PCR in a panel of different tissues, confirming expression in the tissues of interest: colon, small bowel, intestinal epithelial cells, and immune-tissues like spleen and leukocytes (fig. 3-9). Strong expression was seen in colon, kidney, prostata, and thymus tissue. Primers covered the constitutively expressed exons 10, 11, and 12 to avoid amplification from genomic DNA. Sequences of the primers are listed in table 8-3, page 163.

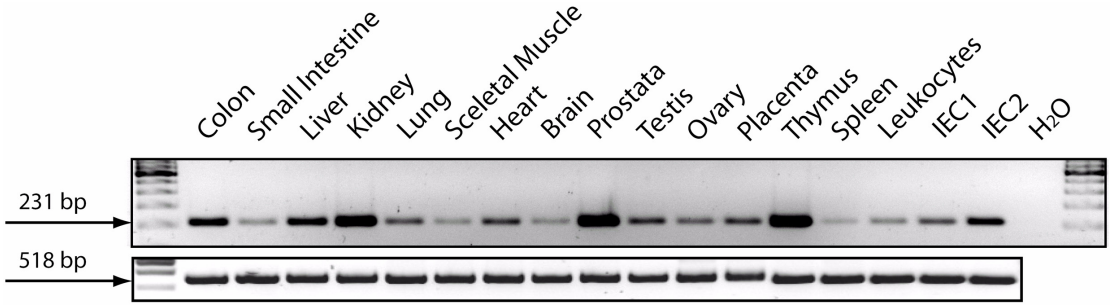

Fig. 3-9 RT-PCR in multiple tissue panel for *ATG16L1*. Expression of *ATG16L1* in a set of different tissues as detected by RT-PCR. The corresponding β-Actin control (518 bp amplicon size) is given below. IEC: intestinal epithelial cells.

3.6.3 Expression in stimulated cell lines

It is not fully clear yet, which pathways ATG16L1 is involved in. Therefore, it was evaluated whether *ATG16L1* expression is regulated in response to different stimuli. As ATG16L1 is involved in autophagy and thus might play a key role in bacterial defense, cells were mainly stimulated with bacterial antigens, such as flagellin, LPS, and *Listeria monocytogenes*. In addition, substances, which are involved in inflammatory cascades, e.g. TNF-α and IFNγ, were used for stimulations

Stimulation of different cell lines and real-time PCR was carried out as described in 2.10.1 on page 71. Fold-changes and standard deviations were computed and the results are plotted in the diagram on the next page.

Overall, most significant changes in *ATG16L1* gene expression were seen after 2 hours. Interestingly, different cell lines respond differently to the same stimulus.

A 2.0-fold downregulation is seen in THP-1 cells after 1 hour, when challenged with *Listeria monocytogenes*. Downregulation increases with time and peaks at 2 hours after stimulation (3.5-fold change). This is supported by a small deviation from the mean of two independent stimulation experiments and is specific for THP-1 cells, as no significant changes are observed for Caco-2, HeLa, and HT-29 cell lines. TNF-α has a similar effect on *ATG16l1* gene expression as *Listeria*, but the the downregulation is not as strong with a 2.3-fold change after 2 hours of incubation. The third stimulation experiment that exceeded a threshold of 1.5-fold, was the stimulation of HT-29 cells with EGF. After 2 hours, an almost 2 times higher expression level is seen than in the 0 hour control experiment.

In summary, no clear pattern is observed and only three of the twenty-three examined combinations provide a small insight into the *ATG16L1* pathway. Further studies are required to replicate these findings and to improve the understanding of the autophagy pathway.

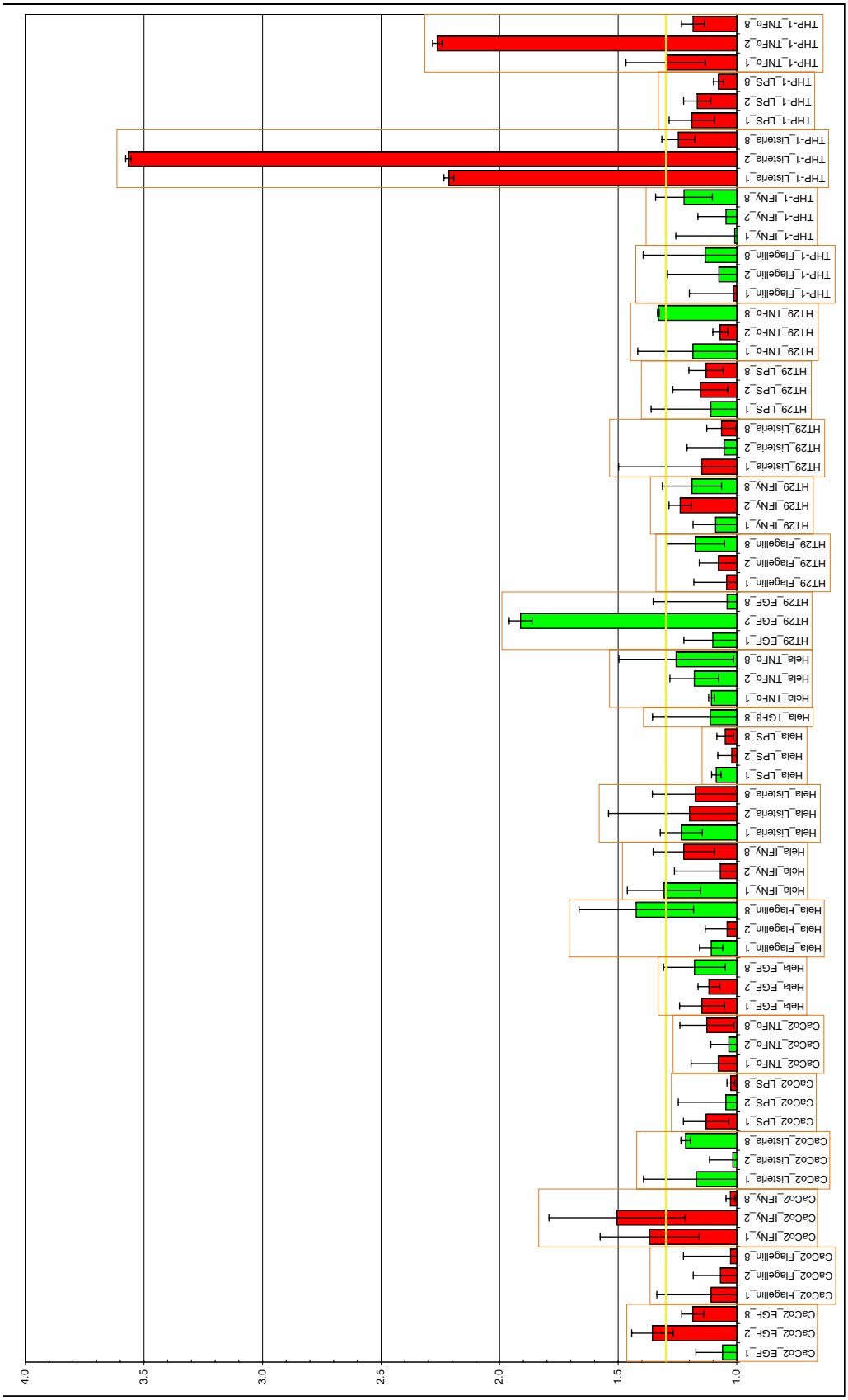

Fig. 3-10 Fold-changes of *ATG16L1* gene expression after different stimuli. Fold-changes are shown on the y-axis and different stimulation experiments are listed along the x-axis with related stimulations framed. A yellow line is shown to indicate the fold-change threshold of 1.3. Green bars highlight an upregulation, while red bars indicate downregulation. Standard deviations are plotted with each bar.

3.6.4 Immunohistochemistry

In a Western Blot from colon tissue (fig. 3-11), a dominant 68.2 kD protein band was identified corresponding to the annotated coding sequence AY398617, the predominant splice variant alpha. The derived protein sequence was therefore used for the modelling of the ATG16L1 protein (see fig. 3-7).

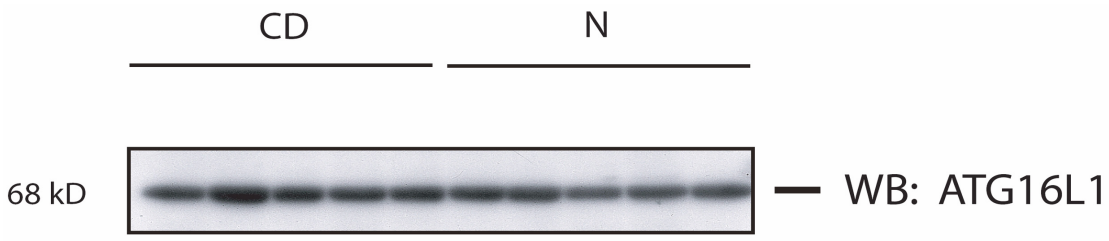

Fig. 3-11 Western blot analysis of ATG16L1 in colonic mucosa. Proteins (15 µg) from rectal mucosal biopsies of CD patients and normal controls (N) were separated by denaturing SDS-PAGE, transferred onto PVDF membranes and probed for presence of ATG16L1 using a specific primary antibody and horseradish-peroxidase-coupled secondary antibody. ATG16L1 is present in the mucosa of CD patients and healthy controls at the same level.

Presence of ATG16L1 in the intestinal epithelium was shown by immunohistochemistry staining (fig. 3-12). No significant difference in expression level and protein distribution was detected between normal and patient tissue.

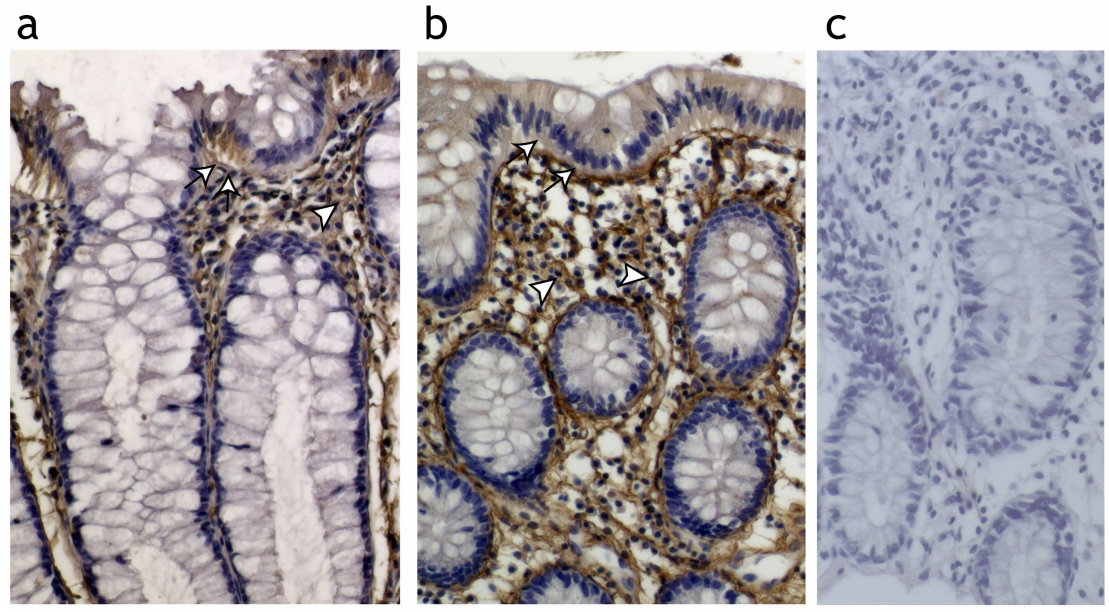

Fig. 3-12 Expression and localization of the ATG16L1 protein in colonic tissue. (A) CD patient, (B) normal control, (C) control staining without the primary antibody in a CD sample. Intestinal epithelial cells are marked with arrows, mononuclear cells are highlighted with arrowheads.

3.7 Replication of previously reported NOD1/*CARD4* polymorphisms

Recently, an association between IBD and variants in the *CARD4* gene, which encodes the protein NOD1, has been shown in the British population (McGovern *et al.*, 2005). In that study, McGovern and colleagues identified the common deletion allele of the marker ND_1+32656 as a risk factor for IBD. Conversely, the minor ND_1+32656 allele, in haplotype combination with its adjacent marker rs2907748, exhibited a protective effect.

CARD4 is located on the minus strand of chromosome 7p15.1 between 30,430,675 and 30,484,790 bp (NCBI build 35). Fourteen exons code for the 953 aa large protein NOD1, that is a structural homologue of NOD2. As *CARD15* replicated in both genome wide screenings and *CARD4* appeared not among the top significant hits, its association with IBD in the German population was evaluated in more detail. SNP genotyping was undertaken on a case-control and an independent family-based panel:

- 1015 IBD patients (676 CD, 344 UC, 886 unrelated controls)
- 775 trios (328 CD, 447 UC)

Given these sample sizes, the study provided >96% power to detect an OR of 2.0, assuming a risk allele frequency of 79% and 1% type I error. However, no significant association was detected in the German population between IBD and the previously reported variants.

Genotyping was performed with TaqMan® and in addition to ND_1+32656, the adjacent single nucleotide polymorphisms ND_1+27606 (rs2075822) and ND_1+45343 (rs2907748) were included for haplotype analyses. ND_1+233 and ND_1+21984, which were only weakly associated with IBD in the study of McGovern *et al.*, were not examined here.

Phenotype	Variant	$MAF_{controls}$	MAF_{cases}	CCA P-value	CCG P-value	TDT P-value
CD	rs2075822	0.81	0.80	0.569	0.308	1.000
	ND_1+32656	0.79	0.79	0.847	0.977	0.091
	rs2907748	0.79	0.77	0.420	0.578	0.306
UC	rs2075822	0.81	0.78	0.152	0.357	0.480
	ND_1+32656	0.79	0.76	0.124	0.295	0.951
	rs2907748	0.79	0.75	0.073	0.195	0.956
IBD	rs2075822	0.81	0.79	0.263	0.327	0.580
	ND_1+32656	0.79	0.78	0.510	0.804	0.248
	rs2907748	0.79	0.77	0.132	0.279	0.547

Table 3-18 Summary of single-marker association statistics for *CARD4*. All three assays had a call rate >95% and showed no significant deviations from Hardy-Weinberg equilibrium. P-values for alleles (CCA), genotypes (CCG), and the TDT are listed.

Variations	Haplotype	$f_{controls}$	f_{cases}	CC P-value	f_T	f_{NT}	TDT P-value	D'
rs2075822 + ND_1+32656	1 – 1	0.771	0.762	0.846	0.455	0.486	0.094	0.91
	1 – 2	0.042	0.040		0.140	0.101		
	2 – 1	0.013	0.015		0.026	0.057		
	2 – 2	0.174	0.184		0.380	0.357		
ND_1+32656 + rs2907748	1 – 1	0.784	0.775	0.727	0.473	0.519	0.719	0.98
	1 – 2	0.002	0.002		0.011	0.011		
	2 – 1	0.006	0.004		0.014	0.008		
	2 – 2	0.208	0.220		0.503	0.462		

Table 3-19 Two-marker haplotype frequencies, transmission, and association statistics for *CARD4* in IBD. Allele 1 is defined as the major allele. P-values resemble global significances after 10,000 permutations with COCAPHASE or TDTPHASE. Estimated frequencies for cases and controls, as well as transmitted and non-transmitted haplotype frequencies are shown. D' was calculated for unrelated controls. Highlighted in blue is the protective haplotype identified by McGovern *et al.* (2005).

McGovern *et al.* also reported strong association with specific disease subgroups. Therefore, the patient sample was stratified for patients with early-onset IBD (n = 449 <25 years) and patients with fistulae and stenoses (n = 491). Analyses of both subgroups showed no significant association (data not shown).

In conclusion, there is no evidence for an association of the IBD phenotype with the previously reported *CARD4* susceptibility variants in the German population (Franke *et al.*, 2006) suggesting ethnic differences as seen for other disease-associated SNPs previously (Mori *et al.*, 2005). This finding is consistent with the results of the two genome-wide screenings. Seven SNPs in the *CARD4* region were located on the Affymetrix 100k gene chip and only one (rs563092) proved to be borderline significant with a p-value of 0.02 in the allelic case-control comparison.

4 Discussion

The two performed genome-wide association screenings of this thesis have led to the identification of two replicable disease susceptibility genes and one region. As the initial screenings identified hundreds of significant SNPs, among which only a few replicated in the follow-up studies, possible influencing and disturbing factors require detailed consideration. Furthermore, the potential roles of *ATG16L1* and *NELL1* in disease-associated pathways, and how this adds to the understanding of inflammatory bowel diseases, need to be discussed.

4.1 Potential methodological pitfalls

4.1.1 Multiple testing and false positive results

Statisical testing of thousands of SNPs in a genome-wide association study, as done in this thesis, is the most likely reason for a large number of false positve results that simply arise by chance. Besides the real association between a disease and a causative marker, or a marker in LD with the latter, false positive association signals can result by several reasons. False positive findings or type I errors describe the scenario that the Null hypothesis of "no association" is incorrectly rejected, leading to a "false alarm". On the $\alpha = 0.05$ significance level, 5% of the results are likely to be false-positives. Since α has to be greater than 0, as for $\alpha = 0$ the null hypothesis will always be accepted, a type I error will always exist.

	Reality	
Result of test	**H_0 is true (NO association)**	**H_0 is false (association)**
H_0 is accepted	correct decision with $1-\alpha$	type II error with β
H_0 is rejected	type I error with α	correct decision with power $1-\beta$ (see 4.1.3 on page 128)

Table 4-1 Test results and reality. *According to Sachs (2003).*

It is generally agreed that α must be defined before the actual testing. Therefore, the p-value is used as an *a posteriori* measure of significance, i.e. it is is the probability to err if the null hypothesis is rejected. A p-value smaller or equal to 5% is generally considered significant. If thousands of alleles are tested, then a nominal α equal 0.05 is unacceptable, as far too many results will simply arise by chance statistical fluctuation, which has economical consequences in terms of minimizing costs per true positive. Decreasing α will only result in a loss of power (see 4.1.3 on page 128) and therefore, different methods exist to estimate the rate of false positives by interpretation of the p-value.

A simple and analytically straightforward approach is to use Bonferroni correction that assesses the probability of a hypothesis being correct by incorporating the prior probability of the hypothesis and the experimental data supporting the hypothesis.

Individual p-values are corrected for multiple-hypothesis testing by multiplying the p-value with the number of tests "n" (Plenge *et al.*, 2006; Bland *et al.*, 1995):

$$P_{corrected} = n \times P_{uncorrected}$$

This would mean a corrected value of p = 0.05 for a p-value of 5×10^{-7} in a genome-wide screen of 100,000 SNPs. As this is overly conservative, if an allele has a very high probability of being a true-positive result, replication studies were used throughout this thesis to filter out false positive results rather than Bonferroni correction. Only marker rs2076756 in the *CARD15* gene had a p-value below 5×10^{-7} in the 100k scan and would have been considered a truly significant association finding. In April 2005, a broad consensus emerged at a workshop for genome-wide association studies (Thomas *et al.*, 2005) that multi-staged designs are the most reasonable way to conduct whole-genome scans. The different levels of replication allow to correct for type I error rates, besides reducing genotyping costs without any loss of power. Sobell *et al.* (1993) first suggested a two-stage design for association studies, which has recently been extended to a genome-wide scale by Satagopan and colleagues (2002; 2003; 2004), Lowe *et al.* (2004), and van den Oord *et al.* (2003).

Another effect that can cause false positive or false negative results are ethnic differences or admixture among the group being studied, so-called population stratification (discussed in section 4.1.4 on page 130). Misclassifications of patients and selection bias are also confounding factors, that researchers need to be aware of. Selection bias arises when the case and control groups are not truly comparable and a difference exists between what is actually estimated and the true effect of interest. It is mainly due to effects on personality leading to "volunteer bias", which might affect a limited number of loci, and more general effects that are due to differences in population substructure between cases and controls. A "healthy volunteer bias" could be an issue of the used control group in the replication study as it partially consists of blood donors. The latter are considered to be exceptionally healthy persons that are self-selected on the basis of better lifestyles (Vineis *et al.*, 1998).

Although not really affecting the results of this study, publication bias (Colhoun *et al.*, 2003) is also a pitfall in association studies. This so-called "Winner's Curse" phenomenon describes the fact that the original study often overestimates the true magnitude of the genetic effect, since larger sample collections are used in following replication studies. Furthermore, mostly positive results are submitted for publication, let alone accepted.

4.1.2 Coverage (pitfalls of auto-calling)

LD-based whole-genome scans strongly rely on a good genomic coverage of markers. This means that the used SNP set must capture as much of the total genomic variation as possible. It must also be mentioned that genome copy number variations (CNVs) are far more frequent than originally expected (Aitman *et al.*, 2006; Sebat *et al.*, 2004), and several genetic disorders are the result of CNVs. However, methods to evaluate CNVs are just beginning to evolve and it should be kept in mind that the performed SNP screenings are not suitable to detect CNVs and their contribution to disease susceptibility. It has been shown just recently that a lower human beta-defensin 2 (*HBD-2*) gene copy number in the beta-defensin locus predisposes to colonic CD (Fellermann *et al.*, 2006). The ultimate strategy to identify genetic variations that influence disease is complete resequencing in clinical samples, thereby comprehensively testing all variants across the genome. Resequencing thousands of genomes is not yet feasible because of economic and technical constraints, although the 1000$ genome is within reach (Bennett *et al.*, 2005). Therefore, genomic coverage of the used SNP sets of this thesis, which need to represent the entire genomic variation, must be carefully evaluated. Barrett *et al.* (2006) have calculated coverage rates for the currently available technologies for GWS. They determined a coverage rate of 31% for the Affymetrix 100k GeneChip®. This low coverage and efficiency is mainly due to the underlying LD "agnostic" marker set, i.e. SNPs were randomly selected instead of using tagSNPs. Therefore, it can be assumed that many gaps exist, where no information about any association can be derived. Nicolae *et al.* (2006) observed that SNPs in the 100k set are undersampled from coding regions and oversampled (57.5% vs. 59.6%) from regions outside genes relative to the distribution of SNPs in the HapMap. Adding to the problem of low coverage, marker numbers for a real whole-genome scan are calculated to be 500,000 and even a lot more, when randomly distributed SNPs are used. Therefore, the screening described in this thesis is more a "genome-wide" scan than a "whole-genome" study and certainly, a lot more susceptibility loci could be identified with another technology. However, random SNP sets offer a protective redundancy against the failure of any given SNP and their disadvantage in coverage compared to population-specific tagSNP sets is reduces, when other populations are used that have a different LD pattern. Finally, when this study started, the Affymetrix 100k GeneChip® was the first and only publicly available technology that made it possible to conduct genome-wide association mapping.

Since non-synonymous SNP sets are designed to test functional variants directly, coverage rates are not of major concern for these SNP sets. However, to conduct a genuine whole-genome nsSNP study, all known and validated nsSNPs should be typed and as many as possible of the 28,000 existing genes and more should be "covered" (Roest Crollius *et al.*, 2000; Ewing *et al.*, 2000). In the current study, 19,779 non-synonymous SNPs were genotyped, of which

7,159 were informative. Because the initial SNP set "covered" less than 10,000 genes, the term genome-wide experiment is more appropriate in this context.

As existence of linkage disequilibrium between the genotyped SNPs and potentially disease-causing variants is assumed, LD-disturbing factors can lead to false-negative or fale-positive outcomes. Ideally, LD decays with time since the alleles evolved and distance between the loci, but empirical studies have shown that the pattern of LD is very complex and poorly understood. Sometimes LD can be detected between distant markers, while there may be little or no LD between very closely located markers. It has been estimated that physical distance can explain about half of the variation in LD (Abecasis *et al.*, 2001). There are a number of different factors that contribute to the extent of LD in a population, like recombination and mutation rates, gene conversion, selection, genetic drift, growth rate, migration and population structure (Ardlie *et al.*, 2002; Pritchard *et al.*, 2001). Below, these factors are discussed in somewhat more detail.

Recombination

Regional variability of recombination rates has been observed in the human genome. Since LD is highly dependent upon the recombination rate, the pattern of LD will vary between regions of high recombination activity (expected low LD) and regions of less recombination (higher LD is expected). Furthermore, recombination, even within shorter regions, may not be randomly distributed, but instead be restricted to shorter regions of high recombination, so-called recombination hotspots, surrounded by more conserved regions with little or no recombination (Daly *et al.*, 2001; Jeffreys *et al.*, 2001). This has been convincingly shown to be the case in the HLA class II region and in the mouse MHC (Jeffreys *et al.*, 2001; Yauk *et al.*, 2003; Cullen *et al.*, 2002; Kauppi *et al.*, 2003). It appears that at least some hotspots are found across populations and haplotypes (global hotspots; Kauppi *et al.*, 2003), while evidence from the HLA seems to support that hotspot activity can be specified for certain haplotype constellations, i.e. certain haplotypes might be more prone to recombination than others (Kauppi *et al.*, 2003).

Mutation

The mutation rate may vary between different genetic sequences. Recurrent mutations are probably very rare with some exceptions: microsatellites and variable number of tandem repeats for example, are more prone to mutations than SNPs (Brinkmann *et al.*, 1998). Such differences may affect the decline of LD.

Gene conversion

When a short stretch of DNA on one chromosome is transferred/copied to the other chromosome during meiosis, this is termed gene conversion. The event appears to be more common

among humans than previously thought, and is probably relatively important for short-scale LD decay, since gene conversion is able to break down LD between very tightly linked markers without affecting long-range LD between markers located outside of the converted segment (Ardlie *et al.*, 2001; Frisse *et al.*, 2001).

Selection

If there is strong selection for or against a certain variant at a locus, the surrounding haplotype may considerably increase or decrease its population frequency. If the rise in allele frequency occurs over such short time that recombination is unable to break down the haplotype carrying the selected mutation, LD can become very strong for this allele. It has been suggested that as a response to pathogens, selection has been an important factor in maintaining high levels of LD in the HLA-complex (Hedrick *et al.*, 1991; Huttley *et al.*, 1999).

Random genetic drift

Random genetic drift is the stochastic process which describes the change in allele and haplotype frequencies between generations in a finite sample. For example, in a very small population there is a chance that the allele frequency of a specific allele will change somewhat in the next generation, just by the chance transmission of a few more, or less, alleles of one kind. Such changes are unlikely to have an impact on large populations but they can have drastic effects in very small populations where these chance effects will occur until one of the alleles gets extinct, and the other alllele becomes fixed in the population. Genetic drift can also occur through accidents, or so-called “bottle-necks” that leave only a few individual survivors with a limited gene pool. An extreme bottle-neck refers to the founder effect, the latter occurs if only a few founders form a new population. Then the genetic make-up in the new population may be very different from the original population within which the founders lived.

Population growth rate

Rapid growing populations reduce genetic drift, hence there is a decrease in LD.

Population structure and migration

Population substructure (see also 4.1.4 on page 130) can produce significant LD even if no linkage exists.

The extent of LD in the human genome has been highly debated and estimates range between 3 kb and several 100 kb (Abecasis *et al.*, 2001; Ardlie *et al.*, 2002; Pritchard *et al.*, 2001; Reich *et al.*, 2002; Kruglyak *et al.*, 1999; Reich *et al.*, 2001). However, as several authors pointed out, knowledge of the average extent of LD between random markers may not be a very useful tool for designing association studies, since there is such great variance of LD in different regions (Nordborg *et al.*, 2002).

4.1.3 Power

In contrast to false positive results that are of minor importance in the described multi-tiered scans, overlooking true association signals poses a greater threat. While the type I error rate α is the probability that a true null hypothesis is incorrectly rejected, the type II error β describes the failure to detect a false null hypothesis ("a chance is missed"). Power is the probability to reject H_0, if H_A is correct and it is calculated as 1-β. All three measures are correlated as with decreasing error probability α, type II error β increases to result in decreasing power. A summary of this relationship is given in the following diagram:

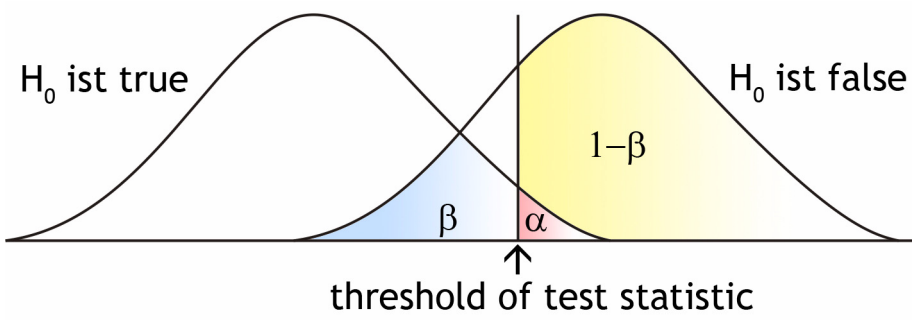

Fig. 4-1 Threshold of test statistic in dependence of α and β. With decreasing α, power (1-β) decreases.

To circumvent the problem of being "underpowered", thus missing true association signals, sample sizes must be increased. Given the sample sizes of the initial screenings, power can be calculated to detect SNPs of different allele frequencies and different effect sizes:

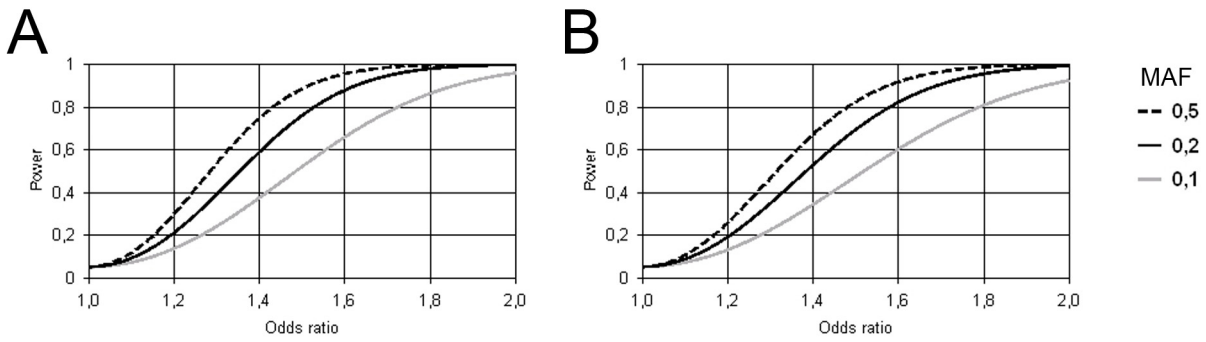

Fig. 4-2 Power estimation. Graphical representation of a power estimation in the sample size at a significance level of α = 0.05 for a two-sided test over an odds ratio range of 1 to 2. The graph was generated using PS-power (Dupont *et al.*, 1997) and shows the test power as a function of the odds ratio (x-axis). (A) Shows the power calculation for the cSNP and (B) of the LD-based scan. Different minor allele frequencies (MAF) were used.

As depicted in fig. 4-2, both screenings provided sufficient power of 80% to detect SNPs with a minor allele frequency of 20% and an odds ratio of 1.6. Of course, it would be desirable to have sufficient power to detect even low frequency SNPs with moderate to low effects. Many investigators argue that complex diseases are caused by numerous loci with small effects and few genetic loci with large effects. The potential for a large number of variants with small in-

dividual contribution to human phenotypes is further supported by recent findings that allelic variation frequently affects gene expression and exon splicing (Pagani *et al.*, 2004; Hoogendoorn *et al.*, 2003; Lo *et al.*, 2003; Mira *et al.*, 2004), which is likely to have smaller effects than polymorphisms that affect the coding sequence. A genome-wide association study that could detect rare variants (MAF = 0.05) with small effects (OR = 1.3) would need large sample numbers of 4,282 cases and 4,282 controls to provide a descent power of 80% ($\alpha = 0.05$). Although this is theoretically and technical feasible, few laboratories possess this large sample collections for a certain disease, which is in some cases also rare in the population. In addition, genotyping of 8,000 samples and more can only be afforded by large research centers with sufficient funding. The Institute of Clinical Molecular Biology in Kiel has for example a CD patient panel of app. 3,000 samples, which is close to the required number of 4,282 samples. According to Wang *et al.* (2005) unrealistically large sample sizes of even more than 10,000 cases and 10,000 controls would be required to achieve convincing statistical support for a disease association if effect sizes are less than 1.3 and allele frequencies below 10%. Therefore, the two screenings described in this thesis are a compromise between sufficient power, genotyping costs, limited sample size, and it is likely that rare (MAF < 20%) associated SNPs with an OR below 1.6 were missed. Using larger numbers of patients and controls for the replication experiments, minimizes the chance of missing associations that were found in the previous scan. Replication panel B consists of 878 cases and 1032 healthy controls and has 80% power to detect an odds ratio of 1.36 (MAF = 0.2; $\alpha = 0.05$). For the equivalent conditions, the 380 used trios have sufficient power to detect an effect size of OR = 1.51.

4.1.4 Population stratification

As even a small extent of population admixture could undermine the association studies of this thesis and lead to false positive results, it has to be discussed whether the two screenings could have been affected by this confounding factor. The term population stratification describes the confounding effect that a population under study consists of a mixture of two or more subpopulations that have different allele frequencies and disease risks. An example is given in the following figure:

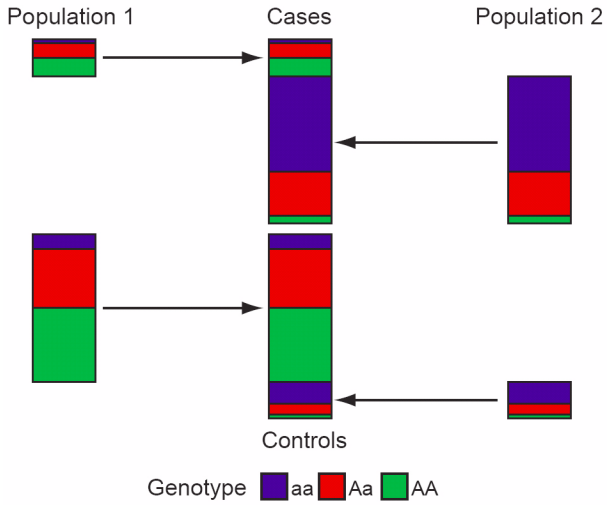

Fig. 4-3 Effects of population structure at a SNP locus. If the study population consists of subpopulations that differ genetically, and if disease prevalence also differs across these subpopulations, then the proportions of cases and controls sampled from each subpopulation will tend to differ, as will allele or genotype frequencies between cases and controls at any locus at which the subpopulations differ. *Illustration from Marchini et al. (2004).*

The extent to which population stratification poses problems to association studies has been widely debated (Wacholder *et al.*, 2000 and 2002; Thomas *et al.*, 2002; Cardon *et al.*, 2003; Freedman *et al.*, 2004). Population substructure can lead to three distinct problems: confounding, cryptic relatedness (resulting in overdispersion of the test statistic), and selection bias. Unlike some other biases, these problems do not become smaller with increasing sample size, on the contrary, type I error rate even increases.

For these case-control experiments of this thesis, control individuals were mainly drawn from the POPGEN biobank (Krawczak *et al.*, 2006), thus having a Northern German background. Since the sampling area is restricted to Kiel and surrounding cities that are "enclosed" by the canal in the south, the Danish border in the north, and the sea to the west and east, this area is thought to consist of a very homogenous population. Although samples from Northern German patients were preferentially selected to match the control population, a large proportion of samples was used from patients that live in different parts of Germany. Further concern of an inflated type I error rate due to population stratification is driven by

the few significant SNPs that were replicated in the second tier. Evidence of population stratification can be obtained by genomic control experiments (GC; Devlin *et al.*, 1999 and 2001), structured association (Pritchard *et al.*, 2000), and simple logistic regression (Tang *et al.*, 2004). For the the still ongoing German genomic control study (Lamina *et al.*, 2005), 720 individuals were randomly drawn from the POPGEN (Kiel, Northern Germany; Krawczak *et al.*, 2006), KORA (Augsburg, Southern Germany; Holle *et al.*, 2005), and SHIP (Greifswald, Eastern Germany) sample and genotyped for 210 SNPs. Of the latter, 140 were chosen from non-functional region of the genome ("genomic desert"), since these loci are presumably not under selective pressure and thus, are only subject to neutral processes of drift and migration in demographic history. The other 70 SNPs were chosen from intragenic regions. First results show very low and non-significant F_{ST} values of less than 0.001 for the three sample collections under study (Caliebe and Krawczak, *personal communications*). Therefore, no significant population stratification seems to exist in Germany compared to, for example, populations from the Caucasus that have F_{ST} values of 0.113 (Nasidze *et al.*, 2001). Interestingly, the European population in total is genetically homogenous (F_{ST} = 0.017), in contrast to the African (F_{ST} = 0.086) or American (F_{ST} = 0.038) population (Stoneking *et al.*, 1997; Nasidze *et al.*, 2001). Since no significant population substructure seems to exist in Germany and therefore is likely to be absent in the used case-control panels, plus the fact that the family-based TDT test was employed, which is robust against such substructure, type I errors have arisen rather by multiple testing, i.e. chance, or other selection biases.

Clinical heterogeneity is also of importance in a genetic study, if a genetic variant influences a disease subset. This confounding becomes even more complex, if the screening group differs significantly in its clinical presentation from the replication sample. Because inflammatory bowel diseases present a variety of distinct clinical phenotypes, it is important to investigate the extent to which this heterogeneity impacts variability in the risk of developing common diseases. This was done for the *ATG16L1* mutation rs2241880, by testing its association in different CD subphenotypes, but no differences were seen for different clinical outcomes. It could be considered to analyze associations with subphenotypes on the genome-wide level, but as this significantly reduces the already small sample size of the patient group by 30% and more, power would drastically drop, thus running into problems of overlooking true association signals. Furthermore, since the number of possible subgroup analyses that can be undertaken is large, significant results obtained in such analyses should be treated with even more scepticism than tests for the average genetic effect across all subgroups combined. The selection of "genetically enriched" cases on the basis of severity, as done for most cases of the 100k scan, could also have the counterproductive effect of enriching environmental as well as genetic factors.

4.1.5 Transmission distortion

In the present study, more association signals, which were significant in the case-control replication panel, could not be replicated in the family-based study group composed of trios than vice versa. For the cSNP scan eleven SNPs replicated in the case-control analysis and not in the TDT, while six were only significant in the TDT. The corresponding numbers were 17 and 12 for the 100k scan, thus less SNPs are replicated in the TDT analysis. Either population stratification among cases and/or controls or transmission distortion could be a result of this imbalance. Several known biological processes lead to skewed transmission probabilities among surviving offspring and departure from Mendelian expectation has been observed by several groups (Zöllner *et al.*, 2004; Friedrichs *et al.*, 2006; Warburton *et al.*, 1983). Processes that result in transmission distortion are (according to Pardo-Manuel del Villena *et al.*, 2000 and 2001):

- Meiotic drive - biased segregation during meiosis
- Gametic selection - differential success of gametes in achieving fertilization
- Postzygotic viability (embryonic mortality) selection for or against particular genotypes, so-called *in utero* selection

It is known that a large fraction of conceptions, perhaps as many as 75%, end with early embryonic loss (Edmonds *et al.*, 1982; Regan *et al.*, 2000; Macklon *et al.*, 2002). Some fractions of these losses are presumably due to genetic factors, including genetic incompatibility between the mother and fetus or inviability of the fetus's genotype. These forms of selection would skew the observed transmission probabilities among those offspring that do survive to term, with fitter genotypes being transmitted more frequently. Interestingly, an increased activity of CD in the beginning of pregnancy causes high rates of prematurity, spontaneous abortions, and stillbirths (Briese *et al.*, 1993). Friedrichs *et al.* (2006) also observed gender-specific transmission distortion of the R30Q risk allele, a variant that is present in the gene *DLG5* and is associated with CD. This male-specific association might also explain, why none of the 4 SNPs in the *DLG5* gene were not significant in the 100k genome-wide scan.

Transmission ratio distortion might also explain why the TDT test was not significant for the *NELL1* mutation rs8176785, which was highly significant in the case-control analysis. *NELL1* knock-out mice embryos can only be rescued by caesarian delivery and die a few hours after birth (Desai *et al.*, 2006). This finding is further discussed in 4.3.2 on page 139.

4.1.6 Interaction

In recent years it has become obvious that for common diseases there may be more complex interactions among genes with and without strong independent main effects. The presence of epistasis is a particular cause of concern, if the effect of one locus is altered or masked by effects at another locus. These effects are more difficult to detect using traditional methodologies and the number of studies documenting epistasis is small (Ritchie *et al.*, 2005). The difficulty in detecting epistasis stems from the "curse of dimensionality" (Bellman *et al.*, 1961). Stated simply, as the number of parameters increases, the number of interaction terms grows exponentially and most parametric statistical methods are limited to deal with interaction data involving many simultaneous factors. Although recent publications show that it is computationally feasible to detect multiple loci that influence complex diseases on a genome-wide level (Marchini *et al.*, 2005), Krawczak and colleagues (*personal communications*) have calculated that it takes 3 days, 264 years, and 7×10^6 years to evaluate 2, 3, and 4 SNP-interactions, respectively, in a genome-wide screening of 100,000 SNPs using a Pentium 4 PC with a 2 GHz processor (4 GFLOPS). Even on the grid project GIMPS with 17,000 GFLOPS it would last 2,000 years to calculate all four-wise SNP interaction terms. Nevertheless, the question of true biological interaction remains of paramount interest, but may ultimately be better answered via molecular, rather than statistical, investigation. A good review of statistical methods for epistasis is given by Cordell (2002).

In summary, genetic association studies become complicated and computationally intensive when the simple correspondence between genotype and phenotype breaks down due to effects of chance, environmental factors, incomplete penetrance of mutations with moderate effects, interaction phenomena, or genetic heterogeneity.

4.2 Common disease, common variant

It is still discussed whether one should intend to search for common polymorphisms having main effects on disease risk or those having modifying effects on other genes or environmental factors, as well as prior beliefs about the "common disease/common variant" (CD/CV; Cargill *et al.*, 2000; Reich *et al.*, 2001; Lohmueller *et al.*, 2003) versus "multiple rare variant" (Pritchard *et al.*, 2001; Pritchard *et al.*, 2002; Wang *et al.*, 2003; Fearnhead *et al.*, 2004) hypotheses. The most prominent example for the "multiple rare-variant" theorem is *CARD15*, where the risk alleles are rare among the general population (1.3-4.7%; see table 1-4, page 13). The allelic spectra of most common diseases probably fall between these two extremes (Wang *et al.*, 2005). Since in this study a common variant associated with CD was found, the "common disease/common variant" hypothesis is strongly supported. 27% of controls (vs. 36% in cases) carry the risk genotype "GG" of the non-synonymous SNP rs2241880 in the *ATG16L1* gene. Adding evidence comes also from a study by Valentonyte *et al.* (2005), who found a truncating splice site mutation (rs2076530) in *BTNL2*, which is associated with sarcoidosis and has a minor allele frequency of 42% in unaffected individuals. Other examples are SNP R30Q associated with CD (Stoll *et al.*, 2004) with a MAF of 10% in healthy normals and SNP rs7566605 that is associated with obesity and has a MAF of 37% (Herbert *et al.*, 2006). Finally, the identification of rather common genetic risk factors in complex diseases partially explains the steep increase in prevalence of barrier diseases since the beginning of the 20th century, i.e. actual genetic predisposition develops in response to the drastic environmental changes ("yesterday's neutral SNP, today's risk factor"). Rare and highly penetrant risk alleles that exert their effect "agnostic" to environmental influences are subject of negative selection as they significantly reduce the reproductive fitness of an individual. These kind of mutations are rather known from rare Mendelian diseases, e.g. cystic fibrosis, than for common complex traits.

4.3 Possible roles of *ATG16L1*, *NELL1*, and the 5p13 locus in IBD

4.3.1 ATG16L1 is involved in autophagosome formation

In this thesis, a new susceptibility variant for CD in *ATG16L1* was identified and disease association was replicated in independent German and UK samples. The mutation rs2241880, which leads to an amino acid substitution, yielded a highly significant result both in a case-control comparison (overall $p = 4.0 \times 10^{-8}$) and by the TDT ($p = 2.7 \times 10^{-5}$). The failure to detect any other disease-associated coding or regulatory variants in the *ATG16L1* gene in a systematic mutation search, and in regression and haplotype analyses, suggests that the CD risk at the *ATG16L1* locus is conferred mainly by rs2241880 alone, and that additional variants do not contribute significantly to the observed association. The disease-associated “G” allele is frequent (0.53 in Germans) and shows odds ratios of 1.45 for carriership and 1.77 for homozygosity. It therefore represents an example of the "common disease - common variant" paradigm (Reich *et al.*, 2001). A statistical interaction between rs2241880 and the established *CARD15* disease mutations was noted (table 3-16, page 110).

The disease-associated variant rs2241880 leads to an amino acid exchange (Thr to Ala) at position 300 of the N-terminus of the WD-repeat domain (see fig 3-7, page 112) in ATG16L1 (other aliases: APG16L, WDR30). The interaction partner of the WD domain in ATG16L1 has not yet been identified experimentally, but during their studies of mouse ATG16L1, Mizushima *et al.* (2003) found a p144 protein that remains a good candidate. It is clear, however, that in yeast the ATG12-ATG5 conjugate, which is required for autophagy, assembles in a multimeric complex via the coiled-coil region of ATG16 (Mizushima *et al.*, 2003; Kuma *et al.*, 2002). ATG16 interacts with the conjugate through ATG5, and ATG16 homo-oligomers formed by the coiled-coils connect multiple ATG12-ATG5 conjugates. Furthermore, it has been shown that the ~350 kDa large ATG12-ATG5·ATG16 complex is necessary for autophagosome formation and localizes to the so-called preautophagosomal structure (Mizushima *et al.*, 2003). In metazoans, this conjugate is contained in a ~800 kDa protein complex and yeast ATG16 is known as ATG16-like protein 1 (ATG16L1) because it contains an additional WD-repeat domain of as yet unidentified function at the C-terminus (see fig 3-5, page 111; Zheng *et al.*, 2004; Mizushima *et al.*, 2003). In most WD-repeat proteins, seven or eight copies of the WD-repeat form a β-propeller domain structure with blades consisting of four-stranded anti-parallel β-sheets (Li *et al.*, 2001). Due to the circular arrangement of the propeller blades, the N-terminal strand β1 is included in the C-terminal blade, and this stable β-propeller structure provides an extensive surface for molecular interactions. These interactions may be impaired by the conformational change resulting from the T300A substitution.

Except for the *ATG5*$^{-/-}$ knock-out mice, no other mouse models have been reported for any other *ATG* gene. *ATG5*-deficient mice die within 1 day of delivery due to energy depletion. Therefore, the degradation of "self" proteins seems to be essential during the early neonatal starvation process.

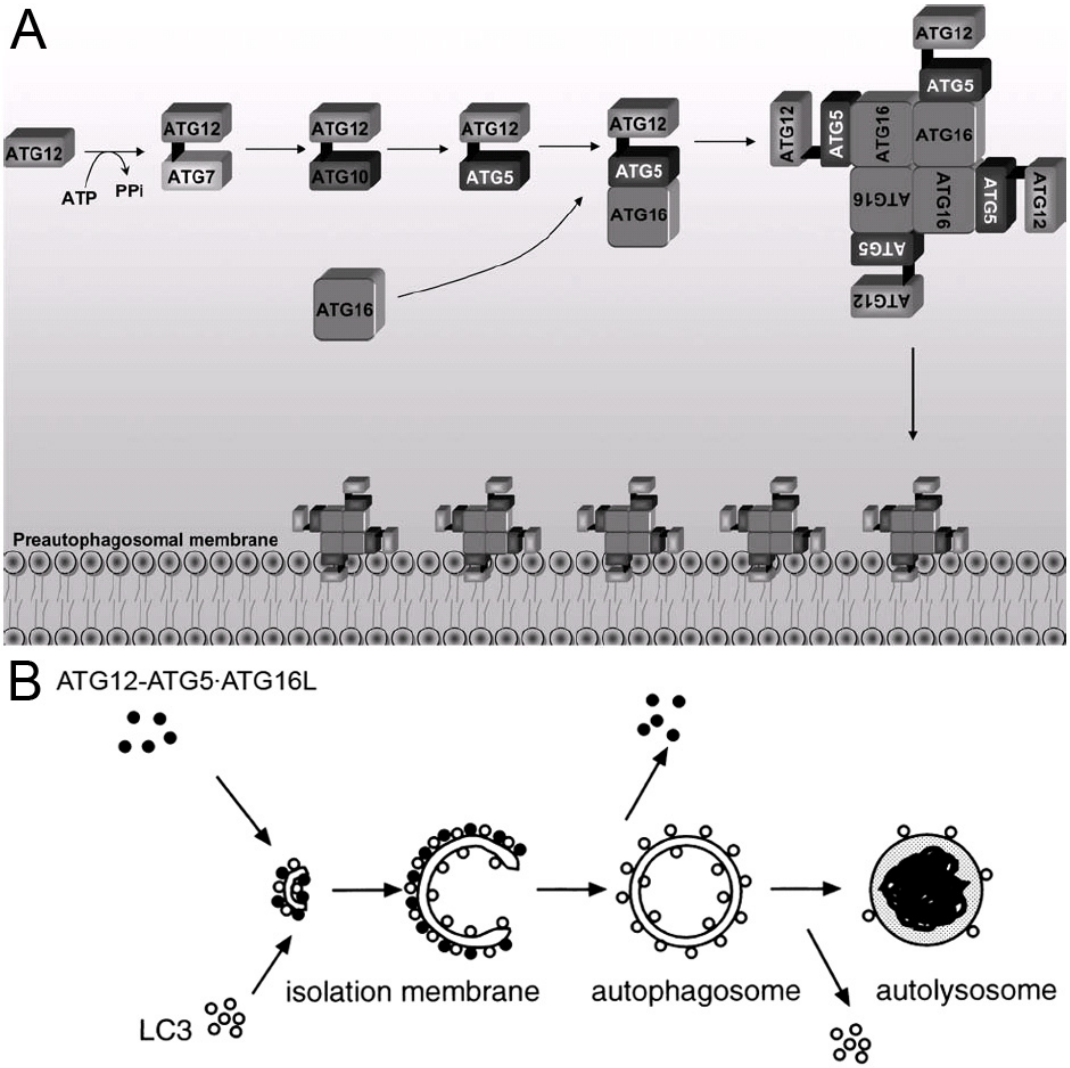

Fig. 4-4 The ATG12-ATG5·ATG16L ubiquitin-like system. (A) In *Saccharomyces cerevisiae*, a chain of reactions that are similar to ubiquitinylation are necessary to form the ATG-complex, which subsequently locates to the preautophagosomal membrane. This complex then promotes autophagosome formation. Tetramer-formation is facilitated through the ATG16 protein, which forms homo-oligomers. *Illustration from Marino et al. (2004).* (B) Model of autophagosome formation in mammalian cells. The ATG12-ATG5·ATG16L conjugate localizes to the preautophagosomal isolation membrane througout its elongation process. LC3, which is homologous to yeast ATG8, is recruited to the membrane in the ATG5-dependent manner. ATG12-ATG5 and ATG16L dissociate from the membrane upon completion of the autophagosome formation, while LC3 remains on the membrane. Therefore, LC3 is commonly used as a marker for autophagosomes. *Illustration from Mizushima et al. (2002).*

The single columnar epithelial lining of the small intestine is the primary barrier restricting transgression by luminal bacteria, endotoxin, bile, digestive enzymes, and antigens while also serving as the principal site for selective transport of electrolytes and nutrients. Bacteria that are able to invade the cytoplasm of host cells are recognized by the innate immune system. Proteins from the NLR (NACHT/LRR or NOD-like receptors) family recognize patho-

gen-associated molecular patterns (e.g. peptidoglycan) and lead to the activation of the innate immune defense, mainly in phagocytes and epithelial cells (Girardin *et al.*, 2003; Inohara *et al.*, 1999; Chamaillard *et al.*, 2003). In particular, NOD2, the protein encoded by *CARD15*, plays a pivotal role in the detection of cytosolic muramyl-dipeptide (MurNAc-L-Ala-D-isoGln; MDP), a fragment of the bacterial cell wall. Autophagy (Greek word for self-digestion) is a fundamental molecular machinery of eukaryotes for bulk protein degradation. It has been implicated in diverse physiological processes such as organelle turnover, starvation response, cell death and defence against invading bacteria, a process referred to as xenophagy (Codogno *et al.*, 2005; Mizushima *et al.*, 2005; Deretic *et al.*, 2005; Swanson *et al.*, 2005). Three differenct mechanisms of autophagy are known: chaperone-mediated autophagy, microautophagy, and macroautophagy (Schmid *et al.*, 2006). The latter mechanism is known to play a major role in several diseases, such as cancer (tumor cells can survive in nutrient-poor environment), muscular disorders (accumulation of autophagosomes), and neurodegeneration (accumulation of aggregated proteins) [Shintani *et al.*, 2004].

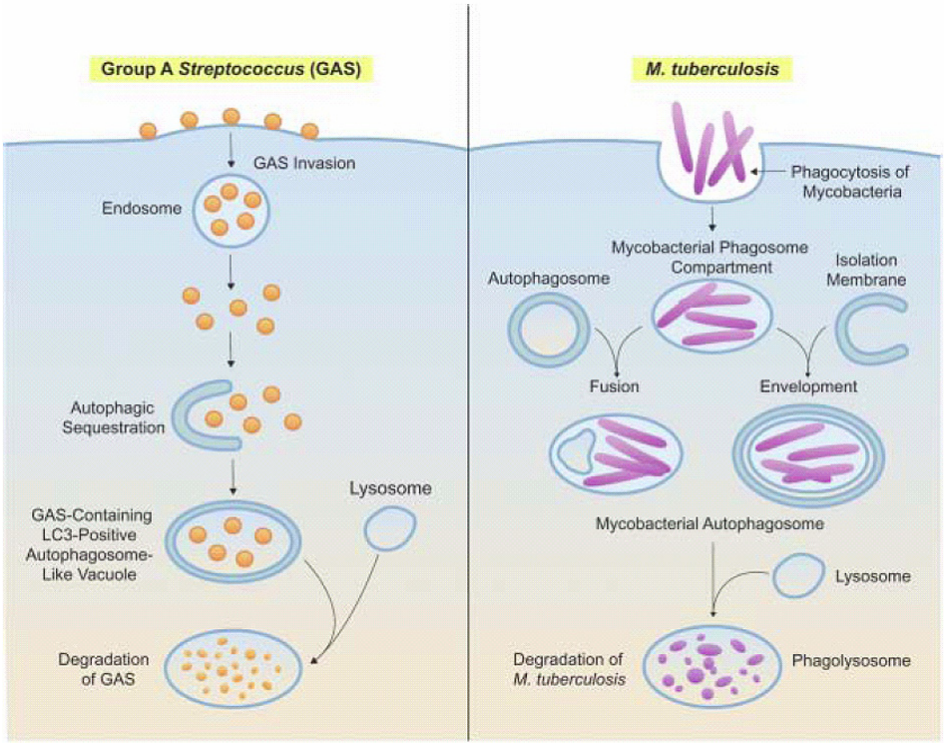

Fig. 4-5 Xenophagy. Cellular defense against invading extracellular pathogens, Group A *Streptococcus* (GAS), in a non-phagocytic cell and an intracellular pathogen, *M. tuberculosis*, in a phagocytotic cell. *Illustration from Levine et al., 2005.*

Evidence exists that xenophagy plays a role in the degradation of both extracellular bacterial pathogens (Nakagawa *et al.*, 2004) and true intracellular bacterial pathogens (Gutierrez *et al.*, 2004; Ogawa *et al.*, 2005). Pathogens trapped by the autophagic membrane are ultimately targeted to the autophagolysosome compartment (Kirkegaard *et al.*, 2004; Ogawa *et al.*, 2005), a pathway of which the ATG16L1 protein is part of. This fusion with a lysosyme (inner pH < 5.0) leads eventually to the degredation of the content by several hydrolytic enzymes (see fig. 4-5).

A genetic interaction between rs2241880 and risk variants in *CARD15* may reflect a connection between the two encoded proteins at the functional level. Both proteins are part of molecular defence pathways and could be additively involved in the interaction with intracellular bacteria. Hence, variants in the encoded proteins could result in an incapacitation of phagocytic cells. Since the risk variants in both *CARD15* and *ATG16L1* are only associated with CD but not with UC, one can hypothesize that an additive functional defect caused by both variant genes could be a disease specific characteristic leading to the pathology found in CD.

While only the risk conferred by SNP rs2241880 in the *ATG16L1* gene could be replicated in other samples, some of the remaining, apparently significant, variants listed in table 3-9, page 96 may still represent important candidates for disease susceptibility. The requirement of a significant replication ($p < 0.05$) in both the TDT and the case-control comparison may lead to the exclusion of relevant variants due to a lack of power. These variants should therefore be evaluated in further, independent large patient samples. While the present study still falls short of comprehensively assessing all putative functional variation in the human genome, it is the largest cSNP study hitherto performed, both in terms of the number of SNPs included and of the size of the database from which they were selected. Our results, together with those of other recent studies, demonstrate that the direct association analysis using cSNPs is a meaningful complement to genome-wide LD-based association studies as the one that has been reported for CD previously (Yamazaki *et al.*, 2005) and the one that was performed in this study.

4.3.2 The place of the nel-like 1 protein in the etiology of IBD

The *NELL1* gene was identified as a susceptibility gene for CD in the 100k association screening of this thesis. Further genotyping efforts in other IBD patient panels replicated the initial finding and showed an association with UC as well. Of the four discovered non-synonymous SNPs in the present study, the mutation Arg82Gln (rs8176785) showed the strongest association.

The neural epidermal growth-factor-like (*nel*) gene was first isolated from an embryonic chicken cDNA library and so named because it was found to be predominantly expressed in neural tissues (Matsuhashi *et al.*, 1995 and 1996). The *NELL1* gene encodes a polypeptide (810 aa) that is glycosylated and processed in the cytoplasm and then secreted as a 400 kDa homotrimer. The protein contains thrombospondin-like, laminin G, von Willebrand factor-like repeats, and epidermal growth-factor (EGF)-like domains (Kuroda *et al.*, 1999; see also fig. 4-6). EGF-like domains are composed of 40-50 amino acid residues, including six conserved cysteine residues. It has been revealed that these domains participate in the Ca^{2+}-dependent protein-protein interactions. The protein binds to and is phosphorylated by PKC-β1 via the EGF-like domains, suggesting that NELL1 represents a novel class of cell-signaling ligand molecules critical for growth and development. Furthermore, the NELL proteins possess heparin-binding activity (Kuroda *et al.*, 1999), which implies an interaction of the NELL1 protein with the heparan sulfate proteoglycan on the cell surface.

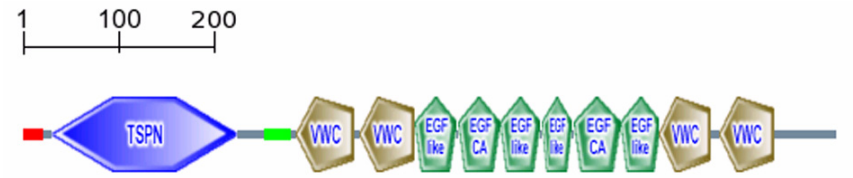

Fig. 4-6 NELL1 protein domain structure. Highlighted in red is the hydrophobic N-terminal secretion signal sequence and in green the coiled-coil region. TSPN: Thrombospondin N-terminal-like domain; VWC: von Willebrand factor type C domain; EGF-like: epidermall growth-factor like domain; EGF CA: Calcium-binding EGF-like domain. *SMART accession number Q92832.*

Has *NELL1* previously be considered a neuronal-restricted gene, Luce and colleagues (1999) found that NELL1 *mRNA* is transcribed during a narrow window of pre-B cell development. Sequence identities between NELL1 and its homologue NELL2 are rather low (40% on protein level), suggesting that these proteins have diverged in an early stage of molecular evolution (Kuroda *et al.*, 1999). NELL1 shares many protein motifs with thrombospondin-1 (TSP-1), which could mean they also have similar molecular functions. TSP1 interacts with various receptors, cytokines, proteases, and extracellular molecules, and exhibits versatile cell-specific effects on adhesion, migration, and proliferation (Bornstein *et al.*, 1995). Mouse models have significantly contributed to the understanding of the NELL1 function, for example,

over-expression of *NELL1* leads to craniosynostis (CS) in transgenic mice (Zhang *et al.*, 2002). CS is a significant medical condition that is caused by the premature closure of the cranial suture, which can severely constrain the growth of the brain. This premature closure is caused by an increase in osteoblast differentiation, apoptosis, and mineralization (Zhang *et al.*, 2003). NELL1's importance for osteoblast differentiation was recently supported by *NELL1*-deficient mice, which manifest skeletal defects in the vertebral column and ribcage, and an altered cranial morphology (Desai *et al.*, 2006). Interestingly, these knock-out mice showed a significantly reduced expression of extracellular matrix (ECM) proteins critical for chondrogenesis and osteogenesis, e.g. tenascin (Tnxb), collagen 5 alpha 3 subunit (Col5a3), and thrombospondin 3 (Thbs3). Furthermore, the prostaglandin E receptor 4 (*PTGER4*), which is an interesting candidate of the 5p13.1 region (see 4.3.3 on page 141), is 1.23-fold downregulated in *NELL1*-deficient mice. Desai and colleagues (2006) reported further that *NELL1*-knockout mice develop to late gestation (E19 days) but do not survive the physical trauma of birth. Mutant neonates were alive after recovery by caesarean section, but quickly succumbed, because they were unable to breathe. However, heterozygote mice survive to adulthood and breed normally, with no readily visible phenotypic differences when compared with wild-type mice.

Given the variety of known *NELL1* functions, but missing a connection to the pathogenesis of IBD, it is difficult to speculate about a potential role of this susceptibility gene. Nevertheless, it is intriguing that *NELL1* is involved in bone formation (Cowan *et al.*, 2006) and that a higher than average prevalence of osteoporosis exists among IBD patients (Abitbol *et al.*, 1995; Roux *et al.*, 1995; Compston *et al.*, 1987). Furthermore, the protein tenascin, mentioned before to be significantly downregulated in *NELL1*-deficient mice, is thought to be involved in stricture formation in CD patients (Geboes *et al.*, 2001).

Before functional or clinical studies are carried out to further clarify the role of *NELL1* in IBD, more efforts are necessary to identify (or verify) the underlying causative mutation(s). Insignificant TDT results for the non-synonymous mutations could be due to transmission ratio distortion (TRD; see 4.1.5 on page 132). Negative selective pressure that might result in TRD is possible, since NELL1 is a crucial protein for the organism as shown by *NELL1*-knockout mice that die during birth.

4.3.3 Regulatory elements of *PTGER4* might be disturbed in CD patients

A broad CD-specific association signal of more than 160 kb was identified on chromosome 5p13.1. Unfortunately, none of the 28 typed SNPs indicated an association of the genes nearby. The signal clearly localizes 125 kb upstream of the next gene *PTGER4*, the latter coding for the prostaglandin E receptor 4 (suptype EP4). Enhancer or silencer elements for *PTGER4* could reside in the peak region, as it is known from some examples that regulatory elements can be very distant from the gene of which transcription is influenced (Hardison *et al.*, 2000).

The ligand of the EP4 seven-transmembrane domain, G-protein-coupled receptor is known to be prostaglandin E_2 (PGE_2). In the prostaglandin (PG) pathway, the PG endoperoxide intermediate PGH_2 is synthesized from arachidonic acid by the enzyme cyclooxygenase (COX; also known as PTGS2). Several PGE synthases exist that convert PGH_2 into PGE_2, which is known to have a relatively short half-life and to act over short distances in an autocrine or paracrine manner (Hull *et al.*, 2005). EP4 receptor signaling generates increased intracellular cyclic AMP (cAMP) levels via coupling to G_s proteins. Interestingly, the EP4 receptor is highly expressed in the intestinal epithelium (*see GeneCards® entry*). Potential physiologic roles for EP receptors in vivo can also be hypothesized from in vitro data. For example, the EP4 receptor mediates mucus production and protects against apoptosis in colonic and gastric epithelial cells (Sheng *et al.*, 1998; Belley *et al.*, 1999 ; Hoshino *et al.*, 2003). Therefore, impaired function of the EP4 receptor could promote mucosal injury in vivo (Hatazawa *et al.*, 2006) and increase susceptibility to IBD. Not surprisingly, drugs for IBD treatment are evaluated that alter the PG pathway, e.g. inhibitors of the enzyme COX-2 (McCartney *et al.*, 1999). Prostaglandins are thought to be essential in the process of wound healing in the gastrointestinal tract (Wallace *et al.*, 2001) and use of non-steroidal anti-inflammatory drugs (NSAIDs), which inhibit both the transcription and activity of COX-2, exacerbates the symptoms in UC (Eberhart *et al.*, 1995) and may even activate quiescent IBD (Kaufmann *et al.*, 1987). Thus, the expression of COX-2 in the inflamed intestine might be a protective response within the woundhealing process (Singer *et al.*, 1998). Two polymorphisms in the *COX-2* gene have been associated with IBD by Cox *et al.* (2005), but this finding needs support from further replication studies.

Just recently, Kurz and colleagues (2006) fine mapped a previously known linkage region on chromosome 5p13 that was associated with asthma. Strikingly, besides five other candidate genes, they identified variation in *PTGER4* as a susceptibility factor for asthma. Since prostaglandins have important roles in inflammation and immediate hypersensitivity, the EP4 receptor can be involved in the pathogenesis of asthma and IBD. It has long been known that PGE_2 is produced abundantly in skin upon exposure to antigens (Ruzicka *et al.*, 1982; Eber-

hard *et al.*, 2002) and it is clear since 2003 (Kabashima *et al.*) that PGE_2-EP4 signaling initiates this immune response by stimulating Langerhans cell mobilization, migration, and maturation.

In summary, *PTGER4* seems to be the perfect candidate gene in the 5p13 susceptibility region and *PTGER4* regulatory elements might be altered in CD patients. However, it cannot be ruled out that a potentially existing regulatory element could also influence the expression of a more distant candidate gene, such as *CARD8*. As long as this hypothesis has not been proven, it is also likely that a yet unknown gene exists in the peak region, but evidence in form of spliced ESTs is scarce.

4.4 Concluding remarks and future studies

The genetic findings of the novel IBD susceptibility loci *ATG16L1*, *NELL1*, and 5p13.1 in this study require further investigations, especially on the functional level. Since the non-synonymous SNP rs2241880, which is strongly associated with CD, is the most consistent and unambigous finding of the two screenings, next experiments should focus on the altered function of the *ATG16L1* protein.

The first reasonable experiment, which is already ongoing, is to look for a different cellular localization of the mutant protein compared to the wildtype protein. This can be achieved by examining the two different protein variants, which must be GFP-labelled, using a fluorescence microscope. To this end, the point mutation must be introduced into a full-length sequence clone, then subcloned into a GFP-vector, and subsequently transfected into *ATG16L1* knock-down cells. The latter can be done by transfecting cells with RNAi, which is specific for part of the ATG16L1 untranslated region. By not targeting the actual coding region of *ATG16L1*, transcribed mRNAs from the GFP constructs won't be degraded and only cellular transcript levels are reduced. First experiments with *ATG16L1* knock-down cell lines, which were subsequently not transfected with any other *ATG16L1* constructs, have shown that the viability of the cells is significantly reduced as autophagosomes keep accumulating.

Another good experiment would be to study the impact of the found *ATG16L1* and *NELL1* mutations in a mouse model. Knock-out mice for both genes (see 4.3 on page 135) exist, but the loss-of function phenotypes do not adequately reflect the effect, which an altered protein can have in the pathogenesis of IBD. Therefore, mice strains should be bred that are homozygous for either wildtype or mutant allele of the respective homologous SNP. Afterwards, the different strains could be exposed to mucosal-damaging reagents and subsequently compared for different IBD incident rates and severity of the disease.

Since a statistical interaction between *ATG16L1* and *CARD15* risk alleles has been detected, it would be interesting to further examine this epistatic phenomenon in cell lines or mouse models. For example, a variety of experiments could be carried out with four cell lines, each having a different combination of *CARD15*/*ATG16L1* risk or non-risk genotypes:

	***CARD15* risk background**	***CARD15* wildtype**
***ATG16L1* risk background**	cell line A	cell line B
***ATG16L1* wildtype**	cell line C	cell line D

Table 4-2 Four cell lines with different combinations of genetic risk backgrounds.

So far, associations with IBD have been shown to exist in the German and the British population, but more populations with different ethnical backgrounds should be tested. It is planned to test the associations in a Korean and a Norwegian IBD sample. Other groups will certainly test their sample collections once the data is publicly available. As it is possible that the IBD-associated mutations predispose to other chronic inflammatory diseases as well, e.g. sarcoidosis, psoriasis, or tuberculosis, additional genotyping experiments should include samples of these disorders.

5 Summary

Two independent and hypothesis-free genome-wide association studies were carried out to find novel susceptibility genes for Crohn's disease (CD), which is a chronic inflammatory disorder of the bowel. In a first sequence-based and "direct" scan, approximately 20,000 non-synonymous SNPs, which result in a change in protein sequence, were typed. The second "indirect" map-based scan comprised about 100,000 evenly distributed SNPs.

SNP screen 1

A subset of 19,779 putatively functional SNPs was selected from over 60,000 candidate nsSNPs compiled from public and private sources. Genotyping in 735 patients with CD and 368 controls by SNPlex™ technology identified 7,159 SNPs as being informative in this population, and the top 72 hits were genotyped in an independent German sample of 380 trios, 878 cases, and 1.032 controls. Besides the known *CARD15* variant "SNP 12", the hierarchic analyses identified one significantly disease-associated non-synonymous SNP ($p = 8.4 \times 10^{-6}$ single point allelic case-control; $p = 2.7 \times 10^{-5}$ TDT) in the *ATG16L1* gene on chromosome 2q37, namely rs2241880. The risk to develop CD was calculated to be 1.77-fold (95% CI: 1.43 – 2.18) higher for homozygous carriers of the "GG" genotype compared to homozygotes for the rarer allele. The association was replicated in another British CD sample, but not in German UC patients, thus classifying *ATG16L1* as a CD-specific susceptibility gene. Full mutation detection failed to detect any other cSNP that explained the association. The novel susceptibility variant rs2241880 interacts genetically with known variants in *CARD15* (Breslow Day test, $p = 0.039$ evaluating homozygous and compound heterozygous carriers of *CARD15* risk alleles). The predicted function suggests that the variant leads to an impairment of the killing of intracellular bacteria and the identified novel disease gene adds to the understanding of CD as an innate immune disorder. Further studies are required to establish the effect of this variant within the gastrointestinal tract and the role it plays in IBD behaviour and response to treatment.

SNP screen 2

116,161 SNPs were successfully genotyped in 399 controls and 393 CD cases using the Affymetrix 100k GeneChip®. Among the top 150 significant lead SNPs that were typed in a larger independent patient panel, SNPs in the *NELL1* gene and a gene-free region upstream of *PTGER4* on chromosome 5p13 were replicated. The performed mutation detection for the *NELL1* gene led to the identification of four non-synonymous polymorphisms. CD and UC patients heterozygous for the rs8176785 variant were found to have a 1.23-fold (95% CI: 1.03 – 1.46) or 1.35-fold (95% CI: 1.12 – 1.64) increased risk, respectively, compared to homozygous carriers of the rare allele. In summary, the LD-based screening led to the identification of a genuine IBD susceptibility gene and a CD-predisposing region of yet unknown function.

6 Zusammenfassung

Zwei unabhängige, genomweite Assoziationsstudien wurden durchgeführt, um prädisponierende Genvarianten für Morbus Crohn (MC), eine chronisch entzündliche Darmerkrankung (CED), zu finden. Zum einem wurde ein „direkter“ und sequenzbasierter Scan mit circa 20.000 nicht-synonymen SNPs durchgeführt, zum anderen ein „indirekter“ Scan mit über 100.000 LD-basierten SNPs.

SNP screen 1

Von über 60.000 nicht-synonymen SNPs (nsSNP) aus öffentlichen und privaten Datenbanken wurden 19.779 ausgewählt. Die Genotypisierung von 735 MC-Patienten und 368 gesunden Kontrollpersonen führte zur Identifikation von 7.159 in der Population informativen SNPs. Die 72 am stärksten assoziierten Polymorphismen wurden in einer unabhängigen deutschen Studiengruppe von 380 „Mutter-Vater-Kind Trios“, 878 MC Einzelpatienten und 1.032 unverwandten Kontrollen nachverfolgt. Neben der bekannten *CARD15* Variante „SNP 12“, wurde der hochsignifikante ($p = 8{,}4 \times 10^{-6}$ im allel-basierten χ^2 Test; $p = 2{,}7 \times 10^{-5}$ im TDT) nsSNP rs241880 im *ATG16L1* Gen auf Chromosom 2q37 identifiziert. Das Risiko für homozygote Träger des Genotyps „GG“ an MC zu erkranken, ist dabei 1,77fach (95% KI: 1,43 – 2,18) höher als für Homozygote des seltenen Allels „A“. Das Assoziationssignal ließ sich in einer Britischen MC-Probengruppe bestätigen, jedoch nicht in einer Gruppe aus deutschen Patienten mit Ulcerativer Colitis (UC). Dieses Ergebnis bestätigte eine eindeutige Assoziation der *ATG16L1*-Variante mit MC und nicht mit CED. Eine Resequenzierung der *ATG16L1*-codierenden Regionen für 47 MC Proben führte zu keiner weiteren Identifikation einer zusätzlichen Mutation. Die Suszeptibilitätsvariante rs2241880 zeigte auf genetischer Ebene eine statistisch signifikante Interaktion mit den bekannten *CARD15*-Mutationen (Breslow Day Test, $p = 0{,}039$ für untersuchte Homozygote oder kombinierte heterozygote Träger der *CARD15*-Risikoallele). Die Schlüsselrolle von *ATG16L1* bei der Autophagosomenbildung suggeriert eine Beeinträchtigung der Pathogenabwehr durch die Mutation und liefert weitere Argumente für die Sichtweise von MC als eine Barrierekrankheit, die auf eine Störung des angeborenen Immunsystems zurückgeht. Weitere Experimente sind notwendig, um den Effekt der Mutation auf den Darm und die Rolle bei der Krankheitsentstehung zu untersuchen.

SNP screen 2

Bei diesem Experiment wurden 116.161 SNPs in 399 gesunden Kontrollen und 393 MC-Proben mit Hilfe einer chip-basierten Methode typisiert. Unter den 150 am höchsten signifikanten SNPs, die in einer größeren, unabhängigen Studienprobe typisiert wurden, ließen sich Varianten im *NELL1*-Gen und in einer genfreien Region auf Chromosom 5p13.1 replizieren. Die durchgeführte Mutationsdetektion für *NELL1* führte zur Identifikation von vier nicht-synonymem Polymorphismen. Für heterozygote MC- bzw. UC-Patienten der Variante rs8176785 konnte ein 1,23fach (95% KI: 1,03 – 1,46), bzw. 1,35fach (95% KI: 1,12 – 1,64) erhöhtes Risiko berechnet werden, verglichen mit homozygoten Trägern für das seltenere Allel. Zusammenfassend kann gesagt werden, dass der auf Kopplungsungleichgewicht basierende Scan zur Identifikation eines echten CED-Suszeptibilitätsgenes und einer mit MC assoziierten Region führte. Auch für diese Loci sind vor allem funktionelle Studien essentiell, um die genaue Rolle bei der Pathogenese zu bestimmen.

7 References

7.1 Articles

Abecasis, G. R., Cherny, S. S. & Cardon, L. R. The impact of genotyping error on family-based analysis of quantitative traits. *Eur J Hum Genet* 9, 130-4 (2001).

Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 68, 191-197 (2001).

Abitbol, V. *et al.* Metabolic bone assessment in patients with inflammatory bowel disease. *Gastroenterology* 108, 417-22 (1995).

Adams, M. Applied genomics: exploring functional variation and gene expression. *Am J Hum Genet* 71, 203 (2002).

Adler, D. J. & Korelitz, B. I. The therapeutic efficacy of 6-mercaptopurine in refractory ulcerative colitis. *Am J Gastroenterol* 85, 717-22 (1990).

Afonina, I. *et al.* Efficient priming of PCR with short oligonucleotides conjugated to a minor groove binder. *Nucleic Acids Res* 25, 2657-60 (1997).

Agresti, A. & Coull, B. A. Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics* 52, 1103-11 (1996).

Ahmad, T. *et al.* The molecular classification of the clinical manifestations of Crohn's disease. *Gastroenterology* 122, 854-66 (2002).

Ahmad, T. *et al.* The contribution of human leucocyte antigen complex genes to disease phenotype in ulcerative colitis. *Tissue Antigens* 62, 527-35 (2003).

Ahmad, T. *et al.* High resolution MIC genotyping: design and application to the investigation of inflammatory bowel disease susceptibility. *Tissue Antigens* 60, 164-79 (2002).

Ahn, S. J., Costa, J. & Emanuel, J. R. PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. *Nucleic Acids Res* 24, 2623-5 (1996).

Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* 439, 851-5 (2006).

Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 12, 1805-14 (2002).

Akira, S. & Takeda, K. Toll-like receptor signalling. *Nat Rev Immunol* 4, 499-511 (2004).

Akolkar, P. N. *et al.* The IBD1 locus for susceptibility to Crohn's disease has a greater impact in Ashkenazi Jews with early onset disease. *Am J Gastroenterol* 96, 1127-32 (2001).

Albrecht, M., Tosatto, S. C., Lengauer, T. & Valle, G. Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng* 16, 459-62 (2003).

Andreeva, A. *et al.* SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32, D226-9 (2004).

Andus, T. & Gross, V. Etiology and pathophysiology of inflammatory bowel disease--environmental factors. *Hepatogastroenterology* 47, 29-43 (2000).

Angeli, A. *et al.* Modulation by cytokines of glucocorticoid action. *Ann N Y Acad Sci* 876, 210-20 (1999).

Annese, V. *et al.* Genetic analysis in Italian families with inflammatory bowel disease supports linkage to the IBD1 locus--a GISC study. *Eur J Hum Genet* 7, 567-73 (1999).

Appleton, B. A., Wu, P. & Wiesmann, C. The crystal structure of murine coronin-1: a regulator of actin cytoskeletal dynamics in lymphocytes. *Structure* 14, 87-96 (2006).

Ardizzone, S., Molteni, P., Imbesi, V., Bollani, S. & Bianchi Porro, G. Azathioprine in steroid-resistant and steroid-dependent ulcerative colitis. *J Clin Gastroenterol* 25, 330-3 (1997).

Ardlie, K. *et al.* Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am J Hum Genet* 69, 582-9 (2001).

Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 3, 299-309 (2002).

Arnott, I. D. *et al.* NOD2/CARD15, TLR4 and CD14 mutations in Scottish and Irish Crohn's disease patients: evidence for genetic heterogeneity within Europe? *Genes Immun* 5, 417-25 (2004).

Augustin, R. *et al.* Dickkopf related genes are components of the positional value gradient in Hydra. *Dev Biol* (2006).

Bacanu, S. A., Devlin, B. & Roeder, K. The power of genomic control. *Am J Hum Genet* 66, 1933-44 (2000).

Bach, J. F. The effect of infections on susceptibility to autoimmune and allergic diseases. *N Engl J Med* 347, 911-20 (2002).

Baehrecke, E. H. Autophagy: dual roles in life and death? *Nat Rev Mol Cell Biol* 6, 505-10 (2005).

Bamias, G. *et al.* Expression, localization, and functional activity of TL1A, a novel Th1-polarizing cytokine in inflammatory bowel disease. *J Immunol* 171, 4868-74 (2003).

Barany, F. Genetic disease detection and DNA amplification using cloned thermostable ligase. *Proc Natl Acad Sci U S A* 88, 189-93 (1991).

Barmada, M. M. *et al.* A genome scan in 260 inflammatory bowel disease-affected relative pairs. *Inflamm Bowel Dis* 10, 513-20 (2004).

Barnich, N. *et al.* GRIM-19 interacts with nucleotide oligomerization domain 2 and serves as downstream effector of anti-bacterial function in intestinal epithelial cells. *J Biol Chem* 280, 19021-6 (2005).

Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat Genet* 38, 659-62 (2006).

Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263-5 (2005).

Bearden, J., Jr. Isolation of nucleolar proteins by ethanol precipitation of nucleic acids. *Biochim Biophys Acta* 361, 109-13 (1974).

Becker, K. G. *et al.* Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. *Proc Natl Acad Sci U S A* 95, 9979-84 (1998).

Belley, A. & Chadee, K. Prostaglandin E(2) stimulates rat and human colonic mucin exocytosis via the EP(4) receptor. *Gastroenterology* 117, 1352-62 (1999).

Bellman, R. Adaptive Control Processes. (1961).

Bennett, S. T., Barnes, C., Cox, A., Davies, L. & Brown, C. Toward the 1,000 dollars human genome. *Pharmacogenomics* 6, 373-82 (2005).

Bergen, A. W., Qi, Y., Haque, K. A., Welch, R. A. & Chanock, S. J. Effects of DNA mass on multiple displacement whole genome amplification and genotyping performance. *BMC Biotechnol* 5, 24 (2005).

Berthold, V. & Geider, K. Interaction of DNA with DNA-binding proteins. The characterization of protein HD from Escherichia coli and its nucleic acid complexes. *Eur J Biochem* 71, 443-9 (1976).

Beutler, B. TNF, immunity and inflammatory disease: lessons of the past decade. *J Investig Med* 43, 227-35 (1995).

Biagi, F. *et al.* Video capsule endoscopy and histology for small-bowel mucosa evaluation: a comparison performed by blinded observers. *Clin Gastroenterol Hepatol* 4, 998-1003 (2006).

Birney, E. *et al.* Ensembl 2006. *Nucleic Acids Res* 34, D556-61 (2006).

Birrenbach, T. & Bocker, U. Inflammatory bowel disease and smoking: a review of epidemiology, pathophysiology, and therapeutic implications. *Inflamm Bowel Dis* 10, 848-59 (2004).

Bland, J. M. & Altman, D. G. Multiple significance tests: the Bonferroni method. *Bmj* 310, 170 (1995).

Blomqvist, P., Feltelius, N., Lofberg, R. & Ekbom, A. A 10-year survey of inflammatory bowel diseases-drug therapy, costs and adverse reactions. *Aliment Pharmacol Ther* 15, 475-81 (2001).

Blumberg, R. S., Saubermann, L. J. & Strober, W. Animal models of mucosal inflammation and their relation to human inflammatory bowel disease. *Curr Opin Immunol* 11, 648-56 (1999).

Bocci, V. The neglected organ: bacterial flora has a crucial immunostimulatory role. *Perspect Biol Med* 35, 251-60 (1992).

Bonin, A. *et al.* How to track and assess genotyping errors in population genetics studies. *Mol Ecol* 13, 3261-73 (2004).

Bornstein, P. Diversity of function is inherent in matricellular proteins: an appraisal of thrombospondin 1. *J Cell Biol* 130, 503-6 (1995).

Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 33 Suppl, 228-37 (2003).

Bouma, G. *et al.* HLA-DRB1*03, but not the TNFA -308 promoter gene polymorphism, confers protection against fistulising Crohn's disease. *Immunogenetics* 47, 451-5 (1998).

Bouma, G. *et al.* Distribution of four polymorphisms in the tumour necrosis factor (TNF) genes in patients with inflammatory bowel disease (IBD). *Clin Exp Immunol* 103, 391-6 (1996).

Braegger, C. P., Nicholls, S., Murch, S. H., Stephens, S. & MacDonald, T. T. Tumour necrosis factor alpha in stool as a marker of intestinal inflammation. *Lancet* 339, 89-91 (1992).

Brand, S. *et al.* The role of Toll-like receptor 4 Asp299Gly and Thr399Ile polymorphisms and CARD15/NOD2 mutations in the susceptibility and phenotype of Crohn's disease. *Inflamm Bowel Dis* 11, 645-52 (2005).

Brandwein, S. L. *et al.* Spontaneously colitic C3H/HeJBir mice demonstrate selective antibody reactivity to antigens of the enteric bacterial flora. *J Immunol* 159, 44-52 (1997).

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).

Brant, S. R. *et al.* Linkage heterogeneity for the IBD1 locus in Crohn's disease pedigrees by disease onset and severity. *Gastroenterology* 119, 1483-90 (2000).

Breese, E. J. *et al.* Tumor necrosis factor alpha-producing cells in the intestinal mucosa of children with inflammatory bowel disease. *Gastroenterology* 106, 1455-66 (1994).

Briese, V., Muller, H. & Berkholz, A. [Pre-conception counseling and pregnancy in chronic inflammatory bowel diseases--Crohn disease and ulcerative colitis]. *Zentralbl Gynakol* 115, 1-6 (1993).

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J. & Rolf, B. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* 62, 1408-15 (1998).

Bujnicki, J. M., Elofsson, A., Fischer, D. & Rychlewski, L. Structure prediction meta server. *Bioinformatics* 17, 750-1 (2001).

Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* 437, 1153-7 (2005).

Byrne, F. R. & Viney, J. L. Mouse models of inflammatory bowel disease. *Curr Opin Drug Discov Devel* 9, 207-17 (2006).

Campieri, M. & Gionchetti, P. Probiotics in inflammatory bowel disease: new insight to pathogenesis or a possible therapeutic alternative? *Gastroenterology* 116, 1246-9 (1999).

Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* 361, 598-604 (2003).

Cargill, M. & Daley, G. Q. Mining for SNPs: putting the common variants--common disease hypothesis to the test. *Pharmacogenomics* 1, 27-37 (2000).

Carpenter, G. & Cohen, S. Epidermal growth factor. *J Biol Chem* 265, 7709-12 (1990).

Carrasquillo, M. M. *et al.* Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat Genet* 32, 237-44 (2002).

Cavanaugh, J. International collaboration provides convincing linkage replication in complex disease through analysis of a large pooled data set: Crohn disease and chromosome 16. *Am J Hum Genet* 68, 1165-71 (2001).

Cavanaugh, J. A. *et al.* Analysis of Australian Crohn's disease pedigrees refines the localization for susceptibility to inflammatory bowel disease on chromosome 16. *Ann Hum Genet* 62, 291-8 (1998).

Cave, N. J. Chronic inflammatory disorders of the gastrointestinal tract of companion animals. *N Z Vet J* 51, 262-74 (2003).

Cerna, D. & Wilson, D. K. The structure of Sif2p, a WD repeat protein functioning in the SET3 corepressor complex. *J Mol Biol* 351, 923-35 (2005).

Chalifoux, L. V., Brieland, J. K. & King, N. W. Evolution and natural history of colonic disease in cotton-top tamarins (Saguinus oedipus). *Dig Dis Sci* 30, 54S-58S (1985).

Chamaillard, M. *et al.* An essential role for NOD1 in host recognition of bacterial peptidoglycan containing diaminopimelic acid. *Nat Immunol* 4, 702-7 (2003).

Chen, F. C., Vallender, E. J., Wang, H., Tzeng, C. S. & Li, W. H. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J Hered* 92, 481-9 (2001).

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).

Cho, J. H. *et al.* Linkage and linkage disequilibrium in chromosome band 1p36 in American Chaldeans with inflammatory bowel disease. *Hum Mol Genet* 9, 1425-32 (2000).

Cho, R. J. & Campbell, M. J. Transcription, genomes, function. *Trends Genet* 16, 409-15 (2000).

Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* 20, 426-7 (2004).

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

Codogno, P. & Meijer, A. J. Autophagy and signaling: their role in cell survival and cell death. *Cell Death Differ* 12 Suppl 2, 1509-18 (2005).

Colhoun, H. M., McKeigue, P. M. & Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet* 361, 865-72 (2003).

Colombel, J. F. *et al.* Clinical characteristics of Crohn's disease in 72 families. *Gastroenterology* 111, 604-7 (1996).

Compston, J. E. *et al.* Osteoporosis in patients with inflammatory bowel disease. *Gut* 28, 410-5 (1987).

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 38, 75-81 (2006).

Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 11, 2463-8 (2002).

Cordell, H. J., Barratt, B. J. & Clayton, D. G. Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol* 26, 167-85 (2004).

Cordell, H. J. & Clayton, D. G. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet* 70, 124-41 (2002).

Costello, C. M. *et al.* Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays. *PLoS Med* 2, e199 (2005).

Cowan, C. M. *et al.* Nell-1 induced bone formation within the distracted intermaxillary suture. *Bone* 38, 48-58 (2006).

Cox, D. G. *et al.* Haplotype of prostaglandin synthase 2/cyclooxygenase 2 is involved in the susceptibility to inflammatory bowel disease. *World J Gastroenterol* 11, 6003-8 (2005).

Crohn, B. B., Ginzburg, L. & Oppenheimer, G. D. Landmark article Oct 15, 1932. Regional ileitis. A pathological and clinical entity. By Burril B. Crohn, Leon Ginzburg, and Gordon D. Oppenheimer. *Jama* 251, 73-9 (1984).

Croucher, P. J. *et al.* Haplotype structure and association to Crohn's disease of CARD15 mutations in two ethnically divergent populations. *Eur J Hum Genet* 11, 6-16 (2003).

Cuervo, A. M. Autophagy: in sickness and in health. *Trends Cell Biol* 14, 70-7 (2004).

Cullen, M., Perfetto, S. P., Klitz, W., Nelson, G. & Carrington, M. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am J Hum Genet* 71, 759-76 (2002).

Cullen, S. & Chapman, R. Primary sclerosing cholangitis. *Autoimmun Rev* 2, 305-12 (2003).

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).

Curtis, D. & Sham, P. C. A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet* 56, 811-2 (1995).

Daig, R. *et al.* Human intestinal epithelial cells secrete interleukin-1 receptor antagonist and interleukin-8 but not interleukin-1 or interleukin-6. *Gut* 46, 350-8 (2000).

Daly, M. J. *et al.* Association of DLG5 R30Q variant with inflammatory bowel disease. *Eur J Hum Genet* 13, 835-9 (2005).

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nat Genet* 29, 229-32 (2001).

Dausset, J. *et al.* Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* 6, 575-7 (1990).

de Bakker, P. I. *et al.* Efficiency and power in genetic association studies. *Nat Genet* 37, 1217-23 (2005).

De La Vega, F. M. *et al.* New generation pharmacogenomic tools: a SNP linkage disequilibrium Map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies. *Biotechniques* Suppl, 48-50, 52, 54 (2002).

De La Vega, F. M. *et al.* The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res* 15, 454-62 (2005).

De la Vega, F. M., Lazaruk, K. D., Rhodes, M. D. & Wenz, M. H. Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. *Mutat Res* 573, 111-35 (2005).

Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* 99, 5261-6 (2002).

Deretic, V. Autophagy in innate and adaptive immunity. *Trends Immunol* 26, 523-8 (2005).

Desai, J. *et al.* Nell1-deficient mice have reduced expression of extracellular matrix proteins causing cranial and vertebral defects. *Hum Mol Genet* 15, 1329-41 (2006).

Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* 55, 997-1004 (1999).

Devlin, B., Roeder, K. & Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 60, 155-66 (2001).

D'Haens, G. *et al.* Endoscopic and histological healing with infliximab anti-tumor necrosis factor antibodies in Crohn's disease: A European multicenter trial. *Gastroenterology* 116, 1029-34 (1999).

D'Haens, G. R. Infliximab (Remicade), a new biological treatment for Crohn's disease. *Ital J Gastroenterol Hepatol* 31, 519-20 (1999).

Di, X. *et al.* Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* 21, 1958-63 (2005).

Dieckgraefe, B. K., Stenson, W. F., Korzenik, J. R., Swanson, P. E. & Harrington, C. A. Analysis of mucosal gene expression in inflammatory bowel disease by parallel oligonucleotide arrays. *Physiol Genomics* 4, 1-11 (2000).

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

Din, S. *et al.* Proteomic Profiling Identifies Corticosteroid Resistant Patients in Severe Ulcerative Colitis. *Gastroenterology* 128, A310 (2005).

Don, R. H., Cox, P. T., Wainwright, B. J., Baker, K. & Mattick, J. S. 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res* 19, 4008 (1991).

Douglas, J. A., Skol, A. D. & Boehnke, M. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am J Hum Genet* 70, 487-95 (2002).

Dubinsky, M. C. *et al.* Pharmacogenomics and metabolite measurement for 6-mercaptopurine therapy in inflammatory bowel disease. *Gastroenterology* 118, 705-13 (2000).

Dubinsky, M. C. *et al.* Serum immune responses predict rapid disease progression among children with Crohn's disease: immune responses predict disease progression. *Am J Gastroenterol* 101, 360-7 (2006).

Dubinsky, M. C., Taylor, K., Targan, S. R. & Rotter, J. I. Immunogenetic phenotypes in inflammatory bowel disease. *World J Gastroenterol* 12, 3645-50 (2006).

Dudbridge, F. Pedigree disequilibrium tests for multilocus haplotypes. *Genet Epidemiol* 25, 115-21 (2003).

Dudbridge, F., Koeleman, B. P., Todd, J. A. & Clayton, D. G. Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *Am J Hum Genet* 66, 2009-12 (2000).

Duerr, R. H. The genetics of inflammatory bowel disease. *Gastroenterol Clin North Am* 31, 63-76 (2002).

Duerr, R. H. Update on the genetics of inflammatory bowel disease. *J Clin Gastroenterol* 37, 358-67 (2003).

Duerr, R. H. *et al.* Evidence for an inflammatory bowel disease locus on chromosome 3p26: linkage, transmission/disequilibrium and partitioning of linkage. *Hum Mol Genet* 11, 2599-606 (2002).

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).

Dupont, W. D. & Plummer, W. D. PS power and sample size program available for free on the Internet. *Controlled Clin Trials* 18 (1997).

Eberhard, J., Jepsen, S., Pohl, L., Albers, H. K. & Acil, Y. Bacterial challenge stimulates formation of arachidonic acid metabolites by human keratinocytes and neutrophils in vitro. *Clin Diagn Lab Immunol* 9, 132-7 (2002).

Eberhart, C. E. & Dubois, R. N. Eicosanoids and the gastrointestinal tract. *Gastroenterology* 109, 285-301 (1995).

Economou, M., Trikalinos, T. A., Loizou, K. T., Tsianos, E. V. & Ioannidis, J. P. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. *Am J Gastroenterol* 99, 2393-404 (2004).

Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792-7 (2004).

Edmonds, D. K., Lindsay, K. S., Miller, J. F., Williamson, E. & Wood, P. J. Early embryonic mortality in women. *Fertil Steril* 38, 447-53 (1982).

Eskelinen, E.-L. Maturation of autophagic vacuoles in mammalian cells. *Autophagy* 1, 1-10 (2005).

Ewen, K. R. *et al.* Identification and analysis of error types in high-throughput genotyping. *Am J Hum Genet* 67, 727-36 (2000).

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

Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12, 921-7 (1995).

Falk, C. T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 51 (Pt 3), 227-33 (1987).

Fan, J. B. *et al.* Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res* 10, 853-60 (2000).

Farrell, R. J. *et al.* Increased incidence of non-Hodgkin's lymphoma in inflammatory bowel disease patients on immunosuppressive therapy but overall risk is low. *Gut* 47, 514-9 (2000).

Farrell, R. J. *et al.* High multidrug resistance (P-glycoprotein 170) expression in inflammatory bowel disease patients who fail medical therapy. *Gastroenterology* 118, 279-88 (2000).

Farrell, R. J. & Peppercorn, M. A. Ulcerative colitis. *Lancet* 359, 331-40 (2002).

Fearnhead, N. S. *et al.* Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc Natl Acad Sci U S A* 101, 15992-7 (2004).

Feinstein, R. E. & Olsson, E. Chronic gastroenterocolitis in nine cats. *J Vet Diagn Invest* 4, 293-8 (1992).

Feldmann, M. & Maini, R. N. Anti-TNF alpha therapy of rheumatoid arthritis: what have we learned? *Annu Rev Immunol* 19, 163-96 (2001).

Fellermann, K. *et al.* A chromosome 8 gene-cluster polymorphism with low human Beta-defensin 2 gene copy number predisposes to crohn disease of the colon. *Am J Hum Genet* 79, 439-48 (2006).

Ferguson, A. Ulcerative colitis and Crohn's disease. *Bmj* 309, 355-6 (1994).

Fielding, J. F. The relative risk of inflammatory bowel disease among parents and siblings of Crohn's disease patients. *J Clin Gastroenterol* 8, 655-7 (1986).

Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res* 34, D247-51 (2006).

Fisher, S. A. *et al.* Sex stratification of an inflammatory bowel disease genome search shows male-specific linkage to the HLA region of chromosome 6. *Eur J Hum Genet* 10, 259-65 (2002).

Fisher, S. A. *et al.* Direct or indirect association in a complex disease: the role of SLC22A4 and SLC22A5 functional variants in Crohn disease. *Hum Mutat* 27, 778-5 (2006).

Fort, M. M. *et al.* A synthetic TLR4 antagonist has anti-inflammatory effects in two murine models of inflammatory bowel disease. *J Immunol* 174, 6416-23 (2005).

Fowler, E. V. *et al.* TNFalpha and IL10 SNPs act together to predict disease behaviour in Crohn's disease. *J Med Genet* 42, 523-8 (2005).

Franchimont, D. *et al.* Deficient host-bacteria interactions in inflammatory bowel disease? The toll-like receptor (TLR)-4 Asp299gly polymorphism is associated with Crohn's disease and ulcerative colitis. *Gut* 53, 987-92 (2004).

Franke, A. *et al.* No association between the functional CARD4 insertion/deletion polymorphism and inflammatory bowel diseases in the German population. *Gut* 55, 11 (2006).

Franke, A. *et al.* GENOMIZER: an integrated analysis system for genome-wide association data. *Hum Mutat* 27, 583-8 (2006).

Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat Genet* 36, 388-93 (2004).

Fremming, B. D., Vogel, F. S., Benson, R. E. & Young, R. J. A fatal case of amebiasis with liver abscesses and ulcerative colitis in a chimpanzee. *J Am Vet Med Assoc* 126, 406-7 (1955).

Friedrichs, F. *et al.* Evidence of transmission ratio distortion of DLG5 R30Q variant in general and implication of an association with Crohn disease in men. *Hum Genet* 119, 305-11 (2006).

Friedrichs, F. & Stoll, M. Role of discs large homolog 5. *World J Gastroenterol* 12, 3651-6 (2006).

Frisse, L. *et al.* Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet* 69, 831-43 (2001).

Galtier, N., Gouy, M. & Gautier, C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 12, 543-8 (1996).

Gaya, D. R., Russell, R. K., Nimmo, E. R. & Satsangi, J. New genes in inflammatory bowel disease: lessons for complex diseases? *Lancet* 367, 1271-84 (2006).

Gazouli, M., Mantzaris, G., Archimandritis, A. J., Nasioulas, G. & Anagnou, N. P. Single nucleotide polymorphisms of OCTN1, OCTN2, and DLG5 genes in Greek patients with Crohn's disease. *World J Gastroenterol* 11, 7525-30 (2005).

Gazouli, M. *et al.* Association between polymorphisms in the Toll-like receptor 4, CD14, and CARD15/NOD2 and inflammatory bowel disease in the Greek population. *World J Gastroenterol* 11, 681-5 (2005).

Gearry, R. B. *et al.* CARD15 allele frequency differences in New Zealand Maori: ancestry specific susceptibility to Crohn's disease in New Zealand? *Gut* 55, 580 (2006).

Geboes, K. *et al.* Tenascin and strictures in inflammatory bowel disease: an immunohistochemical study. *Int J Surg Pathol* 9, 281-6 (2001).

Gerwitz, A. T. *et al.* Common dominant-negative Tlr5 Polymorphisms Reduces Adaptive Immune Response to Flagellin and Provides Protection from Crohn's Disease. *Gastroenterology* 128; Suppl2:A55 (2005).

Gewirtz, A. T., Navas, T. A., Lyons, S., Godowski, P. J. & Madara, J. L. Cutting edge: bacterial flagellin activates basolaterally expressed TLR5 to induce epithelial proinflammatory gene expression. *J Immunol* 167, 1882-5 (2001).

Giallourakis, C. *et al.* IBD5 is a general risk factor for inflammatory bowel disease: replication of association with Crohn disease and identification of a novel association with ulcerative colitis. *Am J Hum Genet* 73, 205-11 (2003).

Ginalski, K. & Rychlewski, L. Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res* 31, 3291-2 (2003).

Girardin, S. E. *et al.* CARD4/Nod1 mediates NF-kappaB and JNK activation by invasive Shigella flexneri. *EMBO* Rep 2, 736-42 (2001).

Glatt, C. E. *et al.* Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. *Nat Genet* 27, 435-8 (2001).

Goldstein, D. B. & Weale, M. E. Population genomics: linkage disequilibrium holds the key. *Curr Biol* 11, R576-9 (2001).

Goppelt-Struebe, M. Molecular mechanisms involved in the regulation of prostaglandin biosynthesis by glucocorticoids. *Biochem Pharmacol* 53, 1389-95 (1997).

Gordon, D. *et al.* A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur J Hum Genet* 12, 752-61 (2004).

Gordon, D., Matise, T. C., Heath, S. C. & Ott, J. Power loss for multiallelic transmission/disequilibrium test when errors introduced: GAW11 simulated data. *Genet Epidemiol* 17 Suppl 1, S587-92 (1999).

Goring, H. H. & Terwilliger, J. D. Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. *Am J Hum Genet* 66, 1298-309 (2000).

Goulding, N. J., Euzger, H. S., Butt, S. K. & Perretti, M. Novel pathways for glucocorticoid effects on neutrophils in chronic inflammation. *Inflamm Res* 47 Suppl 3, S158-65 (1998).

Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* 185, 862-4 (1974).

Gue, M. *et al.* Stress-induced enhancement of colitis in rats: CRF and arginine vasopressin are not involved. *Am J Physiol* 272, G84-91 (1997).

Gutierrez, M. G. *et al.* Autophagy is a defense mechanism inhibiting BCG and Mycobacterium tuberculosis survival in infected macrophages. *Cell* 119, 753-66 (2004).

Hacia, J. G. *et al.* Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet* 22, 164-7 (1999).

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).

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 (2002).

Hampe, J. *et al.* Association of NOD2 (CARD 15) genotype with clinical course of Crohn's disease: a cohort study. *Lancet* 359, 1661-5 (2002).

Hampe, J. *et al.* The interferon-gamma gene as a positional and functional candidate gene for inflammatory bowel disease. *Int J Colorectal Dis* 13, 260-3 (1998).

Hampe, J. *et al.* Fine mapping of the chromosome 3p susceptibility locus in inflammatory bowel disease. *Gut* 48, 191-7 (2001).

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).

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

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

Hanada, T. & Ohsumi, Y. Structure-function relationship of Atg12, a ubiquitin-like modifier essential for autophay. *Autophagy* 1, 110-8 (2005).

Hanauer, S. B. & Dassopoulos, T. Evolving treatment strategies for inflammatory bowel disease. *Annu Rev Med* 52, 299-318 (2001).

Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* 16, 369-72 (2000).

Hatazawa, R., Ohno, R., Tanigami, M., Tanaka, A. & Takeuchi, K. Roles of endogenous prostaglandins and cyclooxygenase isozymes in healing of indomethacin-induced small intestinal lesions in rats. *J Pharmacol Exp Ther* 318, 691-9 (2006).

Heath, E. M., O'Brien, D. P., Banas, R., Naylor, E. W. & Dobrowolski, S. Optimization of an automated DNA purification protocol for neonatal screening. *Arch Pathol Lab Med* 123, 1154-60 (1999).

Hedrick, P. W., Whittam, T. S. & Parham, P. Heterozygosity at individual amino acid sites: extremely high levels for HLA-A and -B genes. *Proc Natl Acad Sci U S A* 88, 5897-901 (1991).

Heller, R. A. *et al.* Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci U S A* 94, 2150-5 (1997).

Herbert, A. *et al.* A common genetic variant is associated with adult and childhood obesity. *Science* 312, 279-83 (2006).

Hill, W. & Robertson, A. Linkage disequilibrium in finite populations. *Theor Appl Genet* 38, 226-231 (1968).

Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* 38, 82-5 (2006).

Hirakawa, M. *et al.* JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res* 30, 158-62 (2002).

Hochstrasser, M. Lingering mysteries of ubiquitin-chain assembly. *Cell* 124, 27-34 (2006).

Holland, P. M., Abramson, R. D., Watson, R. & Gelfand, D. H. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc Natl Acad Sci U S A* 88, 7276-80 (1991).

Holle, R., Happich, M., Lowel, H. & Wichmann, H. E. KORA--a research platform for population based health research. *Gesundheitswesen* 67 Suppl 1, S19-25 (2005).

Hoogendoorn, B. *et al.* Functional analysis of human promoter polymorphisms. *Hum Mol Genet* 12, 2249-54 (2003).

Hoshino, T. *et al.* Prostaglandin E2 protects gastric mucosal cells from apoptosis via EP2 and EP4 receptor activation. *J Biol Chem* 278, 12752-8 (2003).

Hosking, L. *et al.* Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur J Hum Genet* 12, 395-9 (2004).

Hu, N. *et al.* Genome-wide association study in esophageal cancer using GeneChip mapping 10K array. *Cancer Res* 65, 2542-6 (2005).

Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res* 30, 38-41 (2002).

Hugot, J. P. Inflammatory bowel disease: a complex group of genetic disorders. *Best Pract Res Clin Gastroenterol* 18, 451-62 (2004).

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

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

Hull, M. A., Ko, S. C. & Hawcroft, G. Prostaglandin EP receptors: targets for treatment and prevention of colorectal cancer? *Mol Cancer Ther* 3, 1031-9 (2004).

Humbert, P., Russell, S. & Richardson, H. Dlg, Scribble and Lgl in cell polarity, cell proliferation and cancer. *Bioessays* 25, 542-53 (2003).

Huttley, G. A., Smith, M. W., Carrington, M. & O'Brien, S. J. A scan for linkage disequilibrium across the human genome. *Genetics* 152, 1711-22 (1999).

Hysi, P. *et al.* NOD1 variation, immunoglobulin E and asthma. *Hum Mol Genet* 14, 935-41 (2005).

Ichimura, Y. *et al.* A ubiquitin-like system mediates protein lipidation. *Nature* 408, 488-92 (2000).

Idestrom, M., Rubio, C., Granath, F., Finkel, Y. & Hugot, J. P. CARD15 mutations are rare in Swedish pediatric Crohn disease. *J Pediatr Gastroenterol Nutr* 40, 456-60 (2005).

Inohara, N. *et al.* Nod1, an Apaf-1-like activator of caspase-9 and nuclear factor-kappaB. *J Biol Chem* 274, 14560-7 (1999).

Inoue, N. *et al.* Lack of common NOD2 variants in Japanese patients with Crohn's disease. *Gastroenterology* 123, 86-91 (2002).

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. & Godzik, A. FFAS03: a server for profile--profile sequence alignments. *Nucleic Acids Res* 33, W284-8 (2005).

Jayanthi, V., Probert, C. S., Pinder, D., Wicks, A. C. & Mayberry, J. F. Epidemiology of Crohn's disease in Indian migrants and the indigenous population in Leicestershire. *Q J Med* 82, 125-38 (1992).

Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 29, 217-22 (2001).

Johnson, G. C. *et al.* Haplotype tagging for the identification of common disease genes. *Nat Genet* 29, 233-7 (2001).

Kabashima, K. *et al.* Prostaglandin E2-EP4 signaling initiates skin immune responses by promoting migration and maturation of Langerhans cells. *Nat Med* 9, 744-9 (2003).

Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577-637 (1983).

Kanazawa, N. *et al.* Early-onset sarcoidosis and CARD15 mutations with constitutive nuclear factor-kappaB activation: common genetic etiology with Blau syndrome. *Blood* 105, 1195-7 (2005).

Kaufmann, H. J. & Taubin, H. L. Nonsteroidal anti-inflammatory drugs activate quiescent inflammatory bowel disease. *Ann Intern Med* 107, 513-6 (1987).

Kauppi, L., Sajantila, A. & Jeffreys, A. J. Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum Mol Genet* 12, 33-40 (2003).

Kawasaki, A., Tsuchiya, N., Hagiwara, K., Takazoe, M. & Tokunaga, K. Independent contribution of HLA-DRB1 and TNF alpha promoter polymorphisms to the susceptibility to Crohn's disease. *Genes Immun* 1, 351-7 (2000).

Keane, J. *et al.* Tuberculosis associated with infliximab, a tumor necrosis factor alpha-neutralizing agent. *N Engl J Med* 345, 1098-104 (2001).

Kennedy, G. C. *et al.* Large-scale genotyping of complex DNA. *Nat Biotechnol* 21, 1233-7 (2003).

Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* 12, 996-1006 (2002).

Kerlavage, A. *et al.* The Celera Discovery System. *Nucleic Acids Res* 30, 129-36 (2002).

Kiel, C. & Serrano, L. The ubiquitin domain superfold: structure-based sequence alignments and characterization of binding epitopes. *J Mol Biol* 355, 821-44 (2006).

Kim, I. & Rao, H. What's Ub chain linkage got to do with it? *Sci STKE* 2006, pe18.1-3 (2006).

Kirkegaard, K., Taylor, M. P. & Jackson, W. T. Cellular autophagy: surrender, avoidance and subversion by microorganisms. *Nat Rev Microbiol* 2, 301-14 (2004).

Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385-9 (2005).

Klement, E., Cohen, R. V., Boxman, J., Joseph, A. & Reif, S. Breastfeeding and risk of inflammatory bowel disease: a systematic review with meta-analysis. *Am J Clin Nutr* 80, 1342-52 (2004).

Kohler, J. A. & Grant, D. B. Crohn's disease in Turner's syndrome. *Br Med J (Clin Res Ed)* 282, 950 (1981).

Komatsu, M. *et al.* Tumor necrosis factor-alpha in serum of patients with inflammatory bowel disease as measured by a highly sensitive immuno-PCR. *Clin Chem* 47, 1297-301 (2001).

Kontoyiannis, D., Pasparakis, M., Pizarro, T. T., Cominelli, F. & Kollias, G. Impaired on/off regulation of TNF biosynthesis in mice lacking TNF AU-rich elements: implications for joint and gut-associated immunopathologies. *Immunity* 10, 387-98 (1999).

Korzenik, J. R. Past and current theories of etiology of IBD: toothpaste, worms, and refrigerators. *J Clin Gastroenterol* 39, S59-65 (2005).

Kosiewicz, M. M. *et al.* Th1-type responses mediate spontaneous ileitis in a novel murine model of Crohn's disease. *J Clin Invest* 107, 695-702 (2001).

Koss, K., Satsangi, J., Fanning, G. C., Welsh, K. I. & Jewell, D. P. Cytokine (TNF alpha, LT alpha and IL-10) polymorphisms in inflammatory bowel diseases and normal controls: differential effects on production and allele frequencies. *Genes Immun* 1, 185-90 (2000).

Kouranov, A. *et al.* The RCSB PDB information portal for structural genomics. *Nucleic Acids Res* 34, D302-5 (2006).

Krawczak, M., Ball, E. V. & Cooper, D. N. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am J Hum Genet* 63, 474-88 (1998).

Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet* 9, 55-61 (2006).

Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* 22, 139-44 (1999).

Kugathasan, S. *et al.* Comparative phenotypic and CARD15 mutational analysis among African American, Hispanic, and White children with Crohn's disease. *Inflamm Bowel Dis* 11, 631-8 (2005).

Kuma, A. *et al.* The role of autophagy during the early neonatal starvation period. *Nature* 432, 1032-6 (2004).

Kuma, A., Mizushima, N., Ishihara, N. & Ohsumi, Y. Formation of the approximately 350-kDa Apg12-Apg5·Apg16 multimeric complex, mediated by Apg16 oligomerization, is essential for autophagy in yeast. *J Biol Chem* 277, 18619-25 (2002).

Kuroda, S. *et al.* Biochemical characterization and expression analysis of neural thrombospondin-1-like proteins NELL1 and NELL2. *Biochem Biophys Res Commun* 265, 79-86 (1999).

Kuroda, S. & Tanizawa, K. Involvement of epidermal growth factor-like domain of NELL proteins in the novel protein-protein interaction with protein kinase C. *Biochem Biophys Res Commun* 265, 752-7 (1999).

Kurz, T. *et al.* Fine mapping and positional candidate studies on chromosome 5p13 identify multiple asthma susceptibility loci. *J Allergy Clin Immunol* 118, 396-402 (2006).

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).

Kutyavin, I. V. *et al.* 3'-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Res* 28, 655-61 (2000).

Lamina, C. *et al.* Genetic diversity in german and European populations: looking for substructures and genetic patterns. *Gesundheitswesen* 67 Suppl 1, S127-31 (2005).

Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A ligase-mediated gene detection technique. *Science* 241, 1077-80 (1988).

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

Lange, K. Mendel version 4.0: a complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. *Am J Hum Genet* 69, A1886 (2001).

Langmann, T. *et al.* Loss of detoxification in inflammatory bowel disease: dysregulation of pregnane X receptor target genes. *Gastroenterology* 127, 26-40 (2004).

Lantermann, A. *et al.* Investigation of HLA-DPA1 genotypes as predictors of inflammatory bowel disease in the German, South African, and South Korean populations. *Int J Colorectal Dis* 17, 238-44 (2002).

Lasken, R. S. & Egholm, M. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends Biotechnol* 21, 531-5 (2003).

Lawrance, I. C., Fiocchi, C. & Chakravarti, S. Ulcerative colitis and Crohn's disease: distinctive gene expression profiles and novel susceptibility candidate genes. *Hum Mol Genet* 10, 445-56 (2001).

Lee, F. I., Bellary, S. V. & Francis, C. Increased occurrence of psoriasis in patients with Crohn's disease and their relatives. *Am J Gastroenterol* 85, 962-3 (1990).

Lee, G. H., Kim, C. G., Kim, J. S., Jung, H. C. & Song, I. S. [Frequency analysis of NOD2 gene mutations in Korean patients with Crohn's disease]. *Korean J Gastroenterol* 45, 162-8 (2005).

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

Leong, R. W. *et al.* NOD2/CARD15 gene polymorphisms and Crohn's disease in the Chinese population. *Aliment Pharmacol Ther* 17, 1465-70 (2003).

Lesage, S. *et al.* CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* 70, 845-57 (2002).

Levine, B. Eating oneself and uninvited guests: autophagy-related pathways in cellular defense. *Cell* 120, 159-62 (2005).

Lewontin, R. C. On measures of gametic disequilibrium. *Genetics* 120, 849-52 (1988).

Lewontin, R. C. The Interaction Of Selection And Linkage. Ii. Optimum Models. *Genetics* 50, 757-82 (1964).

Li, D. & Roberts, R. WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cell Mol Life Sci* 58, 2085-97 (2001).

Lindberg, E., Tysk, C., Andersson, K. & Jarnerot, G. Smoking and inflammatory bowel disease. A case control study. *Gut* 29, 352-7 (1988).

Liu, P. *et al.* Candidate lung tumor susceptibility genes identified through whole-genome association analyses in inbred mice. *Nat Genet* 38, 888-95 (2006).

Livak, K. J. Allelic discrimination using fluorogenic probes and the 5' nuclease assay. *Genet Anal* 14, 143-9 (1999).

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).

Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402-8 (2001).

Lo, H. S. *et al.* Allelic variation in gene expression is common in the human genome. *Genome Res* 13, 1855-62 (2003).

Loftus, E. V., Jr. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. *Gastroenterology* 126, 1504-17 (2004).

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 33, 177-82 (2003).

Louis, E. *et al.* Tumour necrosis factor (TNF) gene polymorphism in Crohn's disease (CD): influence on disease behaviour? *Clin Exp Immunol* 119, 64-8 (2000).

Lovmar, L. & Syvanen, A. C. Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. *Hum Mutat* 27, 603-14 (2006).

Lowe, C. E. *et al.* Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes Immun* 5, 301-5 (2004).

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).

Luce, M. J. & Burrows, P. D. The neuronal EGF-related genes NELL1 and NELL2 are expressed in hemopoietic cells and developmentally regulated in the B lineage. *Gene* 231, 121-6 (1999).

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

Macklon, N. S., Geraedts, J. P. & Fauser, B. C. Conception to ongoing pregnancy: the 'black box' of early pregnancy loss. *Hum Reprod Update* 8, 333-43 (2002).

Madisch, A. *et al.* NOD2/CARD15 gene polymorphisms are not associated with collagenous colitis. *Int J Colorectal Dis* (2006).

Mahadeo, R., Markowitz, J., Fisher, S. & Daum, F. Hermansky-Pudlak syndrome with granulomatous colitis in children. *J Pediatr* 118, 904-6 (1991).

Mahida, Y. R. & Johal, S. NF-kappa B may determine whether epithelial cell--microbial interactions in the intestine are hostile or friendly. *Clin Exp Immunol* 123, 347-9 (2001).

Mansfield, K. G. *et al.* Enteropathogenic Escherichia coli and ulcerative colitis in cotton-top tamarins (Saguinus oedipus). *J Infect Dis* 184, 803-7 (2001).

Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 22, 719-48 (1959).

Maraganore, D. M. *et al.* High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet* 77, 685-93 (2005).

Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat Genet* 36, 512-7 (2004).

Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 37, 413-7 (2005).

Marino, G. & Lopez-Otin, C. Autophagy: molecular mechanisms, physiological functions and relevance in human pathology. *Cell Mol Life Sci* 61, 1439-54 (2004).

Marth, G. *et al.* Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat Genet* 27, 371-2 (2001).

Martins Silva, B. *et al.* A whole genome association study in multiple sclerosis patients from north Portugal. *J Neuroimmunol* 143, 116-9 (2003).

Matsuzaki, H. *et al.* Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* 1, 109-11 (2004).

Matsuzaki, H. *et al.* Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res* 14, 414-25 (2004).

Mawdsley, J. E. & Rampton, D. S. Psychological stress in IBD: new insights into pathogenic and therapeutic implications. *Gut* 54, 1481-91 (2005).

Mayberry, J. F. Recent epidemiology of ulcerative colitis and Crohn's disease. *Int J Colorectal Dis* 4, 59-66 (1989).

McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nat Genet* 38, 86-92 (2006).

McCartney, S. A., Mitchell, J. A., Fairclough, P. D., Farthing, M. J. & Warner, T. D. Selective COX-2 inhibitors and human inflammatory bowel disease. *Aliment Pharmacol Ther* 13, 1115-7 (1999).

McGovern, D. P. *et al.* Association between a complex insertion/deletion polymorphism in NOD1 (CARD4) and susceptibility to inflammatory bowel disease. *Hum Mol Genet* 14, 1245-50 (2005).

McGovern, D. P., van Heel, D. A., Ahmad, T. & Jewell, D. P. NOD2 (CARD15), the first susceptibility gene for Crohn's disease. *Gut* 49, 752-4 (2001).

McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404-5 (2000).

McGuigan, F. E. & Ralston, S. H. Single nucleotide polymorphism detection: allelic discrimination using TaqMan. *Psychiatr Genet* 12, 133-6 (2002).

McKnight, A. J. *et al.* A genome-wide DNA microsatellite association screen to identify chromosomal regions harboring candidate genes in diabetic nephropathy. *J Am Soc Nephrol* 17, 831-6 (2006).

Medici, V. *et al.* Extreme heterogeneity in CARD15 and DLG5 Crohn disease-associated polymorphisms between German and Norwegian populations. *Eur J Hum Genet* 14, 459-68 (2006).

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).

Miceli-Richard, C. *et al.* CARD15 mutations in Blau syndrome. *Nat Genet* 29, 19-20 (2001).

Mira, M. T. *et al.* Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature* 427, 636-40 (2004).

Mirza, M. M. *et al.* Genetic evidence for interaction of the 5q31 cytokine locus and the CARD15 gene in Crohn disease. *Am J Hum Genet* 72, 1018-22 (2003).

Mizushima, N. The pleiotropic role of autophagy: from protein metabolism to bactericide. *Cell Death Differ* 12 Suppl 2, 1535-41 (2005).

Mizushima, N. *et al.* Mouse Apg16L, a novel WD-repeat protein, targets to the autophagic isolation membrane with the Apg12-Apg5 conjugate. *J Cell Sci* 116, 1679-88 (2003).

Mizushima, N. *et al.* A protein conjugation system essential for autophagy. *Nature* 395, 395-8 (1998).

Mizushima, N., Ohsumi, Y. & Yoshimori, T. Autophagosome formation in mammalian cells. *Cell Struct Funct* 27, 421-9 (2002).

Mizushima, N., Yoshimori, T. & Ohsumi, Y. Role of the Apg12 conjugation system in mammalian autophagy. *Int J Biochem Cell Biol* 35, 553-61 (2003).

Mori, M., Yamada, R., Kobayashi, K., Kawaida, R. & Yamamoto, K. Ethnic differences in allele frequency of autoimmune-disease-associated SNPs. *J Hum Genet* 50, 264-6 (2005).

Mort, A. J., Zhan, D. & Rodriguez, V. Use of scavenger beads to remove excess labeling reagents from capillary zone electrophoresis samples. *Electrophoresis* 19, 2129-32 (1998).

Morton, N. E. & Collins, A. Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci U S A* 95, 11389-93 (1998).

Morton, N. E. *et al.* The optimal measure of allelic association. *Proc Natl Acad Sci U S A* 98, 5217-21 (2001).

Mow, W. S. *et al.* Association of antibody responses to microbial antigens and complications of small bowel Crohn's disease. *Gastroenterology* 126, 414-24 (2004).

Mueller, J. C. *et al.* Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet* 76, 387-98 (2005).

Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* 155, 335-50 (1987).

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).

Mwantembe, O. *et al.* Ethnic differences in allelic associations of the interleukin-1 gene cluster in South African patients with inflammatory bowel disease (IBD) and in control individuals. *Immunogenetics* 52, 249-54 (2001).

Namkung, J., Kim, Y. & Park, T. Whole-genome association studies of alcoholism with loci linked to schizophrenia susceptibility. *BMC Genet* 6 Suppl 1, S9 (2005).

Nasidze, I. *et al.* Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. *Eur J Hum Genet* 9, 267-72 (2001).

Negoro, K. *et al.* Analysis of the IBD5 locus and potential gene-gene interactions in Crohn's disease. *Gut* 52, 541-6 (2003).

Netea, M. G. *et al.* Nucleotide-binding oligomerization domain-2 modulates specific TLR pathways for the induction of cytokine release. *J Immunol* 174, 6518-23 (2005).

Neurath, M. F. *et al.* Predominant pathogenic role of tumor necrosis factor in experimental colitis in mice. *Eur J Immunol* 27, 1743-50 (1997).

Neurath, M. F., Pettersson, S., Meyer zum Buschenfelde, K. H. & Strober, W. Local administration of antisense phosphorothioate oligonucleotides to the p65 subunit of NF-kappa B abrogates established experimental colitis in mice. *Nat Med* 2, 998-1004 (1996).

Newman, W. G. *et al.* DLG5 variants contribute to Crohn disease risk in a Canadian population. *Hum Mutat* 27, 353-8 (2006).

Nicholas, K., Nicholas, H. & Deerfield, D. GeneDoc: Analysis and visualization of genetic variation. *EMBNEW.NEWS* 4, 14 (1997).

Nicolae, D. L., Wen, X., Voight, B. F. & Cox, N. J. Coverage and characteristics of the Affymetrix GeneChip Human Mapping 100K SNP set. *PLoS Genet* 2, e67 (2006).

Noble, C., Nimmo, E., Gaya, D., Russell, R. K. & Satsangi, J. Novel susceptibility genes in inflammatory bowel disease. *World J Gastroenterol* 12, 1991-9 (2006).

Noble, C. L. *et al.* The contribution of OCTN1/2 variants within the IBD5 locus to disease susceptibility and severity in Crohn's disease. *Gastroenterology* 129, 1854-64 (2005).

Noble, C. L. *et al.* DLG5 variants do not influence susceptibility to inflammatory bowel disease in the Scottish population. *Gut* 54, 1416-20 (2005).

Nomura, E. *et al.* Mapping of a disease susceptibility locus in chromosome 6p in Japanese patients with ulcerative colitis. *Genes Immun* 5, 477-83 (2004).

Nordborg, M. & Tavare, S. Linkage disequilibrium: what history has to tell us. *Trends Genet* 18, 83-90 (2002).

Nordgren, S., Fasth, S. & Hulten, L. Anal fistulas in Crohn's disease: incidence and outcome of surgical treatment. *Int J Colorectal Dis* 7, 214-8 (1992).

O'Callaghan, N. J., Adams, K. E., van Heel, D. A. & Cavanaugh, J. A. Association of TNF-alpha-857C with inflammatory bowel disease in the Australian population. *Scand J Gastroenterol* 38, 533-4 (2003).

Ogawa, M. *et al.* Escape of intracellular Shigella from autophagy. *Science* 307, 727-31 (2005).

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

Ogura, Y. *et al.* Nod2, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF-kappaB. *J Biol Chem* 276, 4812-8 (2001).

Ohmen, J. D. *et al.* Susceptibility locus for inflammatory bowel disease on chromosome 16 has a role in Crohn's disease, but not in ulcerative colitis. *Hum Mol Genet* 5, 1679-83 (1996).

Ohsumi, Y. Molecular dissection of autophagy: two ubiquitin-like systems. *Nat Rev Mol Cell Biol* 2, 211-6 (2001).

Okahara, S. *et al.* Inflammatory gene signature in ulcerative colitis with cDNA macroarray analysis. *Aliment Pharmacol Ther* 21, 1091-7 (2005).

Onnie, C. M. *et al.* Associations of allelic variants of the multidrug resistance gene (ABCB1 or MDR1) and inflammatory bowel disease and their effects on disease behavior: a case-control and meta-analysis study. *Inflamm Bowel Dis* 12, 263-71 (2006).

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

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

Orlicky, S., Tang, X., Willems, A., Tyers, M. & Sicheri, F. Structural basis for phosphodependent substrate selection and orientation by the SCFCdc4 ubiquitin ligase. *Cell* 112, 243-56 (2003).

Ott, J. Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. *Ann Hum Genet* 40, 443-54 (1977).

Ott, J. A chi-square test to distinguish allelic association from other causes of phenotypic association between two loci. *Genet Epidemiol* 2, 79-84 (1985).

Pagani, F. & Baralle, F. E. Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet* 5, 389-96 (2004).

Palmer, C. N. *et al.* Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat Genet* 38, 441-6 (2006).

Pardo-Manuel de Villena, F., de la Casa-Esperon, E. & Sapienza, C. Natural selection and the function of genome imprinting: beyond the silenced minority. *Trends Genet* 16, 573-9 (2000).

Pardo-Manuel de Villena, F. & Sapienza, C. Nonrandom segregation during meiosis: the unfairness of females. *Mamm Genome* 12, 331-9 (2001).

Parkes, M., Satsangi, J. & Jewell, D. Mapping susceptibility loci in inflammatory bowel disease: why and how? *Mol Med Today* 3, 546-53 (1997).

Parkes, M., Satsangi, J., Lathrop, G. M., Bell, J. I. & Jewell, D. P. Susceptibility loci in inflammatory bowel disease. *Lancet* 348, 1588 (1996).

Peeters, M. *et al.* Familial aggregation in Crohn's disease: increased age-adjusted risk and concordance in clinical characteristics. *Gastroenterology* 111, 597-603 (1996).

Peltekova, V. D. *et al.* Functional variants of OCTN cation transporter genes are associated with Crohn disease. *Nat Genet* 36, 471-5 (2004).

Persson, P. G., Ahlbom, A. & Hellers, G. Diet and inflammatory bowel disease: a case-control study. *Epidemiology* 3, 47-52 (1992).

Pickles, L. M., Roe, S. M., Hemingway, E. J., Stifani, S. & Pearl, L. H. Crystal structure of the C-terminal WD40 repeat domain of the human Groucho/TLE1 transcriptional corepressor. *Structure* 10, 751-61 (2002).

Pierik, M., Rutgeerts, P., Vlietinck, R. & Vermeire, S. Pharmacogenetics in inflammatory bowel disease. *World J Gastroenterol* 12, 3657-67 (2006).

Plenge, R. & Rioux, J. D. Identifying susceptibility genes for immunological disorders: patterns, power, and proof. *Immunol Rev* 210, 40-51 (2006).

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

Podolsky, D. K. Inflammatory bowel disease (2). *N Engl J Med* 325, 1008-16 (1991).

Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* 6, 847-59 (2005).

Price, W. H. A high incidence of chronic inflammatory bowel disease in patients with Turner's syndrome. *J Med Genet* 16, 263-6 (1979).

Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 69, 124-37 (2001).

Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant.or not? *Hum Mol Genet* 11, 2417-23 (2002).

Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 69, 1-14 (2001).

Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am J Hum Genet* 67, 170-81 (2000).

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).

Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 29, 137-40 (2001).

Pyo, J. O. *et al.* Essential roles of Atg5 and FADD in autophagic cell death: dissection of autophagic cell death into vacuole formation and cell death. *J Biol Chem* 280, 20722-9 (2005).

Quinton, J. F. *et al.* Anti-Saccharomyces cerevisiae mannan antibodies combined with antineutrophil cytoplasmic autoantibodies in inflammatory bowel disease: prevalence and diagnostic role. *Gut* 42, 788-91 (1998).

Rahman, P. *et al.* CARD15: a pleiotropic autoimmune gene that confers susceptibility to psoriatic arthritis. *Am J Hum Genet* 73, 677-81 (2003).

Regan, L. & Rai, R. Epidemiology and the medical causes of miscarriage. *Baillieres Best Pract Res Clin Obstet Gynaecol* 14, 839-54 (2000).

Reggiori, F. & Klionsky, D. J. Autophagosomes: biogenesis from scratch? *Curr Opin Cell Biol* 17, 415-22 (2005).

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

Reich, D. E., Gabriel, S. B. & Altshuler, D. Quality and completeness of SNP databases. *Nat Genet* 33, 457-8 (2003).

Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet* 17, 502-10 (2001).

Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet* 32, 135-42 (2002).

Rengarajan, K., Cristol, S. M., Mehta, M. & Nickerson, J. M. Quantifying DNA concentrations using fluorometry: a comparison of fluorophores. *Mol Vis* 8, 416-21 (2002).

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).

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

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 (2000).

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

Ritchie, M. D. Bioinformatics approaches for detecting gene-gene and gene-environment interactions in studies of human disease. *Neurosurg Focus* 19, E2 (2005).

Rodriguez, R., Chinea, G., Lopez, N., Pons, T. & Vriend, G. Homology modeling, model and software evaluation: three related resources. *Bioinformatics* 14, 523-8 (1998).

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).

Rosenstiel, P. *et al.* TNF-alpha and IFN-gamma regulate the expression of the NOD2 (CARD15) gene in human intestinal epithelial cells. *Gastroenterology* 124, 1001-9 (2003).

Rost, B., Yachdav, G. & Liu, J. The PredictProtein server. *Nucleic Acids Res* 32, W321-6 (2004).

Roth, M. P. *et al.* Familial empiric risk estimates of inflammatory bowel disease in Ashkenazi Jews. *Gastroenterology* 96, 1016-20 (1989).

Roussomoustakaki, M. *et al.* Genetic markers may predict disease behavior in patients with ulcerative colitis. *Gastroenterology* 112, 1845-53 (1997).

Roux, C. *et al.* Bone loss in patients with inflammatory bowel disease: a prospective study. *Osteoporos Int* 5, 156-60 (1995).

Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, 365-386 (2000).

Ruzicka, T. & Printz, M. P. Arachidonic acid metabolism in skin: experimental contact dermatitis in guinea pigs. *Int Arch Allergy Appl Immunol* 69, 347-52 (1982).

Rybakin, V. & Clemen, C. S. Coronin proteins as multifunctional regulators of the cytoskeleton and membrane trafficking. *Bioessays* 27, 625-32 (2005).

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

Saiki, R. K. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230, 1350-4 (1985).

Sandborn, W. J. A review of immune modifier therapy for inflammatory bowel disease: azathioprine, 6-mercaptopurine, cyclosporine, and methotrexate. *Am J Gastroenterol* 91, 423-33 (1996).

Sandborn, W. J. & Hanauer, S. B. Antitumor necrosis factor therapy for inflammatory bowel disease: a review of agents, pharmacology, clinical results, and safety. *Inflamm Bowel Dis* 5, 119-33 (1999).

Sands, B. E. Therapy of inflammatory bowel disease. *Gastroenterology* 118, S68-82 (2000).

Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74, 5463-7 (1977).

Sartor, R. B. Review article: Role of the enteric microflora in the pathogenesis of intestinal inflammation and arthritis. *Aliment Pharmacol Ther* 11 Suppl 3, 17-22; discussion 22-3 (1997).

Sashio, H. *et al.* Polymorphisms of the TNF gene and the TNF receptor superfamily member 1B gene are associated with susceptibility to ulcerative colitis and Crohn's disease, respectively. *Immunogenetics* 53, 1020-7 (2002).

Sasieni, P. D. From genotypes to genes: doubling the sample size. *Biometrics* 53, 1253-61 (1997).

Satagopan, J. M. & Elston, R. C. Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol* 25, 149-57 (2003).

Satagopan, J. M., Venkatraman, E. S. & Begg, C. B. Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 60, 589-97 (2004).

Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E. & Begg, C. B. Two-stage designs for gene-disease association studies. *Biometrics* 58, 163-70 (2002).

Satsangi, J., Jewell, D., Parkes, M. & Bell, J. Genetics of inflammatory bowel disease. A personal view on progress and prospects. *Dig Dis* 16, 370-4 (1998).

Satsangi, J., Jewell, D. P. & Bell, J. I. The genetics of inflammatory bowel disease. *Gut* 40, 572-4 (1997).

Satsangi, J., Jewell, D. P., Rosenberg, W. M. & Bell, J. I. Genetics of inflammatory bowel disease. *Gut* 35, 696-700 (1994).

Satsangi, J., Parkes, M., Jewell, D. P. & Bell, J. I. Genetics of inflammatory bowel disease. *Clin Sci (Lond)* 94, 473-8 (1998).

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).

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).

Sauer, S. *et al.* Miniaturization in functional genomics and proteomics. *Nat Rev Genet* 6, 465-76 (2005).

Sauer, S. *et al.* A novel procedure for efficient genotyping of single nucleotide polymorphisms. *Nucleic Acids Res* 28, E13 (2000).

Scarpa, R. *et al.* Juvenile rheumatoid arthritis, Crohn's disease and Turner's syndrome: a novel association. *Clin Exp Rheumatol* 14, 449-50 (1996).

Schaible, T. F. Long term safety of infliximab. *Can J Gastroenterol* 14 Suppl C, 29C-32C (2000).

Schaid, D. J. & Jacobsen, S. J. Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions. *Am J Epidemiol* 149, 706-11 (1999).

Schinella, R. A., Greco, M. A., Cobert, B. L., Denmark, L. W. & Cox, R. P. Hermansky-Pudlak syndrome with granulomatous colitis. *Ann Intern Med* 92, 20-3 (1980).

Schlup, M., Maclaurin, B. P., Barbezat, G. O. & de Lambert, B. M. Crohn's disease: a ten year retrospective review at Dunedin hospitals. *N Z Med J* 99, 141-4 (1986).

Schmid, D., Dengjel, J., Schoor, O., Stevanovic, S. & Munz, C. Autophagy in innate and adaptive immunity against intracellular pathogens. *J Mol Med* 84, 194-202 (2006).

Schreiber, S., Nikolaus, S. & Hampe, J. Activation of nuclear factor kappa B inflammatory bowel disease. *Gut* 42, 477-84 (1998).

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).

Schwartz, D. C. & Hochstrasser, M. A superfamily of protein tags: ubiquitin, SUMO and related modifiers. *Trends Biochem Sci* 28, 321-8 (2003).

Scorneaux, B., Ouadrhiri, Y., Anzalone, G. & Tulkens, P. M. Effect of recombinant human gamma interferon on intracellular activities of antibiotics against Listeria monocytogenes in the human macrophage cell line THP-1. *Antimicrob Agents Chemother* 40, 1225-30 (1996).

Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* 305, 525-8 (2004).

Shanahan, F. Nutrient tasting and signaling mechanisms in the gut V. Mechanisms of immunologic sensation of intestinal contents. *Am J Physiol Gastrointest Liver Physiol* 278, G191-6 (2000).

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

Shanahan, F. Inflammatory bowel disease: immunodiagnostics, immunotherapeutics, and ecotherapeutics. *Gastroenterology* 120, 622-35 (2001).

Shanahan, F. Crohn's disease. *Lancet* 359, 62-9 (2002).

Shanahan, F. *et al.* Hermansky-Pudlak syndrome: an immunologic assessment of 15 cases. *Am J Med* 85, 823-8 (1988).

Shapiro, D. J. Quantitative ethanol precipitation of nanogram quantities of DNA and RNA. *Anal Biochem* 110, 229-31 (1981).

Sheng, H., Shao, J., Morrow, J. D., Beauchamp, R. D. & DuBois, R. N. Modulation of apoptosis and Bcl-2 expression by prostaglandin E2 in human colon cancer cells. *Cancer Res* 58, 362-6 (1998).

Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29, 308-11 (2001).

Shiina, T., Inoko, H. & Kulski, J. K. An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens* 64, 631-49 (2004).

Shintani, T. & Klionsky, D. J. Autophagy in health and disease: a double-edged sword. *Science* 306, 990-5 (2004).

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).

Shuai, K., Schindler, C., Prezioso, V. R. & Darnell, J. E., Jr. Activation of transcription by IFN-gamma: tyrosine phosphorylation of a 91-kD DNA binding protein. *Science* 258, 1808-12 (1992).

Singer, II *et al.* Cyclooxygenase 2 is induced in colonic epithelial cells in inflammatory bowel disease. *Gastroenterology* 115, 297-306 (1998).

Singer, V. L., Jones, L. J., Yue, S. T. & Haugland, R. P. Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation. *Anal Biochem* 249, 228-38 (1997).

Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. The WD repeat: a common architecture for diverse functions. *Trends Biochem Sci* 24, 181-5 (1999).

Smyth, D. J. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat Genet* 38, 617-9 (2006).

Spielman, R. S. & Ewens, W. J. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* 62, 450-8 (1998).

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).

Sprague, E. R., Redd, M. J., Johnson, A. D. & Wolberger, C. Structure of the C-terminal domain of Tup1, a corepressor of transcription in yeast. *Embo J* 19, 3016-27 (2000).

Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 21, 577-81 (2003).

Stenzel, A. *et al.* Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Hum Genet* 114, 377-85 (2004).

Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293, 489-93 (2001).

Stokkers, P. C., Reitsma, P. H., Tytgat, G. N. & van Deventer, S. J. HLA-DR and -DQ phenotypes in inflammatory bowel disease: a meta-analysis. *Gut* 45, 395-401 (1999).

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

Stout, C. & Snyder, R. L. Ulcerative colitis-like lesion in Siamang gibbons. *Gastroenterology* 57, 256-61 (1969).

Summers, R. W. *et al.* Trichuris suis seems to be safe and possibly effective in the treatment of inflammatory bowel disease. *Am J Gastroenterol* 98, 2034-41 (2003).

Summers, R. W., Elliott, D. E., Urban, J. F., Jr., Thompson, R. & Weinstock, J. V. Trichuris suis therapy in Crohn's disease. *Gut* 54, 87-90 (2005).

Sutherland, L., Roth, D., Beck, P., May, G. & Makiyama, K. Oral 5-aminosalicylic acid for inducing remission in ulcerative colitis. *Cochrane Database Syst Rev*, CD000543 (2000).

Sutherland, L., Roth, D., Beck, P., May, G. & Makiyama, K. Oral 5-aminosalicylic acid for maintaining remission in ulcerative colitis. *Cochrane Database Syst Rev*, CD000544 (2000).

Suzuki, N. N., Yoshimoto, K., Fujioka, Y., Ohsumi, Y. & Inagaki, F. The crystal structure of plant Atg12 and its biological implication in autophagy. *Autophagy* 1, 119-26 (2005).

Swanson, M. S. & Molofsky, A. B. Autophagy and inflammatory cell death, partners of innate immunity. *Autophagy* 1, 174-6 (2005).

Taberlet, P. *et al.* Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Res* 24, 3189-94 (1996).

Takahashi, M., Ogino, T. & Baba, K. Estimation of relative molecular length of DNA by electrophoresis in agarose gel. *Biochim Biophys Acta* 174, 183-7 (1969).

Takeda, K., Kaisho, T. & Akira, S. Toll-like receptors. *Annu Rev Immunol* 21, 335-76 (2003).

Tamai, I. *et al.* Molecular and functional identification of sodium ion-dependent, high affinity human carnitine transporter OCTN2. *J Biol Chem* 273, 20378-82 (1998).

Tang, H. *et al.* Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet* 76, 268-75 (2005).

Tanida, I., Ueno, T. & Kominami, E. LC3 conjugation system in mammalian autophagy. *Int J Biochem Cell Biol* 36, 2503-18 (2004).

Tenesa, A., Noble, C., Satsangi, J. & Dunlop, M. Association of DLG5 and inflammatory bowel disease across populations. *Eur J Hum Genet* 14, 259-60; author reply 260-1 (2006).

Terwilliger, J. D., Weeks, D. E. & Ott, J. Laboratory errors in the reading of marker alleles cause massive reductions in lod score and lead to gross overestimation of the recombination fraction. *Am J Hum Genet* Suppl 47, A201 (1990).

Teuber, M. *et al.* Improving quality control and workflow management in high-throughput single-nucleotide polymorphism genotyping environments. *Journal of the Association for Laboratory Automation* 10, 43-47 (2005).

Thomas, D. C., Haile, R. W. & Duggan, D. Recent developments in genomewide association scans: a workshop summary and review. *Am J Hum Genet* 77, 337-45 (2005).

Thomas, D. C. & Witte, J. S. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev* 11, 505-12 (2002).

Thomas P.D., G. D. Beyond serendipity. *The Scientist* 16, 12 (2002).

Thomas, P. D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13, 2129-41 (2003).

Thomas, P. D. & Kejariwal, A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci U S A* 101, 15398-403 (2004).

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).

Tobler, A. R. *et al.* The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *J Biomol Tech* 16, 398-406 (2005).

Tokuhiro, S. *et al.* An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat Genet* 35, 341-8 (2003).

Torok, H. P. *et al.* Polymorphisms in the DLG5 and OCTN cation transporter genes in Crohn's disease. *Gut* 54, 1421-7 (2005).

Torok, H. P., Glas, J., Tonenchi, L., Mussack, T. & Folwaczny, C. Polymorphisms of the lipopolysaccharide-signaling complex in inflammatory bowel disease: association of a mutation in the Toll-like receptor 4 gene with ulcerative colitis. *Clin Immunol* 112, 85-91 (2004).

Trachtenberg, E. A. *et al.* HLA class II haplotype associations with inflammatory bowel disease in Jewish (Ashkenazi) and non-Jewish caucasian populations. *Hum Immunol* 61, 326-33 (2000).

Truelove, S. C. & Pena, A. S. Course and prognosis of Crohn's disease. *Gut* 17, 192-201 (1976).

Tsujimoto, Y. & Shimizu, S. Another way to die: autophagic programmed cell death. *Cell Death Differ* 12 Suppl 2, 1528-34 (2005).

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).

Valentonyte, R. *et al.* Sarcoidosis is associated with a truncating splice site mutation in BTNL2. *Nat Genet* 37, 357-64 (2005).

van den Oord, E. J. & Sullivan, P. F. A framework for controlling false discovery rates and minimizing the amount of genotyping in the search for disease mutations. *Hum Hered* 56, 188-99 (2003).

van Heel, D. A. *et al.* The IBD6 Crohn's disease locus demonstrates complex interactions with CARD15 and IBD5 disease-associated variants. *Hum Mol Genet* 12, 2569-75 (2003).

van Heel, D. A. *et al.* Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs. *Hum Mol Genet* 13, 763-70 (2004).

van Heel, D. A., McGovern, D. P. & Jewell, D. P. Crohn's disease: genetic susceptibility, bacteria, and innate immunity. *Lancet* 357, 1902-4 (2001).

van Heel, D. A. *et al.* Inflammatory bowel disease is associated with a TNF polymorphism that affects an interaction between the OCT1 and NF(-kappa)B transcription factors. *Hum Mol Genet* 11, 1281-9 (2002).

Venter, J. C. *et al.* The sequence of the human genome. *Science* 291, 1304-51 (2001).

Vermeire, S. *et al.* Association of organic cation transporter risk haplotype with perianal penetrating Crohn's disease but not with susceptibility to IBD. *Gastroenterology* 129, 1845-53 (2005).

Vermeire, S. *et al.* Genome wide scan in a Flemish inflammatory bowel disease population: support for the IBD4 locus, population heterogeneity, and epistasis. *Gut* 53, 980-6 (2004).

Vineis, P. & McMichael, A. J. Bias and confounding in molecular epidemiological studies: special considerations. *Carcinogenesis* 19, 2063-7 (1998).

von Öhsen, N., Sommer, I., Zimmer, R. & Lengauer, T. Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics* 20, 2228-35 (2004).

Vowinkel, T., Kalogeris, T. J., Mori, M., Krieglstein, C. F. & Granger, D. N. Impact of dextran sulfate sodium load on the severity of inflammation in experimental colitis. *Dig Dis Sci* 49, 556-64 (2004).

Wacholder, S., Rothman, N. & Caporaso, N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 92, 1151-8 (2000).

Wacholder, S., Rothman, N. & Caporaso, N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev* 11, 513-20 (2002).

Waetzig, G. H. *et al.* Soluble tumor necrosis factor (TNF) receptor-1 induces apoptosis via reverse TNF signaling and autocrine transforming growth factor-beta1. *Faseb J* 19, 91-3 (2005).

Wakabayashi, M. *et al.* Interaction of lp-dlg/KIAA0583, a membrane-associated guanylate kinase family protein, with vinexin and beta-catenin at sites of cell-cell contact. *J Biol Chem* 278, 21709-14 (2003).

Wallace, J. L. Prostaglandin biology in inflammatory bowel disease. *Gastroenterol Clin North Am* 30, 971-80 (2001).

Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077-82 (1998).

Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 6, 109-18 (2005).

Wang, W. Y., Cordell, H. J. & Todd, J. A. Association mapping of complex diseases in linked regions: estimation of genetic effects and feasibility of testing rare variants. *Genet Epidemiol* 24, 36-43 (2003).

Warburton, D., Stein, Z. & Kline, J. In utero selection against fetuses with trisomy. *Am J Hum Genet* 35, 1059-64 (1983).

Warner, E. E. & Dieckgraefe, B. K. Application of genome-wide gene expression profiling by high-density DNA arrays to the treatment and study of inflammatory bowel disease. *Inflamm Bowel Dis* 8, 140-57 (2002).

Weckx, S. *et al.* novoSNP, a novel computational tool for sequence variation discovery. *Genome Res* 15, 436-42 (2005).

Weidinger, S. *et al.* Association of NOD1 polymorphisms with atopic eczema and related phenotypes. *J Allergy Clin Immunol* 116, 177-84 (2005).

Weiss, R. Hot prospect for new gene amplifier. *Science* 254, 1292-3 (1991).

Wen, L. Two-step cycle sequencing improves base ambiguities and signal dropouts in DNA sequencing reactions using energy-transfer-based fluorescent dye terminators. *Mol Biotechnol* 17, 135-42 (2001).

Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 76, 887-93 (2005).

Wigley, R. D. & Maclaurin, B. P. A study of ulcerative colitis in New Zealand, showing a low incidence in Maoris. *Br Med J* 5299, 228-31 (1962).

Wilcockson, J. The differential precipition of nucleic acids and proteins from aqueous solutions by ethanol. *Anal Biochem* 66, 64-8 (1975).

Williams, C. N., Kocher, K., Lander, E. S., Daly, M. J. & Rioux, J. D. Using a genome-wide scan and meta-analysis to identify a novel IBD locus and confirm previously identified IBD loci. *Inflamm Bowel Dis* 8, 375-81 (2002).

Wirtz, S. & Neurath, M. F. Animal models of intestinal inflammation: new insights into the molecular pathogenesis and immunotherapy of inflammatory bowel disease. *Int J Colorectal Dis* 15, 144-60 (2000).

Wood, J. D. *et al.* Evidence that colitis is initiated by environmental stress and sustained by fecal factors in the cotton-top tamarin (Saguinus oedipus). *Dig Dis Sci* 45, 385-93 (2000).

Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 34, D187-91 (2006).

Xu, J., Turner, A., Little, J., Bleecker, E. R. & Meyers, D. A. Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? *Hum Genet* 111, 573-4 (2002).

Yamada, R. & Ymamoto, K. Recent findings on genes associated with inflammatory disease. *Mutat Res* 573, 136-51 (2005).

Yamamoto-Furusho, J. K. *et al.* Polymorphisms in the promoter region of tumor necrosis factor alpha (TNF-alpha) and the HLA-DRB1 locus in Mexican mestizo patients with ulcerative colitis. *Immunol Lett* 95, 31-5 (2004).

Yamazaki, K. *et al.* Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum Mol Genet* 14, 3499-506 (2005).

Yamazaki, K. *et al.* Association analysis of SLC22A4, SLC22A5 and DLG5 in Japanese patients with Crohn disease. *J Hum Genet* 49, 664-8 (2004).

Yamazaki, K., Takazoe, M., Tanaka, T., Kazumori, T. & Nakamura, Y. Absence of mutation in the NOD2/CARD15 gene among 483 Japanese patients with Crohn's disease. *J Hum Genet* 47, 469-72 (2002).

Yang, H. & Rotter, J. Genetic aspects of idiopathic inflammatory bowel disease. *Inflammatory Bowel Disease*, 301-31 (1995).

Yang, H., Taylor, K. D. & Rotter, J. I. Inflammatory bowel disease. I. Genetic epidemiology. *Mol Genet Metab* 74, 1-21 (2001).

Yap, L. M., Ahmad, T. & Jewell, D. P. The contribution of HLA genes to IBD susceptibility and phenotype. *Best Pract Res Clin Gastroenterol* 18, 577-96 (2004).

Yauk, C. L., Bois, P. R. & Jeffreys, A. J. High-resolution sperm typing of meiotic recombination in the mouse MHC Ebeta gene. *Embo J* 22, 1389-97 (2003).

Ye, Y. & Godzik, A. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res* 32, W582-5 (2004).

Yorimitsu, T. & Klionsky, D. J. Autophagy: molecular machinery for self-eating. *Cell Death Differ* 12 Suppl 2, 1542-52 (2005).

Yoshitake, S., Kimura, A., Okada, M., Yao, T. & Sasazuki, T. HLA class II alleles in Japanese patients with inflammatory bowel disease. *Tissue Antigens* 53, 350-8 (1999).

Yue, T. L. *et al.* TL1, a novel tumor necrosis factor-like cytokine, induces apoptosis in endothelial cells. Involvement of activation of stress protein kinases (stress-activated protein kinase and p38 mitogen-activated protein kinase) and caspase-3-like protease. *J Biol Chem* 274, 1479-86 (1999).

Zaahl, M. G., Winter, T., Warnich, L. & Kotze, M. J. Analysis of the three common mutations in the CARD15 gene (R702W, G908R and 1007fs) in South African colored patients with inflammatory bowel disease. *Mol Cell Probes* 19, 278-81 (2005).

Zhang, X. *et al.* Overexpression of Nell-1, a craniosynostosis-associated gene, induces apoptosis in osteoblasts during craniofacial development. *J Bone Miner Res* 18, 2126-34 (2003).

Zhang, X. *et al.* Craniosynostosis in transgenic mice overexpressing Nell-1. *J Clin Invest* 110, 861-70 (2002).

Zheng, C. Q., Hu, G. Z., Zeng, Z. S., Lin, L. J. & Gu, G. G. Progress in searching for susceptibility gene for inflammatory bowel disease by positional cloning. *World J Gastroenterol* 9, 1646-56 (2003).

Zheng, H. *et al.* Cloning and analysis of human Apg16L. *DNA Seq* 15, 303-5 (2004).

Zheng, W. *et al.* Evaluation of AGR2 and AGR3 as candidate genes for inflammatory bowel disease. *Genes Immun* 7, 11-8 (2006).

Zollner, S. *et al.* Evidence for extensive transmission distortion in the human genome. *Am J Hum Genet* 74, 62-72 (2004).

Zouali, H. *et al.* CARD4/NOD1 is not involved in inflammatory bowel disease. *Gut* 52, 71-4 (2003).

7.2 Textbooks

Altman DG. 1997. *Practical Statistics for Medical Research.* Chapman & Hall/CRC.
Armitage P, Berry G. 1987. *Statistical Methods in Medical Research.* Oxford: Blackwell
Gordis L. 2004. *Epidemiology.* Elseview Saunders
Hartl DL. 1988. *A primer of Population Genetics.* Sundelrand, Massachussetts: Sinauer Associates
Mülhardt C. 2002. *Der Experimentator (Molekularbiologie/Genomics).* Spektrum Akademischer Verlag.
Ott J. 1999. *Analysis of Human Genetic Linkage.* 3rd edition, The John Hopkins University press
Sachs L. 2003. *Angewandte Statistik.* 11th edition, Springer Verlag
Selvin S. 2004. *Statistical Analyses of Epidemiological Data.* 3rd edition, Oxford University Press
Strachan T. 2004. *Human Molecular Genetics.* 3rd edition, Garland Publishing

8 Materials

8.1 Kits, enzymes, and chemicals

Product	Manufacturer
100 bp DNA ladder	Invitrogen, Karlsruhe, Germany
15 ml reaction tubes	Sarstedt, Nümbrecht, Germany
2.2 ml 96 deep well MTP	ABgene, Epsom, UK
384 deep well storage plate (max. 300 µl)	ABgene, Epsom, UK
384 well PCR MTP	Eppendorf, Köln, Germany
384-well MT plates	Greiner Bio-One GmbH, Frickenhausen, Germany
384-well MT plates	Sarstedt, Nümbrecht, Germany
50 ml reaction tubes	BD Biosciences, Heidelberg, Germany
96-well MT plates	Costar Corning Incorporated, Cambridge, MA, USA
96-well MT plates	Sarstedt, Nümbrecht, Germany
Advantage RT-for-PCR	BD Clontech, Palo Alto, CA, USA
Agarose	Eurogentec, Köln, Germany
AmpliTaq Gold® with GeneAmp 10x PCR Buffer II & $MgCl_2$ solution	Applied Biosystems, Foster City, CA, USA
anti-ATG16L1	ABGENT, San Diego, CA
Bacillol®	Bode Chemie, Hamburg, Germany
BigDye Terminator Ready reaction kit v1.1	Applied Biosystems, Foster City, CA, USA
Biosphere® Filter Tips (10/200/1000 µl)	Sarstedt, Nümbrecht, Germany
Bromphenol blue	Sigma, München, Germany
Cell culture flasks (250 ml; canted neck)	BD Biosciences, Heidelberg, Germany
Cryotubes (2 ml)	Greiner Bio-One GmbH, Frickenhausen, Germany
dNTP set (100 mM solutions, each 100 µM)	GE Healthcare UK Limited, Buckinghamshire, UK and Amersham, Piscataway, NJ
Easy peel heat seal foil	ABgene, Epsom, UK
EDTA	Sigma, München, Germany
EDTA blood vial 10 ml	Sarstedt, Nümbrecht, Germany
EGF	PeproTech, Rocky Hill, NJ, USA
epT.I.P.S Box 100-5000 µl	Eppendorf, Köln, Germany
Ethanol pa	Merck, Darmstadt, Germany
Ethidium Bromide solution (10 mg/ml)	Invitrogen, Karlsruhe, Germany
Exonuclease I	GE Healthcare UK Limited, Buckinghamshire, UK

Table 8-1 Kits, enzymes, and chemicals

Product	Manufacturer
Flagellin	Invitrogen, Karlsruhe, Germany
GeneAmp PCR buffer system	Applied Biosystems, Foster City, CA, USA
GenomiPhi v1, v2, and high yield WGA Kit	GE Healthcare UK Limited, Buckinghamshire, UK
Glycerol	Sigma, München, Germany
goat-anti-rabbit	Vector Laboratory, Burlingame, CA
Human Multiple Tissue cDNA (MTC) Panel I	BD Clontech, Palo Alto, CA, USA
Human Multiple Tissue cDNA (MTC) Panel II	BD Clontech, Palo Alto, CA, USA
IFN-γ	BIOSOURCE, Camarillo, CA, USA
Invisorb Blood Giga Kit	Invitek, Berlin, Germany
Isopropanol	Merck, Darmstadt, Germany
LiChrosolv® double distilled water	Merck, Darmstadt, Germany
MicroAmp® optical 96-well reaction plate	Applied Biosystems, Foster City, CA, USA
MicroAmp® single strips	Applied Biosystems, Foster City, CA, USA
MicroAmp® single tubes	Applied Biosystems, Foster City, CA, USA
Microtiter plates, 96-well, round bottom w/ lid	Sarstedt, Nümbrecht, Germany
Multiscreen column loader	GE Healthcare UK Limited, Buckinghamshire, UK
PicoGreen®	Invitrogen, Karlsruhe, Germany
Proteinase K	Molecular Research Center, OH, USA
Reaction tubes (0.5/1.5/2.0 ml)	Eppendorf, Köln, Germany
SAP	GE Healthcare UK Limited, Buckinghamshire, UK
Sephadex powder (G-50 superfine)	GE Healthcare UK Limited, Buckinghamshire, UK
Sephadex spin column plates MAHVN 4550	GE Healthcare UK Limited, Buckinghamshire, UK
Serological pipettes with filter (5/10/25 ml)	Sarstedt, Nümbrecht, Germany
SmartLadder	Eurogentec, Köln, Germany
SNPlex™ System Core Kit	Applied Biosystems, Foster City, CA, USA
TAE Buffer 25x ready pack	Amresco, Solon, OH, USA
Taq DNA polymerase	Qiagen, Hilden, Germany
TaqMan® Universal PCR Master Mix	Applied Biosystems, Foster City, CA, USA
TBE Buffer 10x ready pack	Amresco, Solon, OH, USA
TGF-β	BIOSOURCE, Camarillo, CA, USA
Thermo-Fast® 384 PCR Plate	ABgene, Epsom, UK
TNF-α	BIOSOURCE, Camarillo, CA, USA
Tris	Merck, Darmstad, Germany
Triton-X	Sigma, München, Germany
Xylene Cyanol FF	Sigma, München, Germany

Table 8-1 Kits, enzymes, and chemicals

8.2 Oligonucleotides

8.2.1 Primers

Region	Primer	Sequence	Amplicon
Promoter	ATG16L_p2_F ATG16L_p2_R	5'-CACGAAAAGCAGCTTAACAATCAAAG-3' 5'-AGTGACGCCAGCCTGTAGCC-3'	828 bp
	ATG16L_p1_F ATG16L_p1_R	5'-CACAGTGCTGACTGCATTACATGG-3' 5'-GCCTCAGGTTCCCGCTGAC-3'	829 bp
Exon 01 (115 bp)	ATG16L_e01_F ATG16L_e01_R	5'-TCCGGCCCTCTCGAAAATC-3' 5'-GGGAAAATCCTCCAAAGATAAAACG-3'	505 bp
Exon 02 (94 bp)	ATG16L_e02_F ATG16L_e02_R	5'-GGGAAGACATTCTTGCAGGTG-3' 5'-TGAATCCTGGCAGGTTAGATGAG-3'	536 bp
Exon 03 (106 bp)	ATG16L_e03_F ATG16L_e03_R	5'-CTGCTGGAGACACCCGAATG-3' 5'-TGGTGATGGGCCTCAATCTG-3'	445 bp
Exon 04 (74 bp)	ATG16L_e04-2_F ATG16L_e04-2_R	5'-TGGCAGGGATAGTTCCCCTTTG-3' 5'-GCTGGTAGAAAAGGATCCCAGAGTG-3'	397 bp
Exon 05 (252 bp)	ATG16L_e05_F ATG16L_e05_R	5'-TTTCCTCTCCTAATGGATTATCCTG-3' 5'-TTGTGGTGTATTTCCTTTTTCTAACTC-3'	600 bp
Exon 06 (66 bp)	ATG16L_e06_F ATG16L_e06_R	5'-TGATGTTATGAGTTTGGGCTTGTG-3' 5'-CATTAGAAGCTATGATCACACCACTGC-3'	388 bp
Exon 07 (87 bp)	ATG16L_e07_F ATG16L_e07_R	5'-TGGCAGCTCTTCCTTTTTCTCC-3' 5'-TGCTTCCCTCCCATTAAGCAG-3'	433 bp
Exon 08 (57 bp)	ATG16L_e08_F ATG16L_e08_R	5'-AGGCTGGGTTTTCCCTTTCC-3' 5'-GCACGCAGCGAGATTAAGAGG-3'	437 bp
Exon 09 (103 bp)	ATG16L_e09_F ATG16L_e09_R	5'-CTCATTTGAGTGAGGGTGCTTTTG-3' 5'-CCATCCCTCATGCTAGCAATCC-3'	537 bp
Exon 10 (106 bp)	ATG16L_e10_F ATG16L_e10_R	5'-AGAATCTTAGTTGACCTGGGCTAGGAG-3' 5'-TGGTCAAACGATCCCTTACATAAAATG-3'	433 bp
Exon 11 (71 bp)	ATG16L_e11_F ATG16L_e11_R	5'-TCATGTTCTCTTTGTCCTGCTATTTTG-3' 5'-GCAGAACCCAAGGGTTTATCAGAG-3'	427 bp
Exon 12 (72 bp)	ATG16L_e12_F ATG16L_e12_R	5'-GCGAGTTGAAGCACACTCACG-3' 5'-GGAAACACAGATTTCCCCAAGG-3'	392 bp
Exon 13 (121 bp)	ATG16L_e13-14_F ATG16L_e13-14_R	5'-GAGTCACTGTGCCTGACCTGTTTC-3' 5'-CAAGCAGAGGCACCAACGTG-3'	548 bp
Exon 14 (106 bp)	ATG16L_e15-2_F ATG16L_e15-2_R	5'-GGCTTCATGTTTAGAGGGGCACTG-3' 5'-TTCATGGGAAAGAACAGCCAAGTG-3'	427 bp
Exon 15 (150 bp)	ATG16L_e16_F ATG16L_e16_R	5'-TGTCTTAGGGTCTGTTGATGGGAAAG-3' 5'-GGGGGTGGGTCACTACTAACCTG-3'	515 bp
Exon 16 (48 bp)	ATG16L_e17-2_F ATG16L_e17-2_R	5'-CCTGAGCTGCTCCCGTGATG-3' 5'-CAATAATGGTGGCCTGCAATTATGAAC-3'	385 bp
Exon 17 (102 bp)	ATG16L_e18_F ATG16L_e18_R	5'-CGGACGGGGCTGAAATACTG-3' 5'-AGTGGCCCCAGCTTCTCTCC-3'	456 bp
Exon 18 (94 bp)	ATG16L_e19_F ATG16L_e19_R	5'-AGTGAGCTCCTGCCTTGTCG-3' 5'-CCCATTCACGGCAAAGCTAC-3'	407 bp

Table 8-2 Primer sequences used for the mutation detection of the *ATG16L1* gene.

Region	Primer	Sequence	Amplicon
Exon 10-12	ATG16L10-12_F ATG16L10-12_R	5'-AACGCTGTGCAGTTCAGTCCAG-3' 5'-CACAGTCCAGATTCGGCTTGC-3'	231 bp
β-Actin 3-4	b-Actin_F b-Actin_R	5'-GATGGTGGGCATGGGTCAG-3' 5'-CTTAATGTCACGCACGATTTCC-3'	518 bp
GAPDH	GAPDH_s GAPDH_a	5'-CCAGCCGAGCCACATCGC-3' 5'-ATGAGCCCCAGCCTTCTCCAT-3'	360 bp

Table 8-3 Primer sequences used in the RT-PCR for *ATG16L1*.

Region	Primer	Sequence	Amplicon
Fragment 1	ATG16L_ex1-9_F ATG16L_1-9-1_R	5'-CGCCACATCTCGGAGCAAC-3' 5'-CCTGGGCTTTCTCAGCCATC-3'	depends on splice variant
Fragment 2	ATG16L_1-9-2_F ATG16L_ex1-9_R	5'-AAATGCAGCGGAAGGACAGG-3' 5'-CAAGGCAGTAGCTGGTACCCTCA-3'	depends on splice variant

Table 8-4 Primer sequences used for splice variant detection of *ATG16L1*.

Region	Primer	Sequence	Amplicon
Promoter	NEL_pro2_F NEL_pro2_R	5'-TGTTAGTAGGACAAATAGGAAGTGGGA-3' 5'-GCCGAGTCGAAAAGCCG-3'	673 bp
	NEL_pro1_F NEL_pro1_R	5'-TGACAGAGCGAATCCCGAGTAAT-3' 5'-AAGCTAGGTGGAAGCAAATGAGC-3'	663 bp
Exon 01 (208 bp)	NELL1_ex01_F NELL1_ex01_R	5'-CGCAACAAGCCACAGTAGCC-3' 5'-CGAGCAGGGCAAAGAGATCC-3'	517 bp
add. Exon (84 bp)	NEL_ade_F NEL_ade_R	5'-GGAGTGGTCTGGAGAACTGGTCT-3' 5'-GTGCCCTCAATCCAGAAAGTATTG-3'	423 bp
Exon 02 (129 bp)	NELL1_ex02_F NELL1_ex02_R	5'-AGCGGGGTAAGGGAGCAAAG-3' 5'-TGATCTCTAATGCCTCCCTCCTG-3'	481 bp
Exon 03 (151 bp)	NELL1_ex03_F NELL1_ex03_R	5'-GAAACTTGCAATCTGGATTCTTTGG-3' 5'-TGGTCCTTGGAGAGCAAACAAG-3'	515 bp
Exon 04 (171 bp)	NELL1_ex04_F NELL1_ex04_R	5'-TGCCACGAGGGTTCTCAGAC-3' 5'-GCATGCCTAAGGTCCCATTTG-3'	515 bp
Exon 05 (97 bp)	NELL1_ex05_F NELL1_ex05_R	5'-GCCCAGATGATGTGTCCAAG-3' 5'-AAACCAGTTTATTCATGTCAAGCAC-3'	464 bp
Exon 06 (73 bp)	NELL1_ex06_F NELL1_ex06_R	5'-GCCAGCTCACCCTTTGAATG-3' 5'-TCTTTCACTATCTCCAGCACCTCAG-3'	331 bp
Exon 07 (83 bp)	NELL1_ex07_F NELL1_ex07_R	5'-TCCAGACCTGAAATCCTCTGTG-3' 5'-GCATCAAAGACAGGAATGGTTATG-3'	430 bp
Exon 08 (135 bp)	NELL1_ex08_F NELL1_ex08_R	5'-TGTGGGCTTGAATGGAAAGC-3' 5'-GTGGTCCTGGAGTGGACTGG-3'	589 bp
Exon 09 (103 bp)	NELL1_ex09_F NELL1_ex09_R	5'-GTGGTCCTGGAGTGGACTGG-3' 5'-CATGTTAAGTGTCTGCCACATTGC-3'	433 bp

Table 8-5 Primer sequences used for the mutation detection of the *NELL1* gene.

Region	Primer	Sequence	Amplicon
Exon 10 (74 bp)	NELL1_ex10_F NELL1_ex10_R	5'-AGGTCTGCCTGGGAGATTTAAGG-3' 5'-ATCCAACCCACCGCAGAGAG-3'	397 bp
Exon 11 (100 bp)	NELL1_ex11_F NELL1_ex12_R	5'-TGTGGCATTATCGTTTTTAGTGGAATC-3' 5'-GGAGCAGCGCACAGAGTTTTG-3'	367 bp
Exon 12 (129 bp)	NELL1_ex12_F NELL1_ex12_R	5'-GGAAAGCCTCTTACGCCTTGG-3' 5'-TTCAAAGGTCTTGCTTTCTCATGC-3'	526 bp
Exon 13 (126 bp)	NELL1_ex13_F NELL1_ex13_R	5'-TCAGCTCAGGATGGGATCTAACG-3' 5'-TGGGAGTGGAAATCAATTATGCAG-3'	500 bp
Exon 14 (123 bp)	NELL1_ex14_F NELL1_ex14_R	5'-GCGCATAGTAAGGTACTGACAAGTGG-3' 5'-TTTCCCTGACGCACAGTACCC-3'	408 bp
Exon 15 (96 bp)	NELL1_ex15_F NELL1_ex15_R	5'-GCATGCCAGCCCTATGCTAAAC-3' 5'-AGCTCCATCCACCGTCACAAC-3'	467 bp
Exon 16 (141 bp)	NELL1_ex16_F NELL1_ex16_R	5'-TGGGATTTGCTTCTTCCAGTGAC-3' 5'-GCATGATCCCTGGCCTAATCC-3'	461 bp
Exon 17 (194 bp)	NELL1_ex17_F NELL1_ex17_R	5'-TCTCCCACACAGAGCAGAACTGAC-3' 5'-TCAAAGCAGGCTTCCCTACCC-3'	551 bp
Exon 18 (177 bp)	NELL1_ex18_F NELL1_ex18_R	5'-CTGCTCTGCAAGCAAAATCAGG-3' 5'-CTTAGCCACAGGCCCCACAG-3'	492 bp
Exon 19 (225 bp)	NELL1_ex19_F NELL1_ex19_R	5'-GATCAGCAAAAGCATTCTGAAAGAAG-3' 5'-GCCAGAGTTTCCACCATGTCC-3'	537 bp
Exon 20 (712 bp)	NELL1_ex20_F NELL1_ex20_R	5'-GGGTCTGTCAGTCCTTCCTTTC-3' 5'-TGCAAATGATCTGATAAGGGAAAC-3'	534 bp

Table 8-5 Primer sequences used for the mutation detection of the *NELL1* gene.

8.2.2 TaqMan® assays

Assay_ID	dbSNP ID	Type	Sequence
Hs99999903_m1	–	GEX	FAM 5'-GCCTCGCCTTTGCCGATCCGCCGCC-3'
Hs00250530_m1	–	GEX	FAM 5'-GCCTTGTGTGTCTTCGATGCACATG-3'
Hs00971083_m1	–	GEX	FAM 5'-TCTAAATCACACTTGCCCAACCTGC-3'
C___9095577_20	rs2241880	AoD	CCCAGTCCCCCAGGACAATGTGGAT [A/G] CTCATCCTGGTTCTGGTAAAGAAGT
–	rs951199	AbD	F 5'-GATGCCATGTACTTTTTGTTACCTGTT-3' R 5'-CCACTTCTCTGAGAGTAAAGCACAA-3' VIC 5'-CATAGGCCACTTAACA-3' FAM 5'-CATAGGCCAGTTAACA-3'
C__11472042_10	rs1992662	AoD	CCTAGACAACATACTATTCTCCTTA [A/G] CATTCCTTTATGTGTGCTCACAATG
C__11472026_10	rs1992660	AoD	GGTAGTATCACTCAAAAATTAGTTA [C/T] CATCTGCATGAGTTCTTCTTCATGG
C____392093_10	rs1793004	AoD	TGGATCAGGGCCACAATCTTGTATC [C/G] GCAGTCACCACCTCTATAAAGCTTG
–	rs2075822	AbD	F 5'-CGGAGGGCATCGGGAA-3' R 5'-CAAGGGCCATGGTCATGAGT-3' VIC 5'-CTGGTCCATGGTGCCA-3' FAM 5'-CTGGTCCATGATGCCA-3'
–	rs2907748	AbD	F 5'-CCTGTACCTGGGCTCCTATTTC-3' R 5'-GGAGGGTGGGCTCCTCTA-3' VIC 5'-CAGGTAGCTGGGCTAA-3' FAM 5'-CAGGTAGCTAGGCTAA-3'
–	ND_1+32656	selfD	F 5'-GTCCTTCTGGTGTACTGATGTATGAAA-3' R 5'-GAACAGCAAATCAATCTCTGAGGTT-3' VIC 5'-CGCCCCCCACACA-3' FAM 5'-CCCCCCCCCCACAC-3'
C___3203197_10	rs8176785	AoD	AGTGAGAAATTAATTCAGCTGTTCC [A/G] GAACAAGAGTGAATTCACCATTTTG
C__32647553_10	rs8176786	AoD	CCAGCGGATTTTAACCAAGAGCTGT [C/T] GGGAATGCCGAGTAAGTGTTAATTT

Table 8-6 TaqMan® assays. Primer/Probe sequences are listed if they were provided by Applied Biosystems. GEX = gene expression assay; AbD = Assay-by-Design; AoD = Assay-on-Demand. In general, allele 1 is labelled with VIC and allele 2 with FAM.

Assay_ID	dbSNP ID	Type	Sequence
–	NELL1_G-A	AbD	F 5'-GCATGGCAGATGGACAATGG-3' R 5'-AGGAGATGAGAGGCGCTAACT-3' VIC 5'-ACAAGGTTGCACTGTC-3' FAM 5'-CACAAGGTTACACTGTC-3'
–	NELL1_G-C	AbD	F 5'-AGGGATGAGATTCGGTATCACTACA-3' R 5'-TGCCATGCGGTAAGGAAGTG-3' VIC 5'-CTCTGTCCTTGGCTTC-3' FAM 5'-CTGTGCTTGGCTTC-3'

Table 8-6 TaqMan® assays. Primer/Probe sequences are listed if they were provided by Applied Biosystems. GEX = gene expression assay; AbD = Assay-by-Design; AoD = Assay-on-Demand. In general, allele 1 is labelled with VIC and allele 2 with FAM.

8.2.3 SNPlex™ Pools

#	Pool ID	First SNP	Description
1	Control A	hCV9627665	Control Pool A
2	w1025013003_0001	hCV1767498	TaqMan® Verification Pool
3	w0506290359_0001	rs9290910	Affymetrix follow-up 1
4	w0506290539_0001	rs7681167	Affymetrix follow-up 2
5	w0506290539_0002	rs1488927	Affymetrix follow-up 2
6	w0506270719_0001	rs210465	Affymetrix follow-up X
7	w0506271309_0001	rs965306	Granuloma
8	w0506271309_0002	rs10499108	Granuloma
9	w0508020231_0001	rs4743484	Affymetrix follow-up As and Bs
10	w0508020231_0002	rs1429794	Chromosome 5 und NELL1 follow-up
11	w0508240811_0001	rs6699460	Genizon/ICMB overlaps
12	w0510100166_0001	rs10487428	Residual Affymetrix SNPs
13	w0510100167_0001	rs1063193	cSNP follow-up 1
14	w0510100167_0002	rs7259764	cSNP follow-up 1
15	w0510100167_0003	rs1785506	cSNP follow-up 1
16	w0510100167_0004	rs2361403	cSNP follow-up 1
17	w0510100167_0005	rs1320305	cSNP follow-up 1
18	w0512100656_0001	hCV11507064	cSNP follow-up 2
19	w0512100656_0002	hCV25995063	cSNP follow-up 2
20	w0512100663_0001	hCV15880968	cSNP follow-up 2
21	w0603102006_0001	rs327025	*ATG16L1*/*NELL1* final fine mapping
22	w0603102006_0002	rs1519735	*ATG16L1*/*NELL1* final fine mapping
23	w0603102006_0003	rs6431660	*ATG16L1*/*NELL1* final fine mapping
24	w0603102006_0004	rs3792106	*ATG16L1*/*NELL1* final fine mapping

Table 8-7 SNPlex™ pools. All used SNPlex™ pools are listed in this table. Corresponding assay information files (AIFs) can be found on the attached DVD (folder: "`AIF`"). The 428 pools of the cSNP experiment are not listed in this table. AIFs of the latter are on the DVD in the folder "`AIFs_cSNP Project for ABI`".

8.3 Machines

8.3.1 Centrifuges

Heraeus Biofuge fresco	Kendro, Hanau, Germany
Heraeus Biofuge pico	Kendro, Hanau, Germany
Heraeus Labofuge 400	Kendro, Hanau, Germany
Heraeus Multifuge 3S-R	Kendro, Hanau, Germany
Heraeus Varifuge 3.2RS	Kendro, Hanau, Germany
Micro Centrifuge	Roth, Karlsruhe, Germany

8.3.2 Thermocyclers

ABI Prism™ 7700 Sequence Detector	Applied Biosystems Inc., Foster City, CA, USA
ABI Prism™ 7900HT Sequence Detection System	Applied Biosystems Inc., Foster City, CA, USA
Biometra® T Gradient	Whatman Biometra GmbH, Göttingen, Germany
Biometra® T1 Thermocycler	Whatman Biometra GmbH, Göttingen, Germany
GeneAmp® PCR System 9700	Applied Biosystems Inc., Foster City, CA, USA

8.3.3 Electrophoresis

BioDoc Analyzer	Biometra, Göttingen, Germany
Gel Doc XR	Bio-Rad, München, Germany
Gibco BRL Electrophoresis Power Supply 250 EX	BioRad, München, Germany
Gibco BRL Horizontal Gel Electrophoresis Apparatus	BioRad, München, Germany
High Performance UV Transilluminator	VWR, Hamburg, Germany
Horizontal Electrophoresis Apparatus	Bio-Rad, München, Germany
KERN 440-47N scale	Kern & Sohn, Balingen, Germany
Microwave R-2V18	Sharp Electronics, Hamburg, Germany
Power Pac 300 Electrophoresis Power Supply	Bio-Rad, München, Germany

8.3.4 Pipetting robots

Hydra 384 Robbins Scientific	Dunn Labortechnik, Asbach, Germany
Hydra 96 Robbins Scientific	Dunn Labortechnik, Asbach, Germany
Power Washer PW384	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Carousel for Evo 150	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Freedom Evo 150	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Freedom Evo 200	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Genesis RSP 150	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Genesis Workstation 150	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Genesis Workstation 200	Tecan, Deutschland GmbH, Crailsheim, Germany
Tecan Spectrafluor Plus	Tecan, Deutschland GmbH, Crailsheim, Germany
Te-MO	Tecan, Deutschland GmbH, Crailsheim, Germany
Te-MO with cooling rack	Tecan, Deutschland GmbH, Crailsheim, Germany
WRC96 washing station	Tecan, Deutschland GmbH, Crailsheim, Germany

8.3.5 Other machines

3700 DNA Analyzer	Applied Biosystems Inc., Foster City, CA, USA
3730xl DNA Analyzer	Applied Biosystems Inc., Foster City, CA, USA
Axiocam	Zeiss, Jena, Germany
Axiophot microscope	Zeiss, Jena, Germany
Bambi Compressor DT/23Q	Bambi, Birmingham, UK
GFL 1086 shaking waterbath	GFL, Burgwedel, Germany
Heraeus 3 incubator	Kendro, Hanau, Germany
Heraeus Kelvitron® t	Kendro, Hanau, Germany
Mini Vortexer VM-3000	VWR, Darmstadt, Germany
PCR chambers	Bä-RO® Technology, Leichlingen, Germany
Platesealer ALPS-300	Abgene, Epsom, UK
Thermomixer 5437	Eppendorf, Köln, Germany
TiMix Control incl. TH15 hood	Edmund Bühler Labortechnik, Hechingen, Germany
Vortex-GENIE 2 G-560E	Scientific Industries, Bohemia, NY, USA

8.4 Electronic data processing

8.4.1 Laboratory information management system (LIMS)

An overview of the in-house LIMS is given in section 2.1 on page 24, Hampe *et al.* (2001), and Teuber *et al.* (2005).

8.4.2 Software

FACTS 4.82	Tecan, Deutschland GmbH, Crailsheim, Germany
fastPHASE 1.1	http://www.uwopendoor.org/ViewSoftware.asp?softwareid=3/
Gemini 4.28	Tecan, Deutschland GmbH, Crailsheim, Germany
Genemapper 4.0	Applied Biosystems, Foster City, CA, USA
GENOMIZER	http://www.ikmb.uni-kiel.de/genomizer/
Haploview 3.32	http://www.broad.mit.edu/mpg/haploview/ Daly lab, Broad institute, Cambridge, MA 02141, USA
novoSNP 2.0.3	http://www.molgen.ua.ac.be/bioinfo/novosnp/
Primer Express 2.0	http://www.applied-biosystems.com Applied Biosystems, Foster City, CA, USA
R	http://www.r-project.org
rcexact.exe for Fisher's exact	http://www.qimr.edu.au/davidD/
SAS/STAT	http://www.sas.com/technologies/analytics/statistics/stat/ Cary, NC, USA
Sequence Detection System 2.1	Applied Biosystems, Foster City, CA, USA
Sequencher 4.2 and 4.5	http://www.genecodes.com Gene Codes Corporation, Ann Arbor, MI, USA)
SNPbrowser 3.5	http://www.allsnps.com Applied Biosystems, Foster City, CA, USA
SPSS 13.0	http://www.spss.com Chicago, IL, USA
UNPHASED package: COCAPHASE v2.403 TDTPHASE v2.403	http://portal.litbio.org/Registered/Help/unphased/
WHAP 2.09	http://pngu.mgh.harvard.edu/~purcell/whap/

8.4.3 Web resources

Statistical analyses

SISA	http://home.clara.net/sisa/
Tagger	http://www.broad.mit.edu/mpg/tagger/

Protein structure prediction

Arby	http://arby.bioinf.mpi-inf.mpg.de/arby/jsp/index.jsp
BioInfoBank	http://bioinfo.pl/meta/
DSSP database	http://www.cmbi.kun.nl/gv/dssp/
Ensembl	http://www.ensembl.org
FATCAT	http://fatcat.ljcrf.edu
FFAS03	http://ffas.ljcrf.edu
Genedoc	http://www.psc.edu/biomed/genedoc/
Jalview	http://www.jalview.org
MUSCLE	http://www.drive5.com/muscle/
Pfam	http://www.sanger.ac.uk/Software/Pfam/
Porter	http://distill.ucd.ie/porter/
POV-Ray	http://www.povray.org
PROFsec	http://www.predictprotein.org
Protein Data Bank	http://www.pdb.org
PSIPRED	http://bioinf.cs.ucl.ac.uk/psipred/
SCOP database	http://scop.mrc-lmb.cam.ac.uk/scop/index.html
SeaView	http://pbil.univ-lyon1.fr/software/seaview.html
Smart	http://smart.embl-heidelberg.de
UniProt	http://www.uniprot.org
WHAT IF	http://swift.cmbi.kun.nl/WIWWWI/
Yasara	http://www.yasara.org
YASPIN	http://ibivu.cs.vu.nl/programs/yaspinwww/

SNP databases

Celera	http://www.celera.com
Genetic Association db	http://geneticassociationdb.nih.gov/cgi-bin/index.cgi
HapMap	http://www.hapmap.org
HGMD	http://www.hgmd.cf.ac.uk/ac/index.php
JSNP	http://snp.ims.u-tokyo.ac.jp
NCBI dbSNP	http://www.ncbi.nlm.nih.gov/SNP/

Dictionaries and Encyclopedias

American Heritage	http://www.bartleby.com/61/
Wikipedia	http://en.wikipedia.org

Other used web resources

BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
CEPH	http://www.cephb.fr
Genecards	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed
Google	http://www.google.de
Google Scholar	http://scholar.google.de
MatInspector	http://www.genomatix.de/products/MatInspector/
MultAlin	http://protein.toulouse.inra.fr/multalin/multalin.html
Primer3	http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi
PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed
RepeatMasker	http://woody.embl-heidelberg.de/repeatmask/
SNPeffect	http://snpeffect.vib.be/index.php
UCSC Genome Browser	http://genome.ucsc.edu

9 Appendix

9.1 Summary of IBD linkage regions

Locus	Location	Reference	Known and proposed candidate genes
–	1p12	Williams *et al.*, 2002 (only UC)	
–	1p13–1q3	Vermeire *et al.*, 2004	*DAP-3, IL-6R, MUC-1*
IBD7	1p32–36	Cho *et al.*, 1998 (UC and CD) Cho *et al.*, 2000 Vermeire *et al.*, 2004	*Ephrin A2, MFAP2, RNU1A, VCAM-1, IL-12Rβ2, TNFRSF1B (TNFR2)*
–	1q43–44	Hampe *et al.*, 1999 (IBD)	
–	2q24–32	Barmada *et al.*, 2004 (only UC)	
–	3p14	Rioux *et al.*, 2000	
IBD9	3p21–26	Satsangi *et al.*, 1996 Hampe *et al.*, 2001 Duerr *et al.*, 2002	*GNAI2, CCR2, CCR5, IL-4RA, IL-5RA, IFN-α A2,*
–	3q24	Williams *et al.*, 2002 (only UC)	
–	3q26–28	Cho *et al.*, 1998 (UC and CD) van Heel *et al.*, 2003	
–	4q21–31	Cho *et al.*, 1998 (UC and CD) Hampe *et al.*, 1999 (only UC) Vermeire *et al.*, 2004	*IL-21, IL-15, IL-21, IL-15*
–	5q14	Barmada *et al.*, 2004 (only CD)	
IBD5	5q31–33	Rioux *et al.*, 2000 (only CD)	*SLC22A4/SLC22A5* *SPINK5, IL-3, IL-4, IL-5, IL-13, CSF-2*
IBD3	6p	Satsangi *et al.*, 1996 (only UC) #2 Satsangi *et al.*, 1996 (only UC) Hampe *et al.*, 1999 (IBD) #2 Hampe *et al.*, 1999 (UC and CD) Stokkers *et al.*, 1999 (metaanalysis) Williams *et al.*, 2002 (IBD) van Heel *et al.*, 2003 (only CD)	MHC region, *TNF-α, complement C4, C2*
–	6q16–27		*IDDM-15*
–	6q24–26	Barmada *et al.*, 2004 (IBD)	
–	7p13–15	Satsangi *et al.*, 1996	*CARD4*; 7p21: *AGR2* (Zheng *et al.*, 2006)
–	7q21	Satsangi *et al.*, 1996	*MUC3, HGF, EGFR*
–	8q12–13	Barmada *et al.*, 2004 (IBD)	

Table 9-1 Summary of IBD loci identified by linkage studies. Most of the IBD linkage literature was reviewed. A few examples of replication failures are listed for *IBD1* on chromosome 16.This failure hints at a particular weakness of the method and genetic heterogeneity.

Locus	Location	Reference	Known and proposed candidate genes
–	10p12–15	Hampe *et al.*, 1999 (only CD) Vermeire *et al.*, 2004	*MRC-1, IL-2R, IL-15Rα*
–	10q23	Hampe *et al.*, 1999	*DLG5* (Stoll *et al.*, 2004)
–	11p15	(Cho *et al.*, 1998) Williams *et al.*, 2002 (IBD)	
–	11q22–23	Williams *et al.*, 2002 (only CD) Vermeire *et al.*, 2004	*MMP-1, MMP-3, MMP-12, MMP-13, MMP-20, API1, API2*
IBD2	12p13.2–q24.1	Satsangi *et al.*, 1996 Duerr *et al.*, 1998 (UC and CD) Curran *et al.*, 1998 (UC and CD) Hampe *et al.*, 1999 Barmada *et al.*, 2004 (IBD and UC)	*IFN-γ, NRAMP2, VDR, STAT6, ITGB7*
–	12q12	Barmada *et al.*, 2004 (IBD and UC)	
IBD4	14q11–12	Ma *et al.*, 1999 Duerr *et al.*, 2000 (only CD) Rioux *et al.*, 2000 Vermeire *et al.*, 2004	*T cell receptor α/δ complex, ISGF3G, GZMB, LTB4R, DAD-1, MMP-14*
–	15q26	Barmada *et al.*, 2004 (IBD and CD)	
IBD8	16p12	Hampe *et al.*, 2002	
IBD1	**16p12–q13**	Hugot *et al.*, 1996 (only CD) Ohmen *et al.*, 1996 Parkes *et al.*, 1996 Brant *et al.*, 1998 Cavanaugh *et al.*, 1998 Curran *et al.*, 1998 (only CD) Cho *et al.*, 1998 (only CD) Hampe *et al.*, 1999 Annese *et al.*, 1999 Williams *et al.*, 2002 van Heel *et al.*, 2003 NOT: Satsangi *et al.*, 1996 (other peak than Hugot *et al.*) Rioux *et al.*, 1998 Rioux *et al.*, 2000 Barmada *et al.*, 2004 Vermeire *et al.*, 2004	*CARD15* *BRD7* (*personal communications*) *CD11 integrin cluster, CD19, IL-4R, SPN*
–	16q23	Williams *et al.*, 2002 (only CD)	
–	17p12	Williams *et al.*, 2002 (only CD)	

Table 9-1 Summary of IBD loci identified by linkage studies. Most of the IBD linkage literature was reviewed. A few examples of replication failures are listed for *IBD1* on chromosome 16.This failure hints at a particular weakness of the method and genetic heterogeneity.

Locus	Location	Reference	Known and proposed candidate genes
–	17q21–q23	Ma *et al.*, 1999 Vermeire *et al.*, 2004	
–	18q21–22	Ma *et al.*, 1999 Duerr *et al.*, 2000	
IBD6	19p13	Rioux *et al.*, 2000 (UC and CD) van Heel *et al.*, 2003 (CD w/o CARD15)	*ICAM1, C3, TYK2, JAK3, TBXA2R, LTB4H*
–	20p12	Vermeire *et al.*, 2004	
–	22q11–12	Hampe *et al.*, 1999 (IBD) Barmada *et al.*, 2004	
–	Xp21	Hampe *et al.*, 1999 (IBD)	
–	Xq21–25	Hampe *et al.*, 1999 (only UC) Rioux *et al.*, 2000 van Heel *et al.*, 2003 Vermeire *et al.*, 2004	*IL-2Rγ, IL-13Rα2*

Table 9-1 Summary of IBD loci identified by linkage studies. Most of the IBD linkage literature was reviewed. A few examples of replication failures are listed for *IBD1* on chromosome 16.This failure hints at a particular weakness of the method and genetic heterogeneity.

9.2 Construction of coding SNP set

9.2.1 SNP database for marker selection

SNP data were obtained from three sources:

1. The Celera Human RefSNP database (CRA, Kerlavage *et al.*, 2002), version 3.4, which included about 2.4 million SNPs discovered during the shotgun sequencing of the Human genome by Celera Genomics (Venter *et al.*, 2001), as well as 2.2 million imported from public sources, mainly dbSNP (Sherry *et al.*, 2001), JSNP (Hirakawa *et al.*, 2002), and HGMD (Stenson *et al.*, 2003).
2. The Applera Corp. SNP Project (ASP) database, (Adams *et al.*, 2002), which consists of 266,135 SNPs discovered in 20 European Americans and 19 African Americans by Sanger sequencing of PCR amplicons overlapping the exons of 23,363 genes annotated by Celera Genomics (Bustamante *et al.*, 2005).
3. SNPs included in NCBI's dbSNP database (Sherry *et al.*, 2001), release 117.

All SNPs were mapped to the Celera Human genome assembly Release 27 and only those that mapped to unique locations, after removing redundancy, were advanced in the process (Table 9-2).

Source	Number of SNPs
Celera RefSNP database	4,039,783
Applera SNP Project database	266,135
NCBI dbSNP release 117	4,006,579
Total non-redundant, uniquely mapped	5,560,475

Table 9-2 Sources for SNPs used in marker selection.

9.2.2 SNP selection and assay development

The SNP selection process was aimed at developing a comprehensive list of common putative functional cSNPs from all possible sources (Thomas *et al.*, 2002), and to avoid putative SNPs that are rare variants or potential sequencing or analysis artifacts (Marth *et al.*, 2001). We thus triaged SNPs based on their measured or expected heterozygosity in populations of European and African descent. For this we used the allele frequency data obtained during the ASP project and from the genotyping of 177,781 SNPs in about 45 samples each of European American, African American, with TaqMan® Validated SNP Genotyping Assays (De La Vega *et al.*, 2002; De La Vega *et al.*, 2005). When no allele frequency information in population panels was available, we looked for evidence of independent discovery (so called "double-hit" SNPs, Reich *et al.*, 2003). We used as evidence the ASP project calls, the donor information of the Celera shotgun reads, and the dbSNP submission handles. SNPs whose minor alleles were observed in at least two distinct donors in either the ASP or Celera shotgun SNP discovery were selected. We identified SNPs discovered independently by Celera or ASP and by the public SNP discovery efforts. We also compared single-donor Celera SNPs to the NCBI human genomic assembly to find cases where the Celera minor allele was confirmed in the public consensus sequence. Finally, for SNPs when the single source was dbSNP, SNPs with at least 3 distinct submission handles were included. We compiled 1,601,782 SNPs that meet these requirements. We then identified non-synonymous SNPs (nsSNPs), either missense or nonsense, by mapping these SNPs to Celera (Kerlavage *et al.*, 2002; Venter *et al.*, 2001), RefSeq/LocusLink (Pruit *et al.*, 2001), and ENSEMBL (Hubbard *et al.*, 2002) transcripts. At the end of the process we obtained 28,709 nsSNPs in at least one transcript to be submitted for assay design (Fig. 9-1).

To avoid designing oligoprobes overlapping other common SNPs, we "masked" the context sequence of target SNPs for other adjacent SNPs using only the list of triaged SNPs described above. Masked context sequence and allele information was submitted to the assay design pipeline of the SNPlex™ Genotyping System (Tobler *et al.*, 2005), version 1.0, in batches no larger than 2,500 SNPs. Design batches were organized to group SNPs belonging to a candidate gene lists related with immunity and inflammation, and by similar molecular function as predicted by the PANTHER protein classification (Thomas *et al.*, 2003), when possible. After removing SNP assays that failed to meet the oligoprobe design or genome specificity rules, we obtained and manufactured 19,779 SNPlex™ SNP genotyping assays distributed in 428 multiplexes with probes to type up to 48 SNPs each. Please refer to file "`Supplemental_Table_01.xls`" on the DVD for details on the distribution and annotations of SNPs for each SNPlex™ multiplex pool.

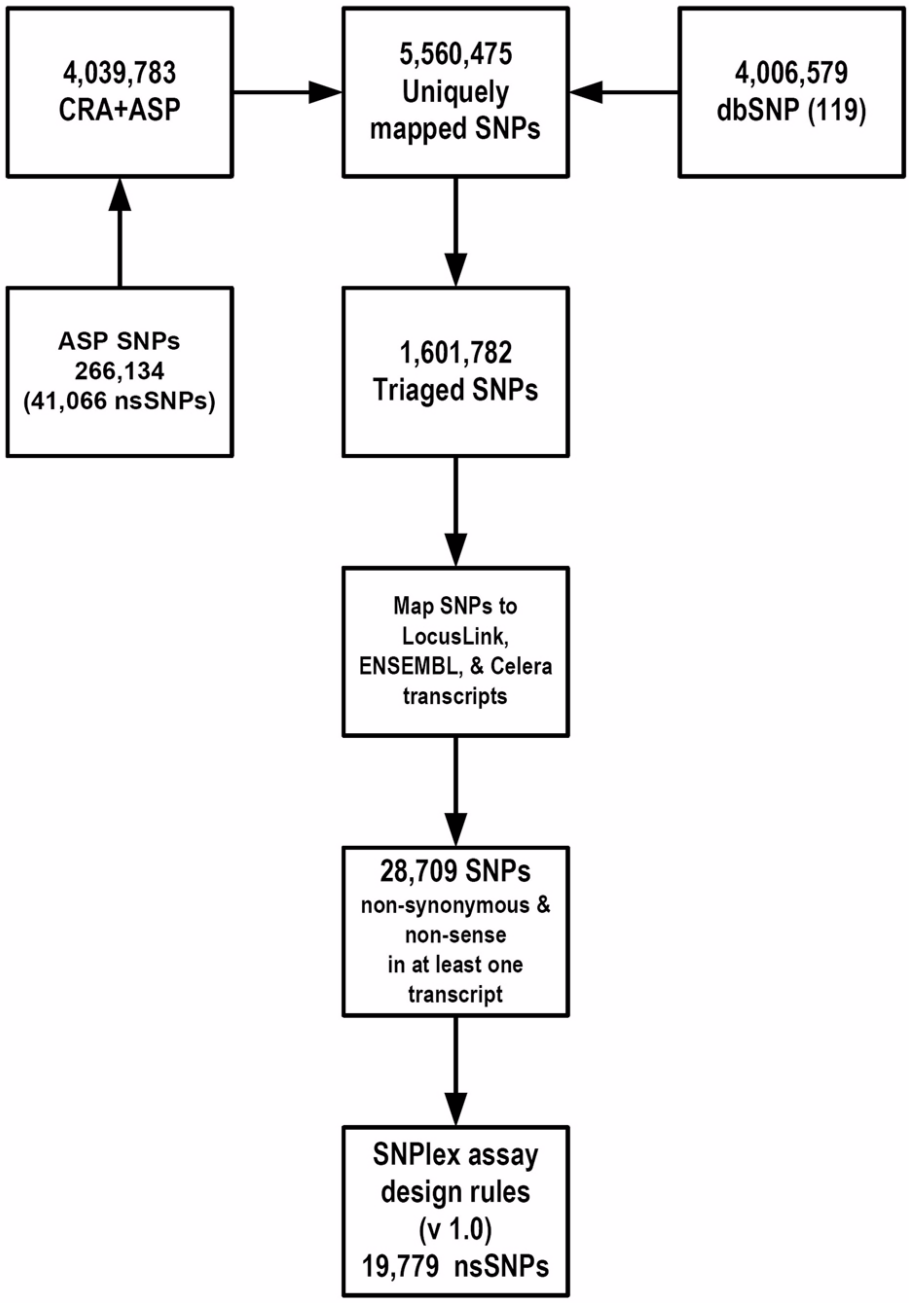

Fig. 9-1 SNP selection and assay development system. For more details see description on page 177.

9.2.3 Distribution and features of nsSNPs panel

Fig. 9-2 shows the distribution of the set of nsSNPs in our final panel across the human chromosomes.

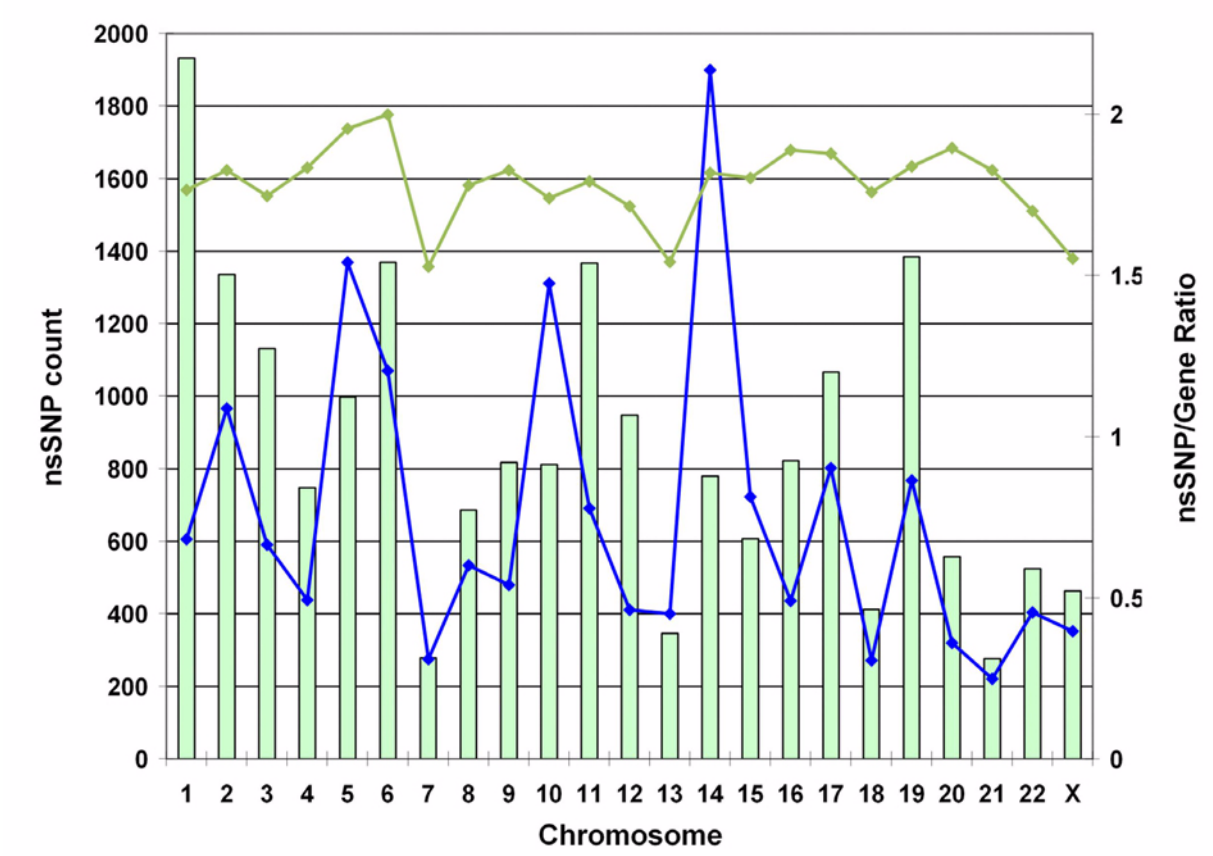

Fig. 9-2 Distribution of nsSNP panel across human chromosomes. The bars show the actual number of nsSNPs per chromosome, and the lines represent the nsSNP/gene ratio, based on the Celera gene annotation (R27). The raw SNP to gene ratio (blue line) shows an apparent higher than average (0.75 nsSNP/gene) value for chromosomes 5, 10, and 14, whereas chromosomes 7, 18, 20, 21, and X show a lower SNP to gene ratio. However, this apparent distortion disappears if we normalize to count only genes with at least one nsSNPs in our set (green line), where the genome average is 1.78 nsSNP/gene. The total number of Celera genes covered by nsSNP in our panel is 9,672 (out of 25,030 genes in the Celera R27 annotation).

We then analyzed the distribution of genes with nsSNPs included in our set using the PANTHER protein function classification (Thomas *et al.*, 2003). Of interest was to ascertain if particular functional classes are unrepresented (i.e. genes classes where common non-synonymous SNPs are rare) or overrepresented (i.e. gene classes with frequent common variation). A binomial distribution test was used to calculate p-value for the observed vs. expected category representation as compared to the entire gene complement (Cho *et al.*, 2000), as per the Celera human gene annotation, Release 27. The analysis shown in table 9-3, page 180 shows that certain molecular function categories are over- or underrepresented in our panel with statistical significance. Note that a large proportion of genes are not currently classified (3373 out of 9672, i.e. 35%).

Category	hCG R27 (n=25030)	nsSNP (n=9672)	Expected	Over (+) Under (-)	p-value
Protein biosynthesis	837	180	321.12	-	2.10E-18
Pre-mRNA processing	226	51	86.71	-	2.16E-05
Chromatin packaging and remodeling	185	43	70.98	-	2.28E-04
Protein folding	164	38	62.92	-	4.73E-04
Cell adhesion	533	313	204.49	+	6.18E-13
Olfaction	296	192	113.56	+	9.44E-12
Chemosensory perception	328	202	125.84	+	1.94E-10
Signal transduction	3387	1507	1299.46	+	7.20E-10
Cell surface receptor mediated signal transd.	1601	748	614.24	+	3.49E-08
Cell adhesion-mediated signaling	292	173	112.03	+	4.65E-08
Sensory perception	588	307	225.59	+	1.09E-07
Cell structure and motility	1064	502	408.21	+	2.45E-06
Proteolysis	734	360	281.61	+	2.93E-06
Cell structure	679	334	260.50	+	5.22E-06
G-protein mediated signaling	897	425	344.14	+	9.83E-06
Cell communication	1250	567	479.57	+	3.58E-05
Extracellular matrix protein-mediated sign.	54	40	20.72	+	1.08E-04
Lipid, fatty acid, and steroid metabolism	678	317	260.12	+	2.92E-04

Table 9-3 Gene representation of nsSNP panel. Highly significant p-values are marked with a yellow background. For further details of the underlying statistics see description on page 179.

Underrepresented in our panel are classes of genes known to be highly conserved and that carry out several fundamental cellular processes (e.g. protein synthesis, chromatin packaging genes), whereas overrepresented gene classes include some classes that are know for presenting higher levels of genetic variation (e.g. olfactory receptors, cell surface/adhesion). This suggest that selection pressure might limit common potentially deleterious polymorphisms in highly conserved genes participating in fundamental cellular processes, whereas at the other extreme selection may favour common functional variation on certain classes of genes that deal with environmental interactions and other functions.

We tailored our SNP selection towards variants that are common in the population, rather than those that might have a potential deleterious impact in protein structure, as it is unclear the importance of such SNPs in common-complex disease (Thomas *et al.*, 2004). To show this, we obtained the PANTHER position-specific evolutionary conservation score (subPSEC) for 9,035 SNPs in our panel. The subPSEC score is the negative logarithm of the probability ratio of the two variant amino acids arising from a cSNP in a particular gene context. More negative values represent mutations with greater potential impact in protein structure since they are strongly avoided in natural sequences. Fig. 9-3 shows that potentially highly deleterious mutations with large negative subPSEC score values are somewhat rare in our set and that, in fact, our panel includes SNPs with a wide range of subPSEC values (Thomas *et al.*, 2004). We hypothesized that this distribution would ensure that our panel have the highest coverage possible of variants that might be involved in complex disease and which are unlikely to be under strong purifying selection, in contrast with variants associated with Mendelian diseases which are more deleterious and rare in the population (Thomas *et al.*, 2004).

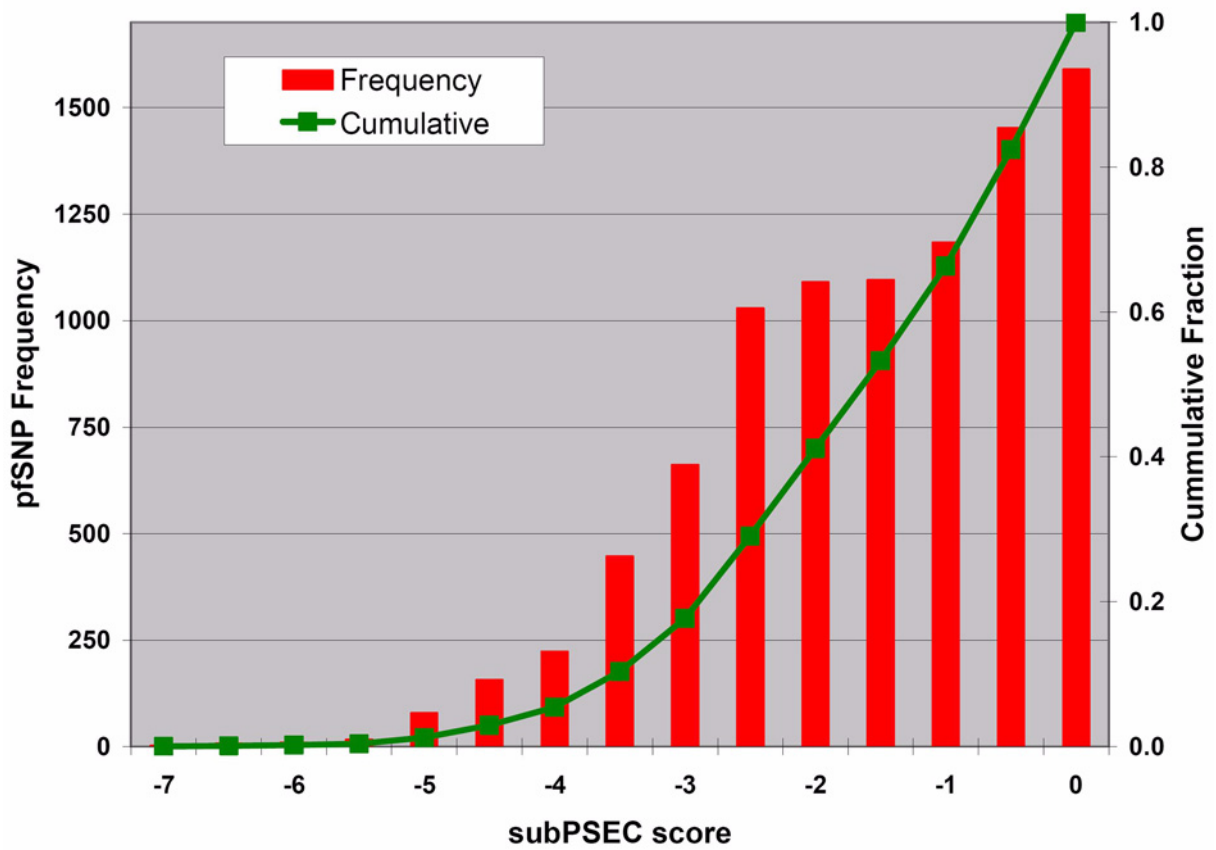

Fig. 9-3 Distribution of subPSEC score for the nsSNPs in panel. For further details see text above.

9.3 GENOMIZER Sample Output

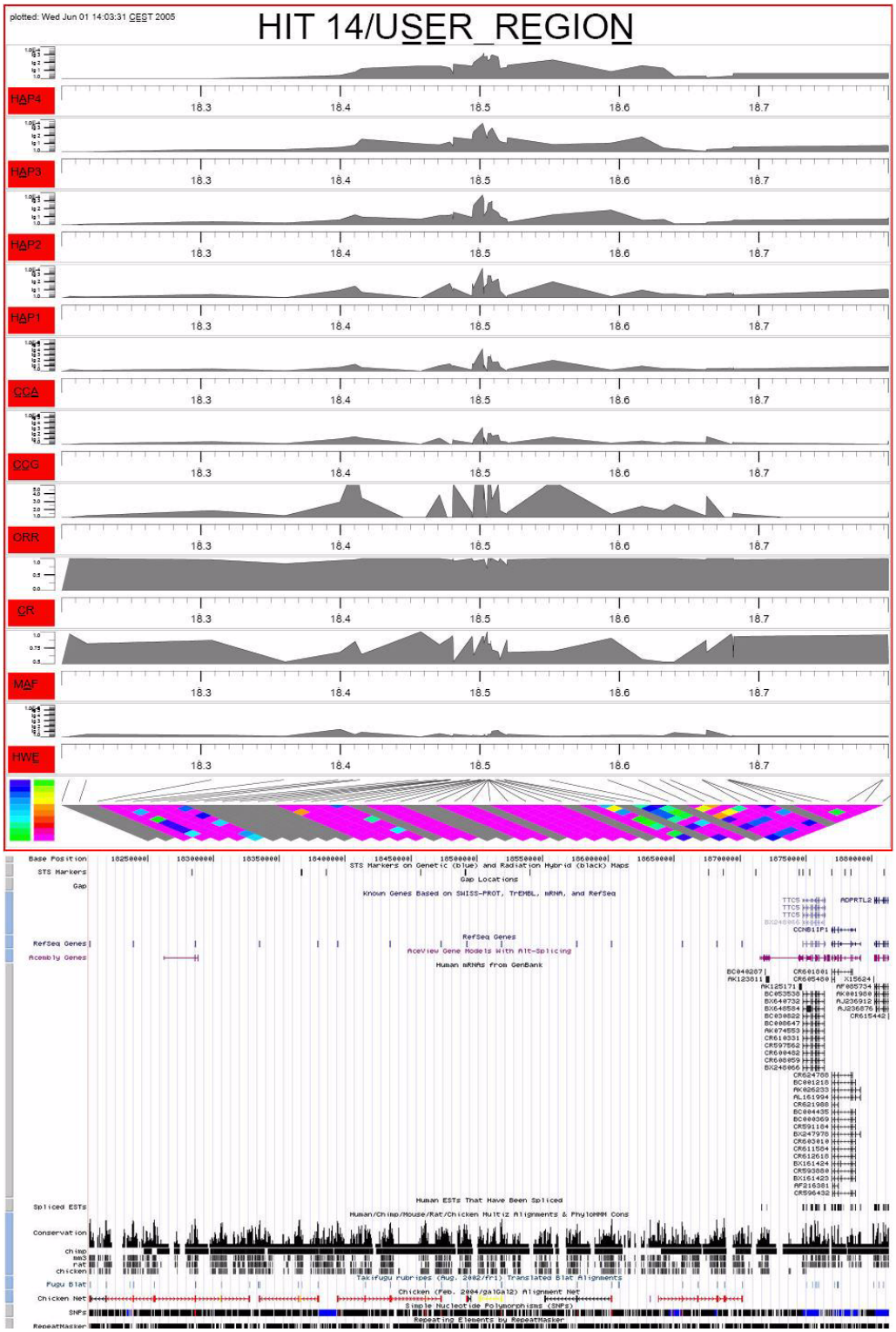

Fig. 9-4 Hit plot of a chromosomal region spanning 600 kbp. The corresponding standard UCSC genome browser view is automatically retrieved using the respective physical map coordinates (NCBI build 34). Pairwise LD is displayed using a color scheme for rho (Morton *et al.*, 2001), where rho = 1 is depicted by purple and rho = 0 by light gray diamonds. Lines above the LD plot relate the physical marker density to the equidistant LD plot. For every analysis type, a separate panel is plotted. HAP1-HAP4: WHAP haplotype analysis including randomization test for the specified sliding window. CCA: χ^2-statistic for alleles; CCG: χ^2-statistic for genotypes; ORR: odds ratio for carriership of less frequent allele; CR: call rate in cases and controls; MAF: major allele frequency; HWE: χ^2-statistic for Hardy-Weinberg equilibrium test. This analysis (n_{cases} = 20, $n_{controls}$ = 20) was generated from the sample data set that is provided with the software and does not relate to any actual phenotype.

9.4 Supplementary figures *in silico* protein analysis

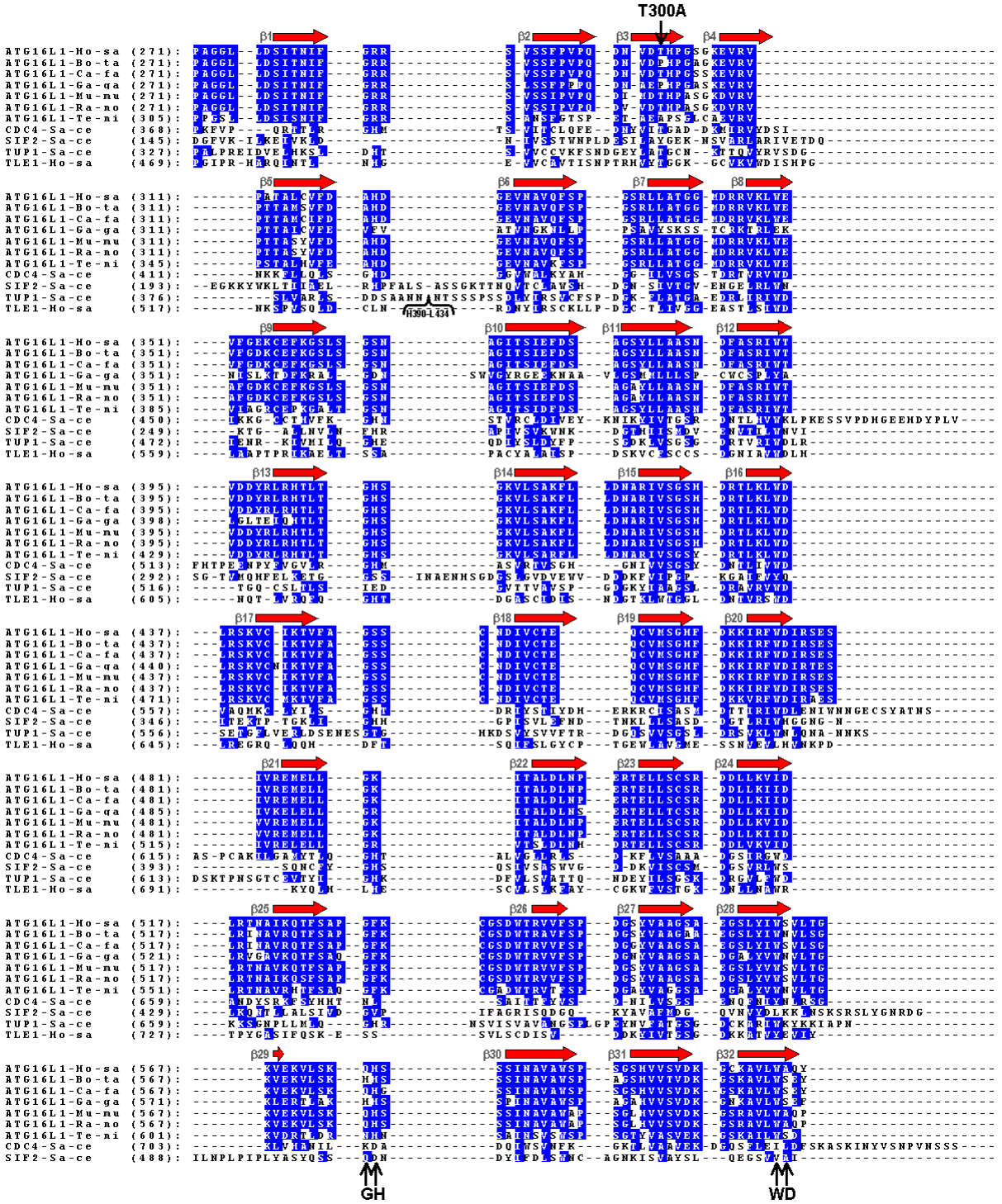

Fig. 9-5 Protein sequence alignment ATG16L1. Structure-based multiple sequence alignment of the WD-repeat domains of ATG16L1 homologs and the related proteins CDC4, SIF2, TUP1, and TLE1 with known 3D structures. Each alignment row contains a single WD repeat frequently characterized by GH and WD dipeptides (bottom annotation). CDC4 and SIF2 comprise eight WD repeats, whereas TUP1 and TLE1 contain seven WD repeats. The secondary structure depicted at the top of each alignment rows is taken from CDC4 and represents the β-strands characteristic of WD repeats. Physicochemically conserved amino acids are highlighted in blue boxes. Residue numbering in the alignment is based on complete protein sequences (table 2-20, page 79). The position of the sequence variant T300A of human ATG16L1 is marked (top annotation).

Fig. 9-6 Exemplary output of a secondary structure prediction for human ATG16L1. Here, the PSIPRED web server produced the depicted prediction.

9.5 Results of the fine mapping and replication of *NELL1* in panel H

#	SNP ID	position (build 35)	MAF	P_{TDT}	P_{TDT3}
1	rs7942430	20,594,635	0.39	0.01	0.02
2	rs3740872	20,605,132	0.42	0.14	0.07
3	rs2000959	20,615,255	0.32	0.30	0.85
4	rs10766710	20,624,679	0.50	0.03	0.54
5	rs7122910	20,636,147	0.29	0.68	0.18
6	rs1792969	20,646,062	0.13	0.28	0.08
7	rs1793004	20,655,505	0.26	0.02	0.03
8	rs870194	20,663,881	0.20	0.35	0.62
9	rs1792983	20,674,751	0.20	0.31	0.17
10	rs11025705	20,688,454	0.15	0.35	0.23
11	rs7952174	20,695,492	0.15	0.05	0.47
12	rs7116986	20,711,993	0.13	0.07	0.67
13	rs1519727	20,721,594	0.37	0.39	0.16
14	rs1554368	20,732,334	0.16	0.07	0.04
15	rs327036	20,742,093	0.19	0.20	0.14
16	rs1949523	20,750,221	0.26	0.02	0.18
17	rs1158547	20,771,723	0.29	0.04	0.21
18	rs4923055	20,779,494	0.27	0.04	0.25
19	rs1519735	20,791,833	0.31	0.04	0.43
20	rs435001	20,800,510	0.12	0.45	0.70
21	rs2680989	20,809,773	0.11	0.60	0.91
22	rs11025788	20,829,897	0.22	0.60	0.87
23	rs1429794	20,838,013	0.19	0.63	0.77
24	rs7109004	20,849,339	0.26	0.90	0.50
25	rs1549717	20,867,190	0.14	0.65	0.81
26	rs4296038	20,878,420	0.21	0.69	0.87
27	rs2082080	20,891,024	0.22	0.84	0.61
28	rs2293241	20,905,691	0.28	0.76	0.47
29	rs1880088	20,915,761	0.26	0.85	0.98
30	rs1880084	20,927,365	0.30	0.76	0.93
31	rs952696	20,935,188	0.20	0.68	0.62
32	rs10833417	20,945,048	0.27	0.91	0.56
33	rs10500884	20,955,364	0.13	1.00	0.92
34	rs1400373	20,956,134	0.13	1.00	0.89
35	rs10766756	20,966,312	0.13	0.75	0.84
36	rs10500886	20,976,742	0.25	0.48	0.30
37	rs12791452	20,986,789	0.15	0.40	0.78
38	rs11025878	20,998,335	0.26	1.00	0.50
39	rs4922728	21,009,790	0.19	1.00	0.20
40	rs7937542	21,017,772	0.46	0.49	0.55
41	rs10833444	21,029,035	0.27	0.57	0.67
42	rs9666195	21,038,467	0.36	0.53	0.54
43	rs1543153	21,047,961	0.25	0.56	0.45
44	rs7932820	21,056,960	0.40	0.82	0.31
45	rs12417046	21,066,312	0.19	0.73	0.30
46	rs2403652	21,076,903	0.48	0.42	0.46
47	rs7933049	21,087,224	0.22	0.69	0.74
48	rs4922753	21,096,666	0.26	0.80	0.53

Table 9-4 Fine mapping of NELL1 in panel H (French-Canadian population). SNPs that are significant in the TDT are highlighted in blue. MAF: minor allele frequency in control individuals. pTDT3: p-value obtained by a three-marker haplotype analysis using TDTPHASE.

#	SNP ID	position (build 35)	MAF	P_{TDT}	P_{TDT3}
49	rs10766767	21,108,451	0.26	0.80	0.58
50	rs6483748	21,137,843	0.29	0.16	0.54
51	rs4475918	21,147,706	0.39	0.37	0.62
52	rs4923403	21,156,380	0.29	0.16	0.07
53	rs1453983	21,166,668	0.26	0.11	0.10
54	rs1454003	21,175,735	0.19	0.41	0.59
55	rs1823843	21,185,111	0.02	0.22	0.63
56	rs1945327	21,190,821	0.09	0.09	0.73
57	rs1453988	21,201,279	0.31	1.00	0.91
58	rs1670638	21,207,776	0.31	0.95	0.12
59	rs1670640	21,218,026	0.48	1.00	0.80
60	rs1454008	21,230,114	0.42	0.61	0.07
61	rs1791822	21,240,131	0.46	0.91	0.50
62	rs1453990	21,249,027	0.47	0.78	0.96
63	rs716577	21,262,140	0.37	0.74	0.40
64	rs6483756	21,267,196	0.38	0.70	0.44
65	rs10833498	21,277,881	0.32	0.61	0.82
66	rs4335544	21,286,758	0.41	0.55	0.67
67	rs1349818	21,292,747	0.47	0.36	0.28
68	rs4399327	21,302,941	0.50	0.36	0.08
69	rs2187522	21,313,688	0.48	0.25	0.74
70	rs11026036	21,323,317	0.35	0.27	0.85
71	rs7126959	21,333,727	0.28	0.17	0.07
72	rs1945404	21,343,840	0.34	0.57	0.46
73	rs10833520	21,352,936	0.21	0.75	0.89
74	rs1945443	21,365,895	0.38	0.56	0.88
75	rs4539321	21,375,322	0.47	0.33	0.09
76	rs11026072	21,385,901	0.34	0.58	0.63
77	rs7943922	21,394,268	0.48	0.79	0.25
78	rs11026079	21,406,963	0.29	0.77	0.45
79	rs7110569	21,418,064	0.13	0.45	0.81
80	rs1945408	21,428,394	0.17	0.71	0.58
81	rs7945802	21,438,700	0.42	0.48	0.75
82	rs4343021	21,446,511	0.16	0.71	0.99
83	rs10766821	21,458,271	0.37	0.65	0.93
84	rs7116826	21,472,159	0.17	0.72	0.96
85	rs10219188	21,479,105	0.48	0.60	0.67
86	rs6483774	21,491,019	0.48	0.32	0.26
87	rs10766829	21,499,146	0.45	0.70	0.23
88	rs6483779	21,509,362	0.06	0.12	0.17
89	rs7927068	21,516,670	0.45	0.70	0.14
90	rs7926887	21,523,755	0.44	0.52	0.43
91	rs4319515	21,534,943	0.43	0.06	0.35
92	rs4320947	21,544,279	0.45	0.08	–
93	rs4922850	21,557,214	0.11	0.26	–

Table 9-4 Fine mapping of NELL1 in panel H (French-Canadian population). SNPs that are significant in the TDT are highlighted in blue. MAF: minor allele frequency in control individuals. pTDT3: p-value obtained by a three-marker haplotype analysis using TDTPHASE.

9.6 Results of fine mapping and replication of the 5p13.1 locus in panel H

#	SNP ID	position (build 35)	p_{TDT}	p_{TDT3}
1	rs13161247	40,093,886	0.03	0.29
2	rs12516769	40,106,140	0.06	0.29
3	rs4957239	40,112,759	0.05	0.23
4	rs17819573	40,130,693	0.06	0.09
5	rs978281	40,134,404	0.22	0.11
6	rs2548166	40,178,916	0.08	0.20
7	rs4455605	40,184,520	0.08	0.33
8	rs17820911	40,191,019	0.17	0.25
9	rs1559992	40,198,499	0.15	0.22
10	rs453254	40,200,091	0.14	0.21
11	rs17821331	40,223,211	0.85	0.41
12	rs353339	40,237,557	0.95	0.35
13	rs353325	40,246,690	0.28	0.24
14	rs350069	40,256,884	0.21	0.10
15	rs353367	40,264,205	0.43	0.38
16	rs183046	40,269,981	0.02	0.21
17	rs17224401	40,275,446	0.44	0.16
18	rs17224723	40,287,513	0.49	0.13
19	rs1445010	40,304,133	0.07	0.09
20	rs1445011	40,315,959	0.08	0.14
21	rs348620	40,322,578	0.34	0.32
22	rs348584	40,329,072	0.22	0.30
23	rs180900	40,342,058	0.19	0.35
24	rs4957127	40,351,767	0.16	0.74
25	rs16869833	40,383,331	0.53	0.86
26	rs7725523	40,407,980	0.43	0.38
27	rs895123	40,419,818	0.67	0.30
28	rs16869934	40,433,109	0.14	0.28
29	rs1025969	40,438,670	0.12	0.41
30	rs10512737	40,445,800	0.66	0.29
31	rs4957288	40,459,597	0.13	0.09
32	rs6451507	40,494,445	0.06	0.21
33	rs7725052	40,523,027	0.02	0.04
34	rs6889990	40,536,424	1.00	0.01
35	rs7718309	40,564,656	0.002	0.03
36	rs7713972	40,588,231	0.001	0.03
37	rs6451525	40,590,026	0.002	0.02
38	rs6864749	40,604,350	0.004	0.12
39	rs6451529	40,611,106	0.03	0.09
40	rs4409138	40,639,362	0.01	0.08
41	rs924967	40,650,879	0.01	0.26
42	rs17238368	40,662,946	0.92	0.21
43	rs7714305	40,677,005	0.86	0.14
44	rs7726237	40,724,309	0.90	0.79
45	rs4957343	40,730,827	0.72	0.90
46	rs249413	40,757,195	0.98	0.97
47	rs2329353	40,784,025	0.70	0.93

Table 9-5 Fine mapping of 5p13.1 in a French-Canadian population (panel H). SNPs that are significant in the TDT are highlighted in blue. MAF: minor allele frequency in control individuals. p_{TDT3}: p-value obtained by a three-marker haplotype analysis using TDTPHASE.

#	SNP ID	position (build 35)	p_{TDT}	p_{TDT3}
48	rs29743	40,791,037	0.87	0.50
49	rs257009	40,806,973	0.81	0.68
50	rs10074991	40,826,308	0.81	0.37
51	rs466108	40,832,503	0.76	0.62
52	rs1070446	40,874,985	0.69	0.51
53	rs16870407	40,888,805	0.86	0.24
54	rs16870425	40,901,620	0.44	0.45
55	rs10941526	40,909,835	0.43	0.19
56	rs1376178	40,944,658	0.50	0.62
57	rs9292793	40,952,442	0.26	0.54
58	rs16870528	40,972,426	0.90	0.77
59	rs2455309	40,979,012	0.60	0.27
60	rs1450657	40,985,533	0.49	0.42
61	rs1063499	40,991,318	0.42	0.28
62	rs1061429	41,017,446	0.21	0.10
63	rs2271707	41,033,753	0.08	0.13
64	rs10941529	41,041,548	0.77	0.05
65	rs1160833	41,046,944	0.60	0.02
66	rs325832	41,069,536	0.002	0.006
67	rs13167868	41,076,088	0.25	0.35
68	rs325872	41,089,680	0.06	0.18
69	rs17257858	41,096,124	0.93	0.11
70	rs1423391	41,106,597	0.47	0.16
71	rs325845	41,120,483	0.05	0.42
72	rs11958837	41,151,982	0.66	0.87

Table 9-5 Fine mapping of 5p13.1 in a French-Canadian population (panel H). SNPs that are significant in the TDT are highlighted in blue. MAF: minor allele frequency in control individuals. p_{TDT3}: p-value obtained by a three-marker haplotype analysis using TDTPHASE.

9.7 Pharmacogenomics and current state of IBD therapy - from bench to bedside

Pharmacogenomics, previously called pharmacogenetics, describes the use of genetics information to predict pharmacological efficacy and toxic effects (Pierik *et al.*, 2006). Examples include the evidence associating failure of steroid treatment with variation in expression of the multidrug resistance gene (MDR; Farrell *et al.*, 2000), and responsiveness or resistance to anti-TNF-α treatment with a polymorphism in the promoter region of TNF-α (Shanahan *et al.*, 2001). Mutations in the enzyme thiopurine methyltransferase (TPMT) change the therapeutic effect of purine analogue drugs, and thiopurine methyltransferase genotyping helps ensure optimum drug dosing and to identify those at increased risk of toxis effects (Dubinsky *et al.*, 2000). Additionally, genetic factors affect expression of subclinical markers of CD such as antibodies to *Saccharomyces cerevisiae* (ASCA), which have been proposed as diagnostic markers and as a way to subclassify IBD (Shanahan *et al.*, 2001; Shanahan *et al.*, 2000). ASCA is present in about 60% of CD patients and less than 5% in UC and non-IBD patients (Dubinsky *et al.*, 2006; Quinton *et al.*, 1998). It is generally agreed that the presence and level of immune responses do not change in a given CD patient (Desir *et al.*, 2004; Mow *et al.*, 2004; Dubinsky *et al.*, 2006).

From a practical point of view, predisposed individuals may benefit from genetic findings because of the following reasons:

1. A more precise diagnosis allows for a more suitable therapy of affected persons.
2. Prophylactic measures might be made to prevent outbreak of disease.
3. Because different “genetic compositions” can lead to the same disease phenotype, but result in distinct drug responses, individual therapies can be carried out.

This prospect of a “personalized medicine“ is also the main impulse of pharmaceutical companies and with increasing knowledge of the inflammatory cascade and disease susceptibility genes, many emerging anti-inflammatory strategies are predictable. Every stage in the inflammatory process can be interrupted with newly devised drugs including biological agents (Sands *et al.*, 2000), gene therapy (Shanahan *et al.*, 2001), or in some cases with drugs of low molecular weight. Only antagonism of TNF-α with the monoclonal antibody infliximab is presently approved for use in treatment of CD. Infliximab is a chimeric antibody composed of murine variable regions specific for human TNF-α, and human IgG1 constant regions. The murine protein component of infliximab results in adverse immunological responses in some patients, e.g. formation of human anti-chimeric antibodies, infusion reactions, and delayed hypersensitivity reactions (Sandborn *et al.*, 1999; Schaible *et al.*, 2000). A general risk with systemic TNF-α inhibition is the activation of latent tuberculosis or other

bacterial infections (Keane *et al.*, 2001). The precise mode of action of the flagship drug infliximab is uncertain but seems to include downregulation of the cytokine cascade, lymphocyte trafficking, and angiogenesis and induction of apoptosis in cells with membrane-bound TNF-α (Feldmann *et al.*, 2001).

Treatment of UC has been extensively reviewed by Hanauer *et al.* (2001), Sands *et al.* (2000), and Sandborn *et al. (*1996). In brief, 5-aminosalicylates (5-ASA) remain the mainstay for inducing remission of mild-to-moderate active UC, and for maintaining remission (Sutherland *et al.*, 2000). Glucocorticoids, a class of steroid hormones, reduce the production of pro-inflammatory cytokines (Sands *et al.*, 2000; Angeli *et al.*, 1999) and inhibit many leukocyte functions, such as adherence, chemotaxis, phagocytosis, and eicosanoid production (Goulding *et al.*, 1998; Goppelt-Struebe *et al.*, 1997). The long-term benefit of azathioprine and 6-mercaptopurine in maintaining remission and avoiding steroid treatments (Sandborn *et al.*, 1996; Ardizzone *et al.*, 1997; Adler *et al.*, 1990) far outweighs potential risks by far (Farrell *et al.*, 2000).

A further pursued strategy in IBD treatment is therapeutic manipulation of the intestinal flora. This is done by probiotics, which are live organisms, usually lactobacilli or bifidobacteria that are given as food supplements either alone or in combination (synbiotics) with certain polysaccharides (prebiotics) that might independently affect the enteric flora. Probiotic treatment seems to be efficacious in patients with UC and pouchitis, and trials in animals with CD have been encouraging (Campieri *et al.*, 1999). Possible mechanisms of probiotic action in inflammatory bowel diseases include production of antimicrobial factors, competitive interactions with pathogens, and signaling with the epithelium and mucosal immune system (Shanahan *et al.*, 2000).

9.8 Patient's questionnaire

Patient(in)

Fragebogen zur Studie bei chronisch entzündlichen Darmerkrankungen (CED)

Haben Sie bereits einmal Blut für eine genetische Studie an uns abgegeben? ❑ Ja ❑ Nein ❑ unbekannt

Haben Sie die Einwilligungserklärung unterschrieben?	❑ Ja ❑ Nein

Persönliche Daten

❑ männlich ❑ weiblich
Geburtsjahr:
Gewicht (kg):
Körpergröße (cm):

In verschiedenen Volksgruppen treten die CED in verschiedener Häufigkeit und Ausprägung auf

Ethnische Herkunft (z.B. deutsch, türkisch, italienisch, polnisch):
Woran sind Sie erkrankt? ❑ nicht erkrankt ❑ Morbus Crohn ❑ CED, aber zwischen M. Crohn/Colitis nicht entschieden ❑ Colitis ulcerosa ❑ Collagene Colitis ❑ Dickdarmkrebs ❑ andere Darmerkrankung, welche:_____
Welcher Ihrer Verwandten (falls zutreffend) hat Morbus Crohn oder Colitis ulcerosa? ❑ unbekannt ❑ keiner
Welcher Ihrer Verwandten hat Symptome für eine chronisch entzündliche Darmerkrankung, ist aber bisher nicht untersucht worden? ❑ unbekannt ❑ keiner
Wie viele Geschwister haben Sie? (Bitte mit Alter und Geschlecht angeben, wenn verstorben bitte Jahr und Lebensalter angeben) ❑ unbekannt
Wie viele Kinder haben Sie selbst? (Bitte mit Alter und Geschlecht angeben) ❑ unbekannt
Rauchen Sie oder haben geraucht? ❑ Ja ❑ Nein Wenn ja, wie viele Jahre: von.....bis..... Wie viele Zigaretten pro Tag: Rauchten Sie zum Zeitpunkt der Diagnose? ❑ Ja ❑ Nein
Trinken Sie Alkohol? ❑ Ja, regelmäßig jeden Tag ❑ Unregelmäßig ❑ Nie

Seite 1 von 6

Fig. 9-7 Questionnaire page 1/6.

In einer englischen Studie wurde ein Zusammenhang zwischen dem Stadtleben und dem Risiko an Morbus Crohn zu erkranken hergestellt:

Wo sind Sie aufgewachsen ?	Wo leben Sie heute?
❑ Dorf	❑ Dorf
❑ Kleinstadt	❑ Kleinstadt
❑ Großstadt	❑ Großstadt
❑ unbekannt	

In einem Haushalt mit wie vielen Personen wuchsen Sie auf?_____
In einem Haushalt mit wie vielen Personen leben Sie heute? _____

Welcher Sanitärkomfort herrschte in Ihrer Kindheit?

Fließendes Wasser	❑ Ja	❑ Nein	❑ unbekannt
Fließendes Warmwasser	❑ Ja	❑ Nein	❑ unbekannt
Wasserklosett	❑ Ja	❑ Nein	❑ unbekannt
Zentralheizung	❑ Ja	❑ Nein	❑ unbekannt

Angaben zur allgemeinen Krankengeschichte:

Besteht bei Ihnen eine Lungenerkrankung? ❑ Ja ❑ Nein ❑ weiß nicht
wenn ja, welche: ❑ COPD (chronische Bronchitis)
❑ Sarkoidose (Morbus Boeck)
❑ Asthma bronchiale
❑ Andere: _____

Haben Sie Husten oder litten Sie innerhalb der letzten 5 Jahre an Husten?
❑ Nein ❑ unbekannt

❑ Ja, wegen ❑ Schnupfen, Nasennebenhöhlenentzündung ❑ akutem Infekt
❑ Sodbrennen/Refluxkrankheit ❑ Herzschwäche
❑ Pneumonie (Lungenentzündung) ❑ Tumor
❑ unbekannt

Dauer des Hustens? ❑ tgl. >3 zusammenhängende Monate innerhalb 2 Jahren
❑ > 2 Wochen
❑ < 2 Wochen
❑ unbekannt

Besteht Auswurf? ❑ Ja, ❑glasig-weißlich
❑ gelblich/grünlich
❑ bräunlich/rötlich
Seit wann?_____ Wie oft am Tag?_____
❑ Nein
❑ unbekannt

Es scheint einen Zusammenhang zwischen chron. Zahnfleischentzündung („Parodontitis“) und CED zu geben.

Ist bei Ihnen eine Zahnfleischerkrankung bekannt?	❑ Ja	❑ Nein	❑ weiß nicht
Haben Sie häufiger Zahnfleischbluten?	❑ Ja	❑ Nein	❑ weiß nicht
Haben Sie lockere, gewanderte oder gekippte Zähne?	❑ Ja	❑ Nein	❑ weiß nicht
Mussten Ihnen schon lockere Zähne gezogen werden?	❑ Ja	❑ Nein	❑ weiß nicht

Eine Verbindung zwischen CED und entzündl. Hauterkrankungen wurde wiederholt vorgeschlagen.

Ist bei Ihnen eine Hauterkrankung bekannt? ❑ Ja ❑ Nein ❑ weiß nicht

Wenn ja, welche:	Neurodermitis (atopisches Ekzem)	❑ Ja	❑ Nein	❑ weiß nicht
	Ekzem	❑ Ja	❑ Nein	❑ weiß nicht
	Psoriasis (Schuppenflechte)	❑ Ja	❑ Nein	❑ weiß nicht
	❑ andere: _____			

Seite 2 von 6

Fig. 9-8 Questionnaire page 2/6.

Haben/ hatten Sie eines der folgenden Hautprobleme?
- ❑ trockene Haut
- ❑ gestörtes Nagelwachstum
- ❑ Kopfschuppen
- ❑ rote schuppende Stellen (z.B. an den Ellenbogen oder am Haaransatz)
- ❑ keines

Leiden Sie häufiger unter Juckreiz?
- ❑ Ja,
  - ❑ gelegentlich
  - ❑ vor allem bei starkem Schwitzen
  - ❑ unangenehmes Gefühl bei direktem Kontakt mit Wolle
  - ❑ juckende Bläschen an Handflächen/ Fußsohlen
- ❑ Nein

Hatten Sie Milchschorf als Säugling? ❑ Ja ❑ Nein ❑ weiß nicht

Leiden Sie unter Allergien? ❑ Ja ❑ Nein ❑ weiß nicht
Wenn ja, welche: _____

Besteht bei Ihnen eine rheumatische Organerkrankung? ❑ Ja ❑ Nein ❑ weiß nicht
Wenn ja, welche: _____
Seit wann:

Leiden Sie unter anderen Erkrankungen wie
z.B. Zuckerkrankheit, Bluthochdruck u.a. ❑ Ja ❑ Nein ❑ weiß nicht
Wenn ja, welche: _____
Seit wann:_____
Nehmen Sie Tabletten gegen Diabetes ❑ Ja ❑ Nein ❑ weiß nicht
Wenn ja, welche _____
Spritzen Sie Insulin? ❑ Ja ❑ Nein ❑ weiß nicht
Wenn ja, wieviel (ca.)? _____

Wurden Sie gegen Tuberkulose (Tbc) geimpft? ❑ Ja ❑ Nein ❑ weiß nicht

Wurden Sie gegen Masern geimpft? ❑ Ja ❑ Nein ❑ weiß nicht

Waren Sie als Kind an Masern erkrankt? ❑ Ja ❑ Nein ❑ weiß nicht

Nahmen oder nehmen Sie Medikamente ein (nur Medikamente, die nicht für die Behandlung von CED verwendet werden) ? ❑ Ja ❑ Nein
Wenn ja, welche und wie lange:

Die folgenden Fragen bitte beantworten, wenn Sie selbst an einer chronisch entzündlichen Darmerkrankung erkrankt sind:

Welche Darmteile sind bei Ihnen betroffen? ❑ unbekannt
- ❑ Dünndarm
- ❑ Übergang vom Dünndarm zum Dickdarm
- ❑ Teile des Dickdarms - welche: _____
- ❑ gesamter Dickdarm (Pancolitis):
- ❑ Enddarm

genauere Angaben:

Seite 3 von 6

Fig. 9-9 Questionnaire page 3/6.

Haben sich während Ihrer Erkrankung Fisteln entwickelt? ❑ Ja ❑ Nein ❑ weiß nicht Wenn ja, welche: Enddarm (Analfisteln) ❑ Ja ❑ Nein Fisteln zwischen Darmschlingen ❑ Ja ❑ Nein Fisteln vom Darm zur Haut ❑ Ja ❑ Nein Fisteln zu anderen Organen ❑ Ja ❑ Nein
Haben sich bei Ihnen im Verlauf der Erkrankung Stenosen (Darmverengungen) ausgebildet? ❑ Ja ❑ Nein ❑ weiß nicht
Haben/Hatten Sie folgende Beschwerden? Gelenkschmerzen ❑ Ja ❑ Nein Hautentzündungen ❑ Ja ❑ Nein Augenentzündungen ❑ Ja ❑ Nein Gallengangsentzündungen ❑ Ja ❑ Nein Analfissur ❑ Ja ❑ Nein Abszesse ❑ Ja ❑ Nein
Seit welchem Kalenderjahr ist die Diagnose klar?
Seit welchem Kalenderjahr hatten Sie erste Beschwerden?
Hat sich die Diagnose während der Erkrankung schon einmal verändert? ❑ Ja ❑ Nein ❑ weiß nicht Wenn ja, von der Diagnose Morbus Crohn zu Colitis ulcerosa ❑ Ja ❑ Nein Von der Diagnose Colitis ulcerosa zu Morbus Crohn ❑ Ja ❑ Nein
Wie wurde Ihre Erkrankung bestätigt? Feingewebeschnitt ❑ Ja ❑ Nein ❑ Weiß nicht Darmspiegelung ❑ Ja ❑ Nein ❑ Weiß nicht Darmröntgen ❑ Ja ❑ Nein ❑ Weiß nicht
Sind aufgrund Ihrer Erkrankung Operationen notwendig gewesen? ❑ Ja ❑ Nein Wenn ja, welche und in welchem Jahr?
Haben Sie zur Zeit einen künstlichen Darmausgang? ❑ Ja ❑ Nein Hatten Sie einmal einen künstlichen Darmausgang? ❑ Ja ❑ Nein
Wie viele stationäre Krankenhausaufenthalte sind bei Ihnen notwendig gewesen? ❑ Nie ❑ Weniger als alle 2 Jahre ❑ ca. 1x pro Jahr ❑ 1-3 x pro Jahr ❑ Mehr als 3 x pro Jahr
Wie lange insgesamt im Krankenhaus? ❑ weniger als 1 Monat ❑ 1-3 Monate ❑ 3-6 Monate ❑ Mehr als 6 Monate

Seite 4 von 6

Fig. 9-10 Questionnaire page 4/6.

Wie lange haben Sie insgesamt Kortison > 10mg pro Tag nehmen müssen?
❑ Mehr als ein Jahr
❑ bis zu einem Jahr
❑ 3-6 Monate
❑ 1-3 Monate
❑ Weniger als 1 Monat
❑ Noch nie

Wie viele Schübe haben Sie in etwa pro Jahr?
❑ Weniger als alle 2 Jahre
❑ ca. 1x pro Jahr
❑ 1-3 x pro Jahr
❑ Mehr als 3 x pro Jahr

Wann war der letzte Schub?

Sind Sie wegen Ihrer CED vorzeitig berentet worden? ❑ Ja ❑ Nein
Wenn ja, in welchem Lebensalter?

Welche Medikamente haben Sie bisher für Ihre chronisch-entzündliche Darmerkrankung erhalten?

	Früher	Aktuell
Mesalazin (Salofalk, Claversal, Pentasa)	❑	❑
Sulfasalazin (Azulfidine, Colo-Pleon)	❑	❑
Cortison	❑	❑
Budesonid (Entocort, Budenofalk)	❑	❑
Azathioprin (Imurek, Zytrim, Azafalk, Azamedac, Colinsan)	❑	❑
6-Mercaptopurin (Puri-Nethol)	❑	❑
Infliximab (Remicade)	❑	❑
Methotrexat (MTX)	❑	❑
Cyclosporin	❑	❑
Tacrolimus	❑	❑
Mutaflor	❑	❑
H15	❑	❑

Studienmedikamente (bitte genauere Angaben):_____
_____

Was würden Sie in der Betreuung von Patienten mit chronisch-entzündlichen Darmerkrankungen verändern?

Seite 5 von 6

Fig. 9-11 Questionnaire page 5/6.

Wären Sie bereit, an einer Studie zu den sozialen und psychischen Faktoren der chronisch entzündlichen Darmerkrankungen teilzunehmen?
❑ Ja ❑ Nein

Nach Eingang der Proben wird jede Verbindung zu Ihrer Adresse und Ihrem Namen aus unseren Unterlagen gelöscht. Aufgrund vieler Rückfragen von Patienten haben wir eine separate Adressenkartei für die Zusendung allgemeiner Informationen zum Fortgang des Projekts.

Ich möchte weiter über die allgemeinen Ergebnisse der Studie (die in ca. 3 Jahren zu erwarten sind) auf dem Laufenden gehalten werden und möchte daher, dass Sie meinen Namen und meine Adresse aufbewahren, um mich über die neuesten Forschungsergebnisse zu informieren. Diese Daten sollen bitte getrennt von den persönlichen Daten auf dem Fragebogen aufbewahrt werden.
❑ Ja ❑ Nein

Ich habe kein Interesse an Informationen über die Studienergebnisse. Bitte löschen Sie meinen Namen und meine Adresse nach Eingang sofort.
❑ Ja ❑ Nein

Vielen Dank für Ihre Mitwirkung!
Bei Rückfragen melden Sie sich bitte gerne unter Telefon 0431-597 1084 (Fr. Timm)

Seite 6 von 6

Fig. 9-12 Questionnaire page 6/6.

9.9 Patient's written consent

Patientenaufklärung / Einwilligungserklärung

Bitte eintragen:

Name:_____

Adresse:_____

_____

Im Rahmen eines Forschungsprojektes zur Entstehung chronisch entzündlicher Erkrankungen (insbesondere Morbus Crohn und Colitis ulcerosa) bitten wir Sie um Zustimmung

☒ **20 ml peripher venöses Blut**

entnehmen zu dürfen. Wir wollen DNA aus Ihrem Blut gewinnen, um die genetischen Ursachen für eine Veranlagung zu chronisch entzündlichen Erkrankungen zu definieren. Die dafür verantwortlichen Gene können sich im gesamten Chromosomenbereich befinden, andere als diese Entzündungsgene werden jedoch nicht untersucht. Wir bitten Sie, Ihre Zustimmung durch Ihre Unterschrift zu bestätigen. Es werden Ihnen keine Nachteile entstehen, falls Sie einer Probenentnahme nicht zustimmen sollten.

Im folgenden möchten wir Ihnen die Vorteile und Risiken einer solchen Probenentnahme darstellen:

Zu erwartende Vorteile für Ihre eigene Behandlung:
Keine unmittelbaren Vorteile. Es werden sich aus den entnommenen Proben keine Rückschlüsse ableiten lassen, die unmittelbaren Einfluß auf Ihre persönliche Behandlung haben werden. Auch aufgrund der Anonymisierung können keine persönlichen Informationen zur Verfügung gestellt werden. In Zukunft erhoffen wir uns jedoch erhebliche Vorteile für alle Patienten Ihrer Krankheitsgruppe, da wir uns durch die weitere Aufklärung immunologischer Mechanismen Fortschritte in der Diagnostik bzw. Therapie versprechen. Diese Erforschung von Krankheitsursachen und -mechanismen kann hier in Kiel erfolgen oder – im Rahmen von Forschungskooperationen – auch an anderen Orten.

Risiken der Blutentnahme:
Wie zu einer Routineblutentnahme werden Ihnen unter sterilen Bedingungen 20 ml Blut aus einer peripheren Vene entnommen. Die Risiken sind identisch mit denen einer Routineblutentnahme: Lokale Infektion ("bakterielle Entzündung, Vereiterung") und Fehlpunktion einer Schlagader. Beide Risiken sind bei sachgemäßer Durchführung extrem selten.

Fig. 9-13 Written consent page 1/3.

Einverständniserklärung 2

Speicherung von Daten:
Ihre persönlichen Daten getrennt von den genetischen Daten gespeichert – sie sind nur bis zur Verarbeitung der Blutprobe verfügbar und werden dann gelöscht

Die gewonnenen Proben werden anonymisiert. Die mit der Probe verbundenen Informationen zur Krankheit (wie das Krankheitsstadium) werden von den persönlichen Informationen unwiderruflich getrennt. Diese Daten werden auf einem Kennwortgeschützten Computer in einer separat abgesicherten Datenbank gespeichert. Eine Weitergabe an unberechtigte Dritte (insbesondere Arbeitgeber, Versicherungen) ist damit vollständig ausgeschlossen. Die Weitergabe von Proben und nichtpersonenbezogenen Informationen an wissenschaftliche Kooperationspartner, die zu einer schnelleren Beantwortung der wissenschaftlichen Fragestellungen führen können, erfolgt ausschließlich in anonymisierter Form.
Für diese spezielle Situation entbinde ich hiermit den behandelnden Arzt von der Schweigepflicht gegenüber Mitarbeitern des Forschungsvorhabens. Bei Beendigung der Forschungsaktivitäten werden die Proben und die dazu gehörenden Daten vernichtet.

Wir würden uns sehr freuen, falls Sie sich entschließen würden, zu diesem für alle Kranken mit Morbus Crohn/Colitis ulcerosa so wichtigen Forschungsprojekt durch Ihre Einwilligung beizutragen.

Kommerzielle Nutzung/ Patent
Es kann sein, dass sich im Rahmen zukünftiger Forschung eine kommerzielle Nutzung der Forschungsergebnisse, basierend auf Ihrem individuellen biologischen oder genetischen Material, oder eine kommerzielle Nutzung des biologischen oder genetischen Materials selbst ergeben wird. Für diesen Fall besteht kein persönlicher Anspruch, dies gilt ebenfalls für Patentrechte.

Benachrichtigung über die Ergebnisse des Gesamtvorhabens
Da alle Proben anonymisiert sind, kann Ihnen Ihr persönliches Ergebnis auch auf Nachfrage nicht mitgeteilt werden. Auf Wunsch teilen wir Ihnen gerne die Gesamtergebnisse der Forschungslabor-Untersuchungen mit. In diesem Fall würden wir Ihren Namen, Vornamen und Geburtsdatum separat ohne Verbindung zur Blutprobe erfassen um Ihnen einen regelmäßigen Informationsbrief zusenden zu können.

Fig. 9-14 Written consent page 2/3.

Einverständniserklärung 3

Ich willige in die Entnahme von

☒ Blut

☐ ein

☐ nicht ein.

☐ Ich habe Zeit und Gelegenheit zur Entscheidung gehabt. Zusätzliche Fragen, falls vorhanden, sind mir ausreichend beantwortet worden.

☐ Ich bin darüber belehrt worden, daß ich die Zustimmung jederzeit verweigern oder widerrufen kann, ohne daß mir Nachteile in der weiteren Diagnostik oder Therapie entstehen.
Dies ist nur **vor** der Anonymisierung der Proben möglich.

☐ Ich möchte auf die Adressenliste für die Aussendung eines Informationsbriefs zu den Gesamtergebnissen der Untersuchung gesetzt werden.

Ort .....

Datum .....

Unterschrift: .....

Fig. 9-15 Written consent page 3/3.

10 Curriculum vitae

Personal Data

Name. . . . . Andre Wilhelm Franke

Address. . . . . Damperhofstr. 6
D-24103 Kiel
Germany

Date of birth . . . . . 16 Oct 1978

Place of birth. . . . . El Paso, Texas, U.S.A.

Citizenship . . . . . German, US-American

Marital status . . . . . Married

Education

1985 - 1986. . . . . Volksschule, Neufahrn (Bayern)

1986 - 1989. . . . . Grundschule, Dieburg (Hessen)

1989 - 1991. . . . . Max-Planck-Gymnasium, Dieburg (Hessen)

1991 - 1992. . . . . Gymnasium, Nordenham (Niedersachsen)

1992 - 1994. . . . . Lothar-Meyer-Gymnasium, Varel (Niedersachsen)

1994 - 1995. . . . . Gymnasium Jungmannschule, Eckernförde (Schleswig-Holstein)

1995 - 1996. . . . . Hillcrest High School, Idaho Falls, U.S.A.

1996 - 1998. . . . . Gymnasium Jungmannschule, Eckernförde
Major in Chemistry und English

Study

10/1998 - 11/2002. . . . . Study of Biology at the Christian-Albrechts University Kiel:
Major in Cell Biology and Computer Sciences

12/2002 - 11/2003. . . . . Diploma thesis at the laboratory of Prof. Dr. Dr. h.c. Thomas C.G. Bosch (Cell and Developmental Biology)
“Development of a high-throughput screening for identification of developmentally regulated genes in *Hydra*”

02/2004 - Today . . . . . PhD work at the Institute for Clinical Molecular Biology of Prof. Dr. S. Schreiber
“A systematic genome-wide association analysis for inflammatory bowel diseases (IBD)”
Scholarship from the "MFG Mucosaimmunologie gemeinnützige Forschungsgesellschaft mbH"

Degrees

06/1997. . . . . Fachhochschulreife (1.8)

06/1998. . . . . Abitur (1.9)

10/2000. . . . . Pre-Diploma in Biology (1.0)

11/2003. . . . . Diploma in Biology (1.0)

Publications

Hampe J[1)], **Franke A**[1)], Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De La Vega F, Briggs J, Günther S, Prescott N, Onnie C, Foelsch UR, Lengauer T, Platzer M, Mathew C, Krawczak M, Schreiber S. A genome-wide association scan using non-synomous SNPs identifies a susceptibility variant for Crohn's disease in the autophagy-related 16-like (ATG16L1) gene. *Nature Genetics 2006; in press.*

Franke A, Ruether A, Wedemeyer N, Karlsen TH, Nebel A, Schreiber S. No association between the functional *CARD4* insertion/deletion polymorphism and inflammatory bowel diseases in the German population. *Gut 2006*; 55(11).

Schafmayer S, Tepel J, **Franke A**, Buch S, Lieb S, Seeger M, Lammert F, Kremer B, Foelsch UR, Faendrich F, Schreiber S, and Hampe J. Investigation of the Lith1 Candidate Genes ABCB11 and LXRA in Human Gallstone Disease; *Hepatology 2006*; 44(3):650-657.

Franke A, Wollstein A, Teuber M, Wittig M, Lu T, Hoffmann K, Nurnberg P, Krawczak M, Schreiber S, Hampe J. GENOMIZER: an integrated analysis system for genome-wide association data. *Human Mutation 2006*; 27:583-8.

Valentonyte R[1)], Hampe J[1)], Huse K, Rosenstiel P, Albrecht M, Stenzel A, Nagy M, Gaede KI, **Franke A**, Haesler R, Koch A, Lengauer T, Seegert D, Reiling N, Ehlers S, Schwinger E, Platzer M, Krawczak M, Muller-Quernheim J, Schurmann M, Schreiber S. Sarcoidosis is associated with a truncating splice site mutation in BTNL2. *Nature Genetics 2005*; 37:357-64.

Augustin R[1)], **Franke A**[1)], Khalturin K[1)], Kiko R[1)], Siebert S[1)], Hemmrich G, Bosch TC. Dickkopf related genes are components of the positional value gradient in Hydra. *Developmental Biology 2006*; 296: 62-70

Presentations

07/2006 . . . . . NGFN Environmental Network Meeting, Berlin
Progress report Crohn's disease (Subproject NUW-S23T06)

06/2006 . . . . . Applied Biosystems (AB) User Meeting, Kiel and Berlin
"Whole genome cSNP association study in Crohn's disease using SNPlex™ genotyping system"

05/2006 . . . . . European Society of Human Genetics (ESHG), Amsterdam,
AB Satellite Meeting
"The coding SNP approach. Straigh to the disease gene?"

12/2005 . . . . . AB seminar in Foster City, CA, USA
"Whole genome association studies in Crohn's disease.
SNP array vs. cSNP using SNPlex™"

04/2005 . . . . . AB SNPlex™ User Meeting, Paris
"Establishment of a high-throughput environment for SNPlex™ Assay genotyping"

1). Equal contribution of authors to the manuscript.

Congress Posters and Abstracts

10/2006 . . . . . ASHG, New Orleans, LA, USA
Franke A, Hampe J, Rosenstiel P, Till A, Huse K, Albrecht M, Mayr G, De La Vega F, Briggs J, Wenz M, Günther S, Gilbert D, Prescott N, Lengauer T, Mathew C, Krawczak M, Schreiber S. A systematic genome wide association study of 19,779 coding SNPs with putative function identifies a novel susceptibility gene for Crohn's disease.

10/2005 . . . . . ASHG, Salt Lake City, UT, USA
Franke A, Wollstein A, Teuber M, Wittig M, Lu T, Hoffmann K, Nurnberg P, Krawczak M, Schreiber S, Hampe J. GENOMIZER: an integrated analysis system for genome-wide association data.

Hampe J, Kwiatkowski R, **Franke A**, Becker C, Krawczak M, Mueller-Quernheim J, Schürmann M, Nuernberg P, Schreiber S. Identification of common disease genes for granulomatous diseases through a two-stage 100k genome-wide association study in sarcoidosis and Crohn's disease

02/2005 . . . . . Labautomation 2005, San Jose, CA, USA
Jung C, Porter G, T.-Clemons L, **Franke A**, Hampe J, Schreiber S. ABI SNPlex™ technology for HT genotyping on Tecan Freedom EVO® robotic platform

11/2004 . . . . . NGFN Meeting, Berlin
Hampe J, **Franke A**, Schreiber S. Genome-wide and pathway-specific coding SNP-sets available for SNPlex™ genotyping within NGFN2

Attendance at Courses and Congresses

03 Oct – 07 Oct 2006 . . . . . Wellcome Trust, Cambridge
Design and analysis of genetic-based association studies
Instructors: Cordell H, Morris A, Marchini J

18 Jan 2006 . . . . . Basic Instrument Training 3730 XL, Applied Biosystems

03 Jun – 04 Jun 2005 . . . . . NGFN Symposium, Kiel
Inflammatory diseases of barrier organs

08 May – 11 May 2005 . . . . CNG, Paris
Paris workshop on molecular and statistical genomic epidemiology

25 Oct – 29 Oct 2004 . . . . . Tecan, Crailsheim
FACTS and Gemini robotics training

02 Aug – 06 Aug 2004 . . . . . GSF, München
Advanced gene mapping/ linkage course
Instructors: Leal S, Clayton D, Cordell H, Hoh J, Almasy L, Ott J, Wienker T

02 Jun – 04 Jun 2004 . . . . . NGFN meeting, Berlin

11 Declaration

Declaration

Apart from the advice of my supervisors, this thesis is completely the result of my own work. No part of it has been submitted to any other board for another qualification. Most of the results have been or are about to be published (see “Publications” on page 201).

Erklärung

Hiermit erkläre ich, daß diese Dissertation, abgesehen von der Beratung durch meine akademischen Lehrer, nach Inhalt und Form meine eigene Arbeit ist. Sie hat weder im Ganzen noch zum Teil an anderer Stelle im Rahmen eines Promotionsverfahrens vorgelegen. Die meisten Ergebnisse dieser Arbeit wurden zur Veröffentlichung eingereicht (siehe “Publications” auf Seite 201).

Kiel, ... (Andre Franke)

12 Acknowledgement

This study would not have been possible without the contribution of many competent and kind persons. Therefore, I am very grateful to...

Prof. Dr. Stefan Schreiber ..... for providing me with excellent working conditions, large resources of patient material, and an attractive research project. I am especially grateful for his continuous encouragement and his trust in letting me deal with the described large and expensive experiments.

Prof. Dr. Dr. h.c. Thomas C.G. Bosch .... for his support and supervision throughout the years and the possibilities to present the progress of my work during the laboratory meetings of his workgroup.

PD Dr. Jochen Hampe ..... for his supervision during the last three years and his introduction into epidemiology and human genetics. His scientific work and project steering provided the basis of this thesis. Furthermore, his time-intensive investments into the laboratory infrastructure and the LIMS are greatfully acknowledged.

Prof. Dr. Ulrich Fölsch ..... for the opportunity to work in his clinic.

Ilona Urbach, Jan Sebastian Tammen, Catharina Fürstenau, Katja Cloppenborg-Schmidt
for the continuous high-quality work of preparing DNA samples.

Tanja Henke, Susan Ehlers, Catharina von der Lancken, Ina Elena Baumgartner, Meike Davids
for ensuring a high-quality genotyping plate production.

Birthe Fedders, Tanja Wesse, Lena Bossen
for the perfect technical assistance and the generation of millions of genotypes.

Anita Dietsch, Melanie Friskovec ..... for carrying out mutation detections and continuous maintenance of the sequencers.

Markus Teuber & Carl Manaster ..... for their introduction into the LIMS and the many helpful tools they programmed.

Michael Wittig ..... for his tools that facilitated the handling of genome-wide data sets.

Marcus Will..... for his IT expertise and maintenance of the server and the network.

Rainer Vogler & Birgitt Timm..... for their help with the database.

Dr. Philip Rosenstiel, Dr. Andreas Till, Sven Künzel, and Tanja Kaacksteen
for their valuable contributions that added to the functional understanding of the found susceptibility regions.

Dr. Rober Häsler & Dorina Ölsner ..... for their expert help with real-time gene expression studies.

Dr. Klaus Huse ..... for his cloning and sequencing efforts that significantly contributed to this project. Furthermore, I would like to thank him for helpful discussions and sharing his wealth of experience.

Prof. Dr. Peter Nürnberg, Christian Becker
for performing the genotyping of the 100k Affymetrix arrays.

Mario Albrecht, Gabriele Mayr ..... for their extremely competent help with protein modeling.

Dr. Almut Nebel..... for carefully proof-reading this thesis and for her intellectual input.

Dr. Tom Hemming Karlsen ..... for his everlasting support with articles and his examplary manner in terms of working and political decisions. It was and still is more than a pleasure to work together with him.

Dr. Andrea Töppel..... for her help with chemistry-related problems and the excellent support for AB products.

Dr. Simone Günther, Dr. Michael Wenz, Jason Briggs, Francisco de La Vega, Dennis Gilbert
for the fruitful cSNP collaboration and weekly telephone conferences.

Genizon BioSciences ..... for sharing their association data, genotyping platform, and Crohn's disease samples during this project.

Prof. Dr. Christopher Mathew, Dr. Natalie J. Prescott
for putting their UK CD sample collection at our disposal.

Dr. Ruta Kwiatkowski, Friederike Flachsbart, Annegret Fischer, Abdou al Sharawy, Weiyue Zheng
for the pleasant and warm atmosphere, though living together in a physically restricted office under occasional time pressure. Ruta Kwiatkowski is especially acknowledged for the introduction into genetic data analyses and SNP selection.

my wife Isabell Franke..... for her love and companionship and her understanding for the long working hours of her husband. I would also like to thank her for the strong support and social backing during stressful phases.

my parents Elke & Lothar Franke ..... for their constant motivation, plus financial and IT support since the beginning of my studies. They were always there when I needed them and they never complained about how infrequently I visited them.

The expert help of all clinicians and nurses, especially the team of the Kiel ambulance in the 1st medical department , and the cooperation of patients, who agreed on the use of their biopsies and blood samples for academic research, is highly appreciated. **Drs. Susanna Nikolaus, Christian Sina, Jochen Hampe, Philip Rosenstiel, and Tanja Kühbacher** deserve special credit for careful evaluation and characterization of patients.

Finally, I thank the entire staff of the Institute of Clinical Molecular Biology for the nice atmosphere and their support during the course of my PhD thesis. In addition, I excuse myself to the persons, who I forgot to mention in this acknowledgement. They should know that these last words are generally written under highly stressfuly circumstances.

This work was funded by the National Genome Research Network (NGFN), the Deutsche Forschungsgesellschaft (DFG), and by a scholarship of the Mucosaimmunologie gemeinnützige Forschungsgesellschaft mbH (MFG; Hamburg, Germany).