LITHUANIAN UNIVERSITY OF HEALTH SCIENCES

Jurgita Šventoraitytė

GENETIC CHARACTERISTICS OF LITHUANIAN AND LATVIAN PATIENTS WITH INFLAMMATORY BOWEL DISEASE

Doctoral Dissertation

Biomedical Sciences, Biology (01 B)

Kaunas, 2011

Dissertation has been prepared at Lithuanian University of Health Sciences during the period of 2006–2010.

Scientific Supervisor

Prof. Dr. **Neringa Paužienė** (Lithuanian University of Health Sciences, Biomedical Sciences, Biology – 01 B)

Consultants:

Prof. Habil. Dr. **Limas Kupčinskas** (Lithuanian University of Health Sciences, Biomedical Sciences, Medicine – 07 B)

Prof. Dr. rer. nat. **Andre Franke** (Christian-Albrechts University in Kiel, Biomedical Sciences, Biology – 01 B)

CONTENT

ABBREVIATIONS

AoD	assay-on-Demand
ARPC	actin-related protein 2/3 complex
ASO	allele-specific oligo
ATG16L1	autophagy-related protein 16-like 1
BSN	bassoon (presynaptic cytomatrix protein)
BTNL2	butyrophilin-like 2
C13orf31	chromosome 13 open reading frame 31
CARD	caspase recruitment domain
CCR6	CC chemokine receptor 6
CD	Crohn's disease
CI	confidence interval
Csk	C-terminal src kinase
ECM1	extracellular matrix protein 1
gDNA	genomic deoxyribonucleic acid
GWAS	genome wide association studies
HLA	human leukocyte antigen
HWE	Hardy Weinberg equilibrium
IBD	inflammatory bowel disease
ICOSLG	inducible T-cell co-stimulator ligand
IFN-γ	interferon-γ
IL	interleukin
IL23R	interleukin 23 receptor
IRGM	immunity-related GTPase family, M
JAK2	Janus kinase 2
LD	linkage disequilibrium
LOD	logarithm of odds
LR	likehood ratio
LSO	locus-specific oligo
MAF	minor allele frequency
MST1	macrophage stimulating 1
NELL1	nel-like 1
NKX2-3	NK2 transcription factor-related locus 3
NOD2	nucleotide-binding oligomerization domain containing 2

NPV	negative predictive value
OR	odds ratio
ORMDL3	orosomucoid1-like 3
OTUD3	OTU domain containing 3
PCR	polymerase chain reaction
PLA2G2E	phospholipase A2, group IIE
PPV	positive predictive value
PTGER4	prostaglandin receptor EP4 gene
PTPN	protein tyrosine phosphatase non-receptor
RNF186	ring finger protein 186
S100Z	S100 calcium binding protein Z
SD	standart deviation
SLC22A4	solute carrier family 22, member 4
SNP	single nucleotide polymorphism
STAT3	signal transducer and activator of transcription 3
TCR	T cell receptor
TE	Tris-ethylenediaminetetraacetic acid
TNFSF15	tumor necrosis factor (ligand) superfamily, member 15
TNF-α	tumour necrosis factor α
UC	ulcerative colitis
WGA	whole genome amplification
WTCCC	Wellcome Trust Case Control Consortium

DNA base nomenclature

A	Adenine
C	Cytosine
G	Guanine
T	Thymine

INTRODUCTION

The gastrointestinal tract is a barrier organ that constitutes one of the largest sites of exposure to the outside environment. The intestinal mucosal surface, which consists of a single-layered epithelium, is continuously exposed to a diverse admixture of commensal bacteria comprised of 500 to 1000 species (reaching up to 10^{11}–10^{12} cells per milliliter or gram of luminal contents) [1, 2], as well as to an enormous antigenic load through dietary and environmental factors [3]. The normal response to penetration of epithelium by antigens as well as commensal and pathogenic microbes is controlled by immune system mediated self-limiting inflammation [3]. Dysregulation of the fine-tuned immune response, leading to chronic inflammation of the gastrointestinal tract and loss of epithelial integrity, results in inflammatory bowel disease (IBD) [4].

The two clinically defined conditions of IBD (OMIM 601458) – Crohn's disease (CD; OMIM 266600) and ulcerative colitis (UC; OMIM 191390) – are chronic remittent and progressive immune-mediated inflammatory disorders. They are characterized by episodes of recurring abdominal pain, diarrhea, rectal bleeding and malnutrition [5]. IBD represents an important public health problem. It tends to afflict young people and has a protracted and relapsing clinical course, affecting patients working abilities, education, social life, and quality of life [6, 7]. These disorders also increase the risk of colon cancer [8]. Although mortality is low, morbidity associated with IBD is substantial [9]. The incidence of the diseases is reported to be the highest in the industrialized Western countries, with prevalence rates in North America and Europe ranging from 21 to 246 per 100,000 inhabitants for UC and 8 to 214 per 100,000 inhabitants for CD [10]. Until recently, only few data was available on the epidemiology of IBD in the developing countries. The recent data from this region indicated low, but gradually rising incidence of IBD [11, 12].

The precise etiologic and pathogenetic mechanisms underlying the pathogenesis of CD and UC remain uncertain. However, the presently available data support the hypothesis of a complex interplay between genetic factors in a fraction of the population and the rather drastic change in environmental conditions that took place over the last century [13, 14]. Geographic differences in disease distribution as well as changes in incidence over time in particular populations suggest a role of certain environmental factors (hygiene, nutritional habits, smoking, the industrialization of both food production and preservation and viral and bacterial agents) [15]. The genetic component of the diseases is

supported by the observations that IBD tends to cluster within families as well as on the increased concordance of pathologic phenotypes in monozygotic versus dizygotic twins [13]. Former studies noted that CD patients' relatives have a 10–fold increased risk to develop the disease compared to controls; whereas the risk for UC has been found to be 8–fold increased [13].

International groups have been searching for IBD susceptibility genes over the past 15 years using linkage stdies, candidate gene approaches, and targeted association mapping [16, 17]. However, the advent of genome-wide association studies in the last 5 years has generated new insights into the genetic basis of disease pathogenesis. Currently, more than 99 genes/loci conferring susceptibility to either CD (*e.g., NOD2, ATG16L1, IRGM, LRRK2, PTPN2, ICOSLG, ORMDL3*), UC (*e.g., ECM1, IL10, IL22, IL26, ARPC2,* OTU domain containing 3 (*OTUD3*)) or both forms of IBD (*e.g., IL23R, JAK2, STAT3, LYRM4, MST1*) are known [17-19]. However, in order to distinguish true positive associations from spurious ones, independent replication of results, preferably in large sample sets with matched controls and disease phenotypes comparable with those used in the initial studies, are required.

Aim and objectives of the study

The aim of the study is to investigate the role of the inflammatory bowel disease associated genetic variants in a subset of Crohn's disease and ulcerative colitis patients from Lithuania and Latvia and to test the relation of genetic markers to disease phenotype.

Objectives of the study:

1. Determine the association of the inflammatory bowel disease associated single nucleotide polymorphisms in the subset of Crohn's disease and ulcerative colitis patients.
2. Evaluate the association of the single nucleotide polymorphisms with the phenotype of the inflammatory bowel disease.
3. Determine the interactions of single nucleotide polymorphisms (SNP-SNP) and their association with inflammatory bowel disease.
4. Evaluate the significance of the combinations of disease associated single nucleotide polymorphisms for diagnosis of inflammatory bowel disease.

Originality of the study

Recently performed numerous genome-wide and linkage studies have identified and replicated significant associations between inflammatory bowel disease development and polymorphisms of genes attributed to recognition of bacterial products, adaptive and acquired immune responses, autophagy pathways, *etc*. Given the heterogeneity in allele frequencies reported for the genetic factors involved in the pathogenesis of inflammatory bowel disease in different populations, the thorough replication of the study results in different populations is essential.

Compared to the western countries Baltic countries still observe low IBD incidence rates, especially for CD in their populations. Therefore, analysis of the genetic contribution to disease susceptibility in this region is of great interest. The genetic studies in the field of IBD in the Lithuanian study population started from year 2002. The research group of the Department of Gastroenterology, Kaunas University of Medicine (Kaunas, Lithuania) in close collaboration with the Institute for Clinical Molecular Biology, Christian-Albrechts University in Kiel (Germany) has performed the first genetic study of

IBD examining the frequencies of the previously described variants in the *NOD2*, *IL23R* and *ATG16L1* genes in a Lithuanian IBD study population (these results are included in the thesis). The research study results have been published in the peer-reviewed journal *World Journal of Gastroenterology* in 2010 [20]. The results of the study indicated that CD in Lithuania has a strong genetic background that relates partially to *NOD2* susceptibility variants, especially Leu1007insC. The relatively high carriership frequency of any of the three *NOD2* alleles in the healthy controls (16.9%) in our study is in contrast with the data of low CD incidence in Lithuania. This indicates the importance of other genetic and/or environmental factors (*e.g.*, diet, lifestyle) in disease development.

Therefore, in the frames of this doctoral thesis further genetic explorations of multiple IBD associated genetic markers in the subset of IBD patients from Lithuania and Latvia was undertaken. This study was implemented in collaboration with the Institute for Clinical Molecular Biology, Christian-Albrechts University in Kiel (Germany), where DNA extraction, whole genome amplification, genotyping, and part of data analysis had been performed. In the study, not only possible single nucleotide polymorphisms associations with the diseases were analysed, but also the possible links to IBD phenotypes. The relationship of genotype to phenotype is a fundamental problem in the genetics of complex disorders. Through these investigations it is hoped that deeper understanding of the phenotypic expression as well as disease susceptibility will be gained. Moreover, this study is one of the first studies analysing the possible interactions of single nucleotide polymorphisms (SNP-SNP) and their association with IBD. There is growing evidence that genetic interactions, whether synergistic or antagonistic, are not only possible but are also ubiquitous [21-23]. The inheritance of combinations of functional and disease-linked commonly occurring SNPs may additively or synergistically disturb the system-wide communication of the biological processes, leading to disease [21]. Finally, we evaluated the significance of the combinations of disease associated single nucleotide polymorphisms for diagnosis of inflammatory bowel disease. Noninvasive genetic risk profiling would be valuable in diagnosis and management of inflammatory bowel disease.

1. LITERATURE REVIEW

1.1. Inflammatory bowel disease

Inflammatory bowel disease (IBD, OMIM 266600) is a relapsing-remitting immune-mediated disorder of the gastrointestinal tract. The two clinically defined subphenotypes of IBD, Crohn's disease (CD; OMIM 266600) and ulcerative colitis (UC; OMIM 191390), are progressive inflammatory disorders that may affect the entire gastrointestinal tract or only the intestinal mucosa. IBD represents an important public health problem, as it tends to afflict young people and has a protracted and relapsing clinical course, affecting patients working abilities, education, social life, and quality of life [6, 7]. Although mortality is low, morbidity associated with this disease is substantial [9]. IBD predominantly is regarded as an idiopathic multifactorial disorder, as the genesis of it is still unclear. However, the presently available data overwhelmingly support a hypothesis centered around a complex interplay between genetic factors in a fraction of the population and the rather drastic change in environmental conditions that took place over the last century [13, 14].

1.1.1. Clinical aspects of Crohn's disease

Colonic "regional ileitis" was not recognized as a separate entity until 1932 when Drs. Crohn, Ginzburg, and Oppenheimer initially described it as a distinct disease [24]. With later knowledge that the disease could also affect other sites of the gastrointestinal tract, the "Crohn's disease" term became accepted.

CD is characterized by a focal or multifocal chronic transmural inflammation extending the entire thickness of the intestinal wall (Fig.1.1.1.1). Areas of deep ulceration can form localized regions of lymphoid aggregates (non-caseating granulomas) or tube-like connections between loops of the intestines or nearby organs (fistulas). These complications can be found in 26%–37% of patients, and may indicate a more aggressive disease course [25]. Another feature of CD is its segmental distribution, *i.e.*, regions of inflammation can be separated by tissue with normal appearance. The inflammation can affect any part of the gastrointestinal tract from the oropharynx to the perianal area [24]. The characteristic histological features of CD are: mucosal inflammation (neutrophil infiltration into the epithelial layer and crypts), chronic mucosal damage, ulceration, transmural inflammation

affecting all layers, and noncaseating granulomas [24]. Signs and symptoms of CD can include diarrhoea, abdominal pain, fever, rectal bleeding, weight loss, clinical signs of bowel obstruction [5, 24]. These symptoms are largely dependent on the location of inflammation and the disease behavior [26].

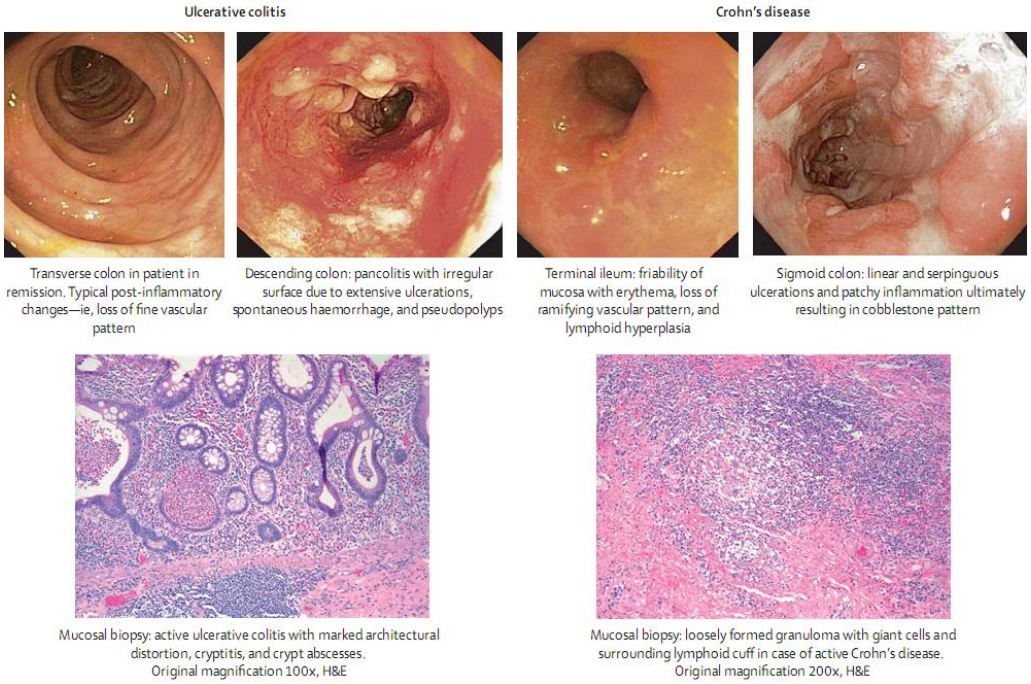

Fig. 1.1.1.1. *Common and distinct features of ulcerative colitis and Crohn's disease* [5]
H&E = haematoxylin and eosin stain.

In 2005, the Montreal clinical classification [27] revising the previously developed Vienna classification [26] was developed to describe the distinct clinical phenotypes of CD based on the anatomical location and behavior of disease. At diagnosis, the disease is located in the terminal ileum (L1) in 47% of cases, the colon (L2) in 28%, the ileocolon (L3) in 21%, and the upper gastrointestinal tract (L4) in 3%. Disease behaviour is classified as non-stricturing and non-penetrating (B1) in 70% of patients, stricturing (B2) in 17%, and penetrating (fistulas or abscesses or both; B3) in 13% of all patients at diagnosis [5]. In addition to inflammation of the intestine, several immune-

related extra-intestinal manifestations are common in CD, such as arthritis, erythema nodosum, pyoderma gangrenosum, aphtous stomatitis and uveitis [28].

The medical treatment approach for CD is individualized based on intestinal location of disease, the severity of symptoms and complications. In active moderate CD, sulfasalazine is effective in left-sided colonic disease management, while budesonide is a first choice in right-sided colonic disease and terminal ileum disease. For patients with active moderate to severe CD, corticosteroids remain a first line treatment. In more severe active disease prednisone is indicated. Chimeric monoclonal antibodies directed against tumour necrosis factor α (TNF-α; Infliximab) are used for patients failing to respond or intolerant to steroid treatment, or when CD is complicated by perianal fistula. Immuno-modulators, like azathioprine, 6-mercaptopurine and methotrexate, are effective to maintain remission after induction therapy, in particular with steroids. However, about 70% of patients, who do not improve with medical therapy, ultimately need surgery during the course of disease [5, 29].

1.1.2. Clinical aspects of ulcerative colitis

UC was first described as a non-infectious disease, *i.e.*, pathology distinct from dysentery, by Wilks and Moxon in 1875; while the term "ulcerative colitis" was introduced by Hale-White in 1888 [30].

In UC, the inflammatory response and morphologic changes remain restricted to the large bowel. The disease typically presents with symptoms of rectal bleeding, abdominal pain and distension, diarrhea, loss of appetite, and weight loss [5, 24]. The inflammation seen in UC is continuous and superficial including the mucosa and submucosa of the intestinal wall (Fig. 1.1.1.1). Histologically active UC typically consists of a neutrophilic mucosal infiltrate, goblet cell depletion, "cryptitis", and prominent crypt abscesses [24]. The disease typically starts from the rectum and extends proximally to include parts of the entire colon [24]. Fulminant colitis is a rare and severe form of the disease that involves the entire intestinal wall and often leads to potentially fatal outcomes such as toxic megacolon, colonic perforation, and peritonitis [5, 24]. UC patients with extensive colon involvement are also subjects to a progressively increasing risk of colorectal cancer after 8 years of disease [8].

The classification system for UC was developed in 2005 by "Montreal Working Party" [27]. UC is classified based on the severity as well as the anatomic extend of inflammation. Approximately 30%–50% of UC patients have disease confined as proctitis (E1; rectum only) at diagnosis, 20%–30% have left-sided disease (E2; up to the splenic flexure) and 20%–30% have extensive colitis or pancolitis (E3; extending beyond the hepatic flexure) [5]. The extra-intestinal manifestations are also present in UC. Most of them are similar to the manifestations present in CD (chapter 1.1.1). Primary sclerosing cholangitis is a serious extra-intestinal manifestation, which is more often associated with UC (2.5%–7.5% of UC also have primary sclerosing cholangitis) than CD affected patients [24, 28].

Therapeutic decisions in UC depend on the anatomic extent and severity of the disease [5]. The first-line therapy for patients with mild to moderate UC consists of 5-aminosalicylates, which can induce and maintain remission. Steroids are effective for patients who are intolerant or are not responding to 5-aminosalicylates. Cyclosporine can be used in patients with severe active UC who are steroid-resistant. Biological therapy (*e.g.*, Infliximab) has been proved to be effective in the management of moderate to severe active UC cases [31]. Usage of this chimeric monoclonal antibody is indicated for patients who fail to respond to therapy with corticosteroids and/or immunomodulators. Surgery treatment is necessary in acute toxic colitis, intractable disease or colorectal cancer [29, 32].

1.1.3. Epidemiology

The epidemiologic and etiologic considerations in UC and CD have many features in common and further will be discussed together. The geoepidemiological picture of IBD varies considerably. The disease is more common in developed, industrialized countries, pointing at urbanization as a potential risk factor. The highest incidence and prevalence rates are reported in Northern and Western Europe and North America, the geographic regions with the earliest described cases of IBD [9, 33]. The incidence rate of CD varies and is approximately 4–10 cases per 100,000 inhabitants annually, whereas the incidence of UC is stable at 6–15 per 100,000 annually. CD and UC have a combined prevalence of 200–300 cases per 100,000 inhabitants and the life-time risk has been calculated to be 0.15% for CD and 0.3% for UC.

The incidence is characterized by a north-south as well as west-east gradients. In Europe, the incidence of CD is 80% higher in the northern countries (Scandinavia) compared with southern countries (Portugal, Greece) (risk ratio = 1.8, 95% CI: 1.5–2.1) [34]. The rates for UC in the northern populations are 40% higher than in the southern (risk ratio = 1.4, 95% CI: 1.2–1.5). Until recently, only a few data was available on the epidemiology of IBD in the East European countries. The recent data from this region indicated low, but gradually rising incidence of IBD (Lithuania (2006) – UC: 9.0 per 100,000, CD: 2.0 per 100,000 [35]; Hungary (1977–2001) – UC: 5.9 per 100,000, CD: 2.2 per 100,000; Croatia (1995–2001) – UC: 4.9 per 100,000, CD: 4.8 per 100,000 [11]; and Estonia (1993–98) – UC: 1.7 per 100,000, CD: 1.4 per 100,000 [36]) and gave the evidence of possible existence of west-east gradient in the European countries [11, 12].

Both UC and CD have a bimodal distribution of the age of disease onset: the first peak occurs in 15 to 30 years old individuals, and a second, smaller peak – in 50 to 70 years aged individuals [37]. About 25% of cases occur in childhood and adolescence. The early-onset IBD is characterized by a rapid and extensive progression that has a detrimental effect on growth and development [38]. UC is slightly more common in males, whereas CD is marginally more frequent in female [9, 39]. Breakdowns by racial and ethnic subgroups indicate that higher rates of IBD occur in people of Caucasian and Ashkenazi Jewish origin than in individuals from other backgrounds, and this is irrespectible of time period and geographic location [10]. However, in the past decades the migration wave to the developed countries caused the increase of incidence in African Americans, in second generation south Asians, and other immigrant groups [33, 40]. Thereby, indicating substantial impact of environmental or lifestyle components to disease risk.

Survival of UC and CD affected patients does not differ from the general population. The risk of colorectal cancer is slightly increased in extensive UC and CD, but the overall survival is similar to the general population [41, 42]. However, in a recent European multicentre study an overall increased mortality was seen ten years after diagnosis, especially in patients diagnosed with CD beyond the age of 40 years and mainly due to gastrointestinal causes [43].

1.1.4. Pathogenesis

Current evidence from research in basic science and clinical trials bring a deeper understanding to the genetically determined interplay between the commensal microbiota, intestinal epithelial cells, and the immune system and the manner in which this interplay might be modified by relevant environmental factors in the pathogenesis of IBD [44]. These studies indicate that CD and UC are heterogeneous diseases characterized by the number of distinct genetic abnormalities that lead to disruption of distinct molecular mechanisms (Fig. 1.1.4.1).

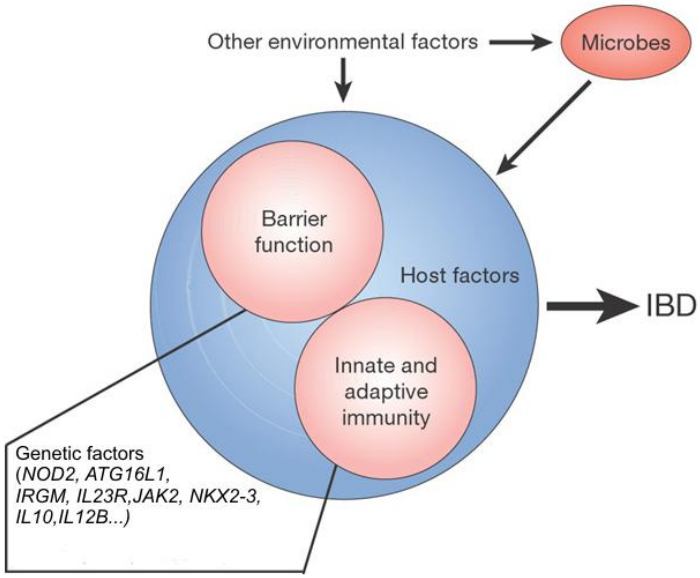

Fig. 1.1.4.1. *A schematic figure of the current genetic model in inflammatory bowel disease*

Specific genes combined with environmental factors lead to either CD or ulcerative colitis UC (figure modified from [44]).

Adaptive immunity. The traditional concept regarding the cause of IBD is an abnormal response of cells from the acquired immune system. This response leads to either an exaggerated aggressive activity of effector lymphocytes (Th1, Th2, and Th17) and excessive production of pro-inflammatory cytokines or deficient regulatory T-cell function (*e.g.*, CD4+CD25-T cells, CD4+CD25+

FOXP3+T cells (Treg)) and reduced secretion of the anti-inflammatory cytokines in response to gut-derived antigens presented by antigen-presenting cells [45–47]. The success of treatment with monoclonal antibodies targeting specific immune components (like anti-interferon γ (anti-IFN-γ), anti-IL12/IL23p40 or anti-TNFα) supports this hypothesis [48]. Immunologically it has been proven that in the intestinal mucosa and peripheral blood mononuclear cells of CD patients and murine models there is an increased production of the Th17 cytokine interleukin-17 (IL17) and Th1 cytokines: IL-12, IFN-γ and TNF-α [49, 50]; in UC, by contrast, there is an increase in IL-17 and atypical set of Th2 cytokines (like IL-4, IL-5, IL-13) [49, 50]. In contrast, the observed numbers of Tregs and the amount of regulatory cytokines (IL10, TGF-β) are reduced in the blood and colon of IBD patients [46, 51]. In addition, mice engineered to lack expression of regulatory cytokines succumb to wasting disease and colitis when disease-triggering bacteria are present in the intestinal flora [46]. Moreover, recent genetic studies demonstrating genes involved in the effector T-cell (e.g., interleukin 23 receptor (*IL23R*), *IL12B,* Janus kinase 2 (*JAK2*), signal transducer and activator of transcription 3 (*STAT3*), CC chemokine receptor 6 (*CCR6*), tumor necrosis factor superfamily, member 15 (*TNFSF15*)) and regulatory T-cell functions (*IL10,* actin-related protein 2/3 complex gene (*ARPC2*)) as IBD susceptibility genes supports the important role of adaptive immunity in disease pathogenesis [17].

Innate immunity. Innate immunity is the first level of defense comprised of physical and biochemical barriers that prevent microbial invasion. It is mediated by a large variety of different cell types including epithelial cells, neutrophils, monocytes, macrophages, dendritic cells, and natural killer cells. Defects in mucosal barrier (in detail see next paragraph "*Epithelium*") and microbial clearance functions have been associated with pathogenesis of IBD [52, 53]. The genetic IBD studies indicated susceptibility loci in genes responsible for recognition of bacterial antigens (toll-like receptor (*TLR*) and nucleotide-binding oligomerization domain (*NOD*) family genes) [54–56] and their intracellular elimination (immunity-related GTPase family, M gene (*IRGM*) and autophagy-related protein 16-like 1gene (*ATG16L1*)) [57–59]. The biochemical and functional studies in epithelial cells revealed that mutations in bacterial antigens recognising genes cause defective bacterial recognition, antigen presentation and cellular response [53]. These findings were confirmed in animal studies [60]. In the recent years, the altered process of intracellular microbial elimination (autophagy) has also been implicated in the pathogenesis

of IBD (*ATG16L1* [57], *IRGM* [58]). Autophagy is an evolutionarily highly conserved innate defense mechanism important for cellular homeostatic functions [61]. The exact role of autophagosomal mutations in IBD pathogenesis is not clear yet. However, animal models indicated that *ATG16L1* deficient mice had an impaired autophagosome function and increased susceptibility to dextran sodium sulfate colitis [62, 63]. Recent studies provided a functional link between bacterial sensing by NOD proteins and autophagy [64, 65]. NOD2 initiates autophagy by recruiting ATG16L1 to the cell membrane at the site of bacterial entry [64]. Dendritic cells from CD patients with *NOD2* or *ATG16L1* mutations are defective in autophagy, bacterial trafficking and antigen presentation [65]. Collectively, these studies suggest that defects in innate immunity causing inappropriate bacterial clearance might be a driver of persistent inflammatory responses in IBD.

Epithelium. The epithelial barrier of intestine is important as it is the first (anatomical) defence level impeding penetration of macromolecules and intact bacteria. Defects in mucosal barrier integrity and repair lead to constant stimulation of the mucosal immune system by luminal antigens [52]. Genetic studies indicated susceptibility loci in genes responsible for regulation of mucosal repair and barrier functions (*e.g.*, prostaglandin receptor EP4 (*PTGER4*), mucin 19 (*MUC19*), x-box binding protein 1 (*XBP1*), extracellular matrix protein 1 (*ECM1*)) [53, 66–68], and transepithelial transport (solute carrier family 22, member 4 (*SLC22A4*) and *SLC22A5* genes) [69]. In IBD patients and their unaffected first-degree relatives, an increased intestinal permeability and defective regulation of tight junctions has been found [44, 52]. In patients with CD and their relatives, this barrier disfunction was associated with *NOD2* polymorphisms [70], and experimental models demonstrated that barrier dysfunction can activate mucosal immune response and sensitize subjects to disease [71]. Moreover, the expression analysis in human mucosa biopsies has demonstrated downregulation of epithelial junctional complexes (E-cadherin and β-catenin) in IBD patients, although the underlying mechanisms are still unknown [44, 52]. The animal models of IBD also provide a strong evidence for epithelial barrier in disease predisposition [60]. It has been reported that defects in epithelial-cell development or proliferation, barrier function, cell-matrix adhesion, endoplasmic reticulum stress, and epithelial restitution after injury increase susceptibility to induction of severe colitis [44, 52, 53]. Thus, the current data strongly support the deranged epithelial function as a critical compenent of IBD pathogenesis.

Bacteria. There is convincing evidence that alterations of the intestinal ecosystem may lead to impairment of the intestinal barrier function and initiation of IBD. It has been assigned that deviation of the faecal stream [72], antibiotic and probiotic treatment (especially in pouchitis) can ameliorate IBD [73]. Although a number of specific pathogens have been associated with the development of IBD (reoviruses, mycobacteria, helicobacters, *Listeria monocytogenes*, *etc.*), none of them have been confirmed as causal; rather, microbial antigens that are normally present in the intestinal lumen seem to drive inflammation in the gut [74]. 16S ribosomal RNA analysis revealed a detectable difference between the number and the diversity of intestinal microbiota in CD and UC compared to healthy controls [74]. IBD patients had a 10–fold lower bacterial load, characterized by depletion of commensal bacteria, notably members of both major classes of commensal phyla, *Firmicutes* and *Bacteroidetes* [53, 75]. The importance of the luminal flora is more directly supported by animal studies using chemically (dextran sodium sulfate) or genetically ($IL10^{-/-}$, $Rag2^{-/-}Tbx21^{-/-}$ (mice deficient recombination activating gene 2 and T-box transcription factor 21 genes)) induced gut inflammation models [53]. It has been demonstrated that in susceptible murine strains even a single species of normal bacteria is possible to induce colitis (*e.g.*, *Bacteroides vulgatus* in the $IL10^{-/-}$ mice) [60]. These studies provided compelling evidence that the nature of the host defence system, rather than the biological properties of the intestinal microbiota *per se*, may determine the functional outcome of that dynamic interaction.

Environment. A number of unrelated environmental factors have been proposed as risk factors for IBD, including smoking, appendectomy, infections ("Hygiene" hypothesis), events in childhood, "Western" nutritional practice, dietary additives, socio-economic changes, drugs, and stress [15]. The hypothesis of environmental involvement in IBD pathogenesis is supported by the observed excess of familial aggregations in CD-affected families [76] as well as by the increasing incidence of disease in the developing countries [11, 12]. However, the potential action mechanisms of environmental factors are very poorly understood. In general, these triggering factors have an effect on the mucosal barrier integrity, immune responses, or luminal microenvironment, leading to the intestinal inflammatory response [15]. Moreover, environmental factors act in the context of genetic risk factors (*i.e.*, gene-environment interactions) are likely to underpin the complexity of disease phenotype.

Genes. The contribution of genetic factors to IBD has long been recognized and the search for disease-causing genes has been of major interest since the first CD gene, the *NOD2* gene, was identified in the year 2001 [54, 55]. The technological advent in the last 5 years has completely changed the landscape of the IBD pathogenesis. Currently, 99 genes/loci conferring susceptibility to CD, UC or both forms of IBD are known [17–19]. The genetic aspects of IBD will be described in more detail in the next paragraphs.

1.2. Genetic mapping in complex human diseases

The genetic basis of IBD has been pursued using genetic linkage and association studies. These two genetic approaches allow finding causal genes without *a priori* knowledge about the underlying biology, the position on the genome or contribution to the disease [77–79]. At a fundamental level, genetic association and linkage analysis rely on similar principles and assumptions. Both rely on the co-inheritance of adjacent DNA variants, with *linkage* capitalizing on this by identifying haplotypes that are inherited intact over several generations (such as in families or pedigrees of known ancestry), and *association* relying on the retention of adjacent DNA variants over many generations (in historic ancestries) [77, 80, 81].

1.2.1. Genome-wide linkage studies

The success of positional cloning of diseases with simple Mendelian inheritance via genome-wide linkage studies has led to increased application of this approach in the analysis of complex genetic traits [82]. This method enables the identification of rare, high-risk, disease-associated mutations, owing to the clear inheritance patterns they display (Fig. 1.2.1.1) [83].

A whole genome linkage scan usually includes typing of about 300–800 evenly distributed tri- or tetranucleotide repeats (*e.g.,* [CAG]n) within cohorts of affected relatives (*e.g.,* affected sibling pairs, pedigrees of independent families). If the marker allele sharing between affected relative pairs is more than 50% (*i.e.,* significantly different from the expected ratio), the general area surrounding the marker is assumed to be disease associated [83]. The logarithm of odds (LOD) score, *i.e.,* the function of the recombination fraction, is used to express the extent by which allelic sharing between individuals is greater than would be expected by chance. The closer the microsatellite is located to the

disease gene, the higher LOD score value is expected because of rarer recombination. A LOD score of 3.0 was proposed as providing a significant evidence of linkage equivalent to P = 10^{-4} [83]. The issue of statistical significance threshold in the whole-genome screens led to the development of stringent criteria defining a significant degree of linkage, *i.e.,* LOD = 3.6 [84].

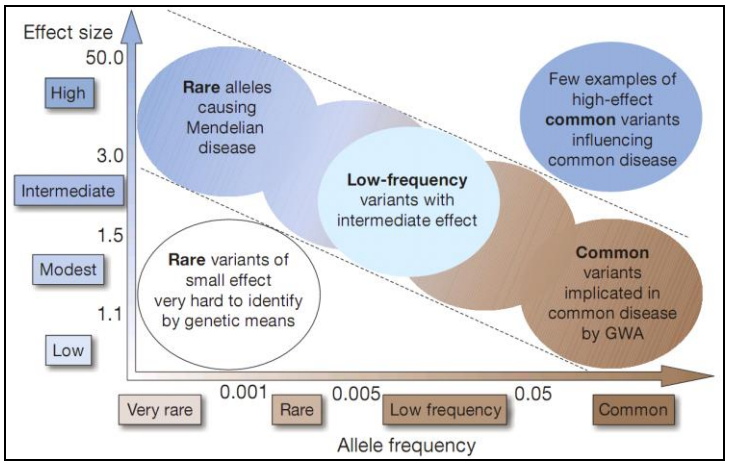

Fig. 1.2.1.1. *Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio)*

Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines [81].

As linkage focuses only on recent, usually observable ancestry, in whom there have been relatively few opportunities for recombination to occur, disease gene regions that are identified by linkage are often large, and can encompass hundreds or even thousands of possible genes across many megabases of DNA [77]. Once the region of interest has been narrowed down to a sufficiently small area, fine mapping approach and genetic association studies (*e.g.*, candidate gene approach) are being used in order to identify the specific trait underlying gene. Identified potential candidate genes are further validated in independent case-control or family-based (using transmission disequilibrium testing) cohorts [85].

Although powerful for detecting genetic loci in single gene disorders, linkage analysis attempts for common, multifactorial disorders has been difficult to replicate, presumably because linkage is less powerful when risk

variants have small effects and there is heterogeneity in the underlying genetic factors in different families [86–88]. Nevertheless, whole genome linkage analysis identified several strong single gene effects (*e.g., NOD2, OCTN etc.*) for complex diseases [54, 55, 69] that have been abundantly confirmed by genetic linkage and association studies and later verified by genome wide association studies.

1.2.2. Genome-wide association studies

The recognized limitations of existing linkage strategies in complex diseases have raised the requirement for a radicaly new methodology in exploration of these disorders [87]. The combination of progress in high throughput genotyping technology [89] and growing knowledge about the human genome through the Human Genome Project [90] and the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) [91] has enabled the development of genome wide linkage disequilibrium mapping, *i.e.*, genome wide association studies (GWAS).

The proposed common-disease common-variant hypothesis was the basic strategy of GWAS [92, 93]. It was assumed that genetic variants displaying neutral or favorable effect with respect to survival became common (allele frequency >5%) and due to changed environmental conditions (*e.g.*, preserving fat during an ice age, but leading to obesity in the fast food era) they had acquired mildly harmful effects [78]. The SNPs, which consist of a change in a single nucleotide at a particular location in the genome, are the most common form of genetic variation with over 23 million present, more than 10 million of which were successfully validated (SNP database dbSNP [94]) [95].

At a fundamental level GWAS rely on the linkage disequilibrium (LD) between the markers and the causal variants. It has been assigned that adjacent alleles assort together non-independently from generation to generation because they are tightly linked (*i.e.*, they form haplotype) and thus less likely to become separated by recombination (Fig. 1.2.2.1) [96]. Therefore, when a functional mutation occurs – perhaps one that contributes to disease – it does so on a haplotype of other pre-existing DNA variants. The deeper insight into the degree of association between the alleles of neighbouring SNPs (*i.e.*, LD) was gained through the International HapMap Project [91, 95, 97]. The development of a high-resolution haplotype map enabled the selection of maximally informative, non-redundant subsets of markers across the regions of interest to

type in GWAS. A wide variety of haplotype-based and pairwise tagging methods were developed, that reduced the number of SNPs genotyped in a study without substantially decreasing the amount of information generated [98–101]. It has been estimated that approximately 300,000–500,000 tag SNPs are required to capture the majority (~70%) of common variation in the human genome of European ancestry [95, 102]. GWAS allow the investigator to narrow an association region to a 10–100 kb length of DNA, in contrast to the 5–10 Mb detected in familial linkage studies [103].

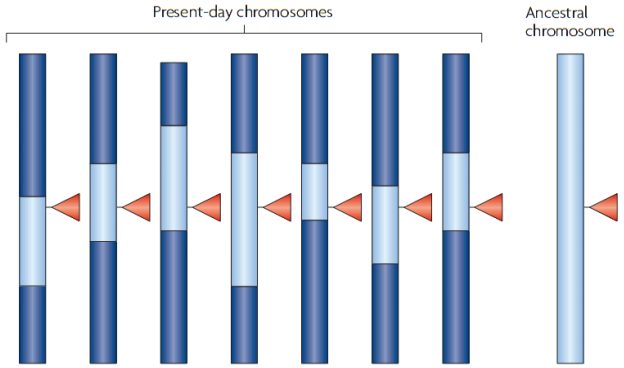

Fig. 1.2.2.1. *Linkage disequilibrium around an ancestral mutation*
The triangle is the mutation that has occurred in the ancestral chromosome. Chromosomal regions that were introduced by recombination are shown in dark color. Markers that are physically close (*i.e.*, within the light-blue regions of present-day chromosomes) tend to remain associated with the ancestral mutation, even as recombination changes the region of association over time [96].

The development of high-throughput genotyping platforms was the last step that allowed the GWAS to revolutionize the current research of complex genetic disorders. Over the last three years, several companies have developed commercial arrays (so called DNA chips) that assay SNP sets with high accuracy (0%–2% missing data, <0.5% errors), at reasonable cost (approximately $500 [U.S.] per subject), and rapidly (>1000 DNA specimens per week) [104, 105]. To date the high density genotyping chips have the potential to assay up to 1 million markers (Affymetrix SNP 6.0 and Illumina 1M) [89]. However, balancing cost and efficiency issues, it has recently been suggested that the most cost-effective way to perform a GWAS is to continue

using the older and cheaper arrays with medium density (300–500k SNPs) and then computationally predict the missing data (untyped SNPs) in the remainder of the genome based upon the observed data by means of the HapMap reference (so called imputation) [106].

The typical GWAS screens the genomes of several hundreds to thousands of subjects (case-control or population-based cohort) with subset of SNPs, followed by a single-locus association test [107]. Subsequent data quality control, as well as large-scale replication in independent sets of patients and controls across similar and diverse populations (using functionally different genotyping technology) ensures the accuracy of the result [107–109]. The GWAS approach has proven itself particular effective at detecting common SNP variants with the modest effects (odds ratio (OR) between 1.11 and 1.29) on phenotype and the modest proportion of heritability for most traits (Fig. 1.2.1.1) [110]. The hard chalange following initial GWAS is fine mapping seeking to determine causative mutations, followed by functional studies to understand the true biology behind the association [17].

One of the important limitations of GWAS includes the generation of false positive associations. One of the main sources of false positive associations is the statistical fluctuations that arise by chance and result in low P–values (which are likely to occur when testing multiple hypotheses). As the large number of SNPs (up to 1 million) is assessed, the application of the traditional P–value cut-off to a GWAS leads to a very large number of false-positive results [111]. At the usual $P < 0.05$ level of significance, a GWAS of one million SNPs will report 50,000 SNPs to associate significantly with disease, almost all spuriously. Therefore, very conservative P–value cut-off using robust tests of multiple corrections (*e.g.*, Bonferroni's method) is applied, *i.e.*, P value of $\leq 10^{-7}$ is required for genome-wide significance [112]. The correction for multiple testing is applied to the independent replication studies as well.

Another cause of the false-positive associations to which GWA studies are prone is population stratification due to ethnic admixture [113,114]. Due to the small affect sizes of the individual SNPs, potentially successful GWAS and replication studies rely on large sample numbers [115]. Therefore, multiple cohorts from different countries are enrolled in the genetic research. Heterogeneity between studied samples can give false-positive results in association studies, as association with the trait may be the result of the systematic ancestry difference in allele frequncies between groups [116]. Techniques have been developed to detect and correct for population stratification [114, 116–119].

The genetic structure of the European populations has been investigated extensively [120–123]. Large-scale studies have demonstrated the existence of the genetic substructures (*e.g.,* clustering of Northern and Southern populations) within Europe and showed that this information can be used for improving error rates caused by population stratification in association studies of candidate genes and in replication studies of GWA scans. However, only the recent study by Nelis *et al.* (2009) [124] performed a detailed analysis of the North Eastern European populations. The study involved 19 cohorts from 16 different Europe countries. The analysis was based on three different measures: the inflation factor λ, fixation index (F_{st}) and principal component (PC). A detailed description of the European population structure revealed that several distinct genetic map regions can be distinguished: (1) Finland, (2) the Baltic region (Estonia, Latvia, Lithuania), Western Russia and Poland, (3) Central and Western Europe, and (4) Italy (Fig. 1.2.2.2). Thereby indicating that the European populations clustering together can be combined in association analysis (correction for the inter-populational differences is needed).

Further limitation of GWAS is their lack of power for identifying associations with rare sequence variants (<1% population frequency), since these are poorly represented on current genotyping platforms, as are structural variants [81]. The 1000 Genomes Project (1kGP) [125] aims at generating a comprehensive catalog of SNPs with a prevalence of 1% to 5%. The acquired knowledge will be usefull for fine-mapping efforts and expansion of genome wide association arrays.

In the past 5 years, nearly 800 significant associations ($P < 5\times10^{-8}$) have been reported in 150 distinct diseases and traits [126,127]. However, variants so far identified by GWAS together explain only a small fraction of the overall inherited risk (*e.g.*, ~20% of variance for CD, ~6% for type 2 diabetes) [81, 128]. As SNPs identified through GWAS do not demonstrate any obvious pattern in terms of gene content (only 12% of SNPs are located in, or occur in tight LD with protein-coding regions of genes, 40% of SNPs are in intergenic regions, and another 40% are in noncoding introns) it is supposed that the detection of the true effect showing variants and characterizion of the effect would increase the overall genetic inheritance fraction of the complex diseases [126].

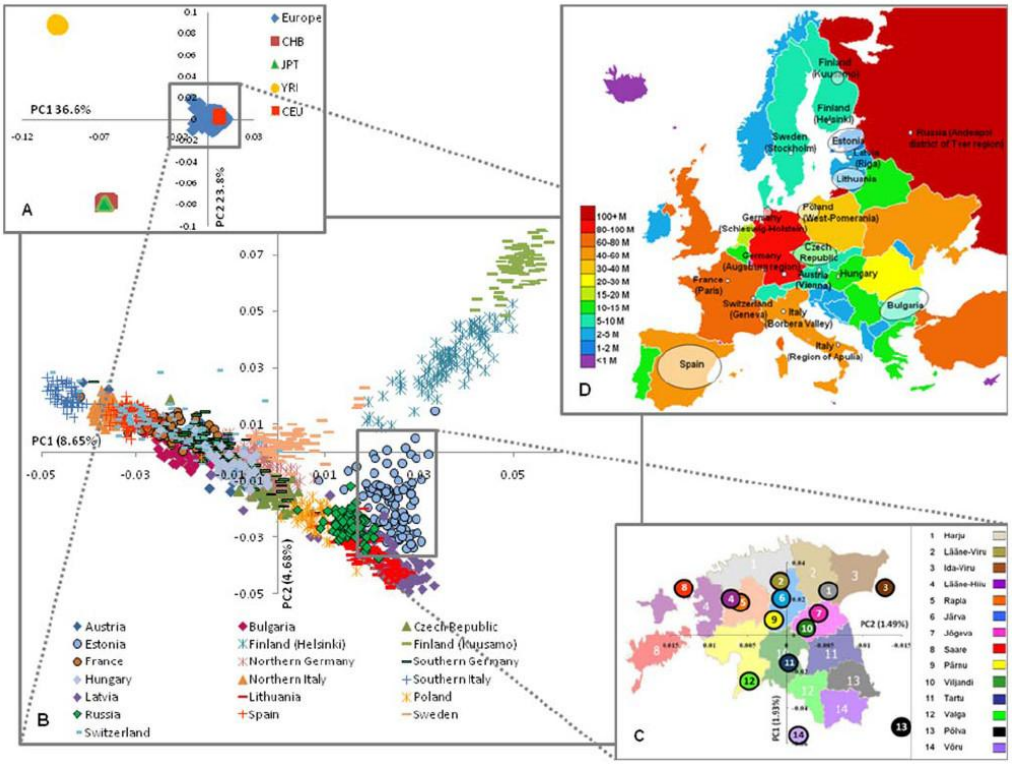

Fig. 1.2.2.2. *The European genetic structure (based on 273,464 SNPs)* [124]
Three levels of structure as revealed by principal component (PC) analysis are shown: A) intercontinental; B) intra-continental; C) inside a single country (Estonia), where median values of the PC1&2 are shown. D) European map illustrating the origin of sample and population size. CEU - Utah residents with ancestry from Northern and Western Europe, CHB – Han Chinese from Beijing, JPT - Japanese from Tokyo, and YRI - Yoruba from Ibadan, Nigeria.

1.3. Epistasis in determining susceptibility to complex human diseases

The past decade has witnessed remarkable success in the identification of low-penetrance, high-frequency susceptibility variants in common, complex diseases [81]. However, a large part of the genetic variance in many of these diseases is still unaccounted for. One of the possible reasons is that complex human diseases result from the poorly understood systematic epistatic interactions of genetic variants [129, 130].

The idea that the effects of a given gene on a trait can be dependent on one or more other genes has been around for at least 100 years. Currently, the growing number of evidences indicated that genetic interactions, whether synergistic or antagonistic, are not only possible but are also ubiquitous [21–23]. The inheritance of combinations of functional and disease-linked commonly occurring DNA sequence variations may additively or synergistically affect proteins that are involved in biological processes ranging from transcription to physiological homeostasis (Fig. 1.3.1). Therefore, the effect might be missed if the gene functioning primarily through a complex mechanism is examined in isolation without allowing for its potential interactions with other genes and, possibly, environmental factors [130]. Disturbance of the system-wide communication of the biological processes leads to disease (*i.e.*, biological epistasis) [21]. Differences in genetical and biological epistasis among individuals in a population give rise to statistical epistasis [129].

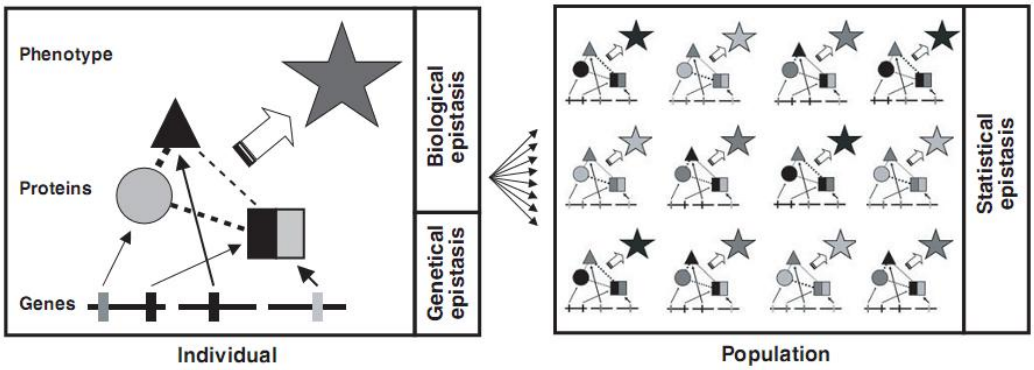

Fig. 1.3.1. *Genetical, biological and statistical epistasis* [129]

For a long time, the type of human data collected was unsuitable for modeling epistasis, but with the advent of the genomic era, a number of studies have demonstrated the presence of gene-gene interactions in complex human diseases. The possible genetic interactions in the association with the IBD have investigated a number of studies. However, the interactions were analysed between pathway-related genes [131–133] or genes that were individually associated with IBD [134–136]. Hypothesis free interaction analysis has been performed only recently. The study performed by Emily *et al.* (2009) [137] analysed SNP-SNP interactions based on the WTCCC genome scale data for

CD [58]. The study has identified the association between the SNP pair (rs6496669 and rs434157) that is in LD with adenomatous polyposis coli (*APC*) and IQ motif containing GTPase activating protein 1 (*IQGAP1*) genes and CD [137]. However, the analysis of both forms of inflammatory bowel disease has not been performed, yet.

1.4. Genetic aspects of inflammatory bowel disease

1.4.1. Genetic epidemiology

Initial evidence for genetic factors to have an important role in determining susceptibility to IBD was based on epidemiological, family and twin studies [13].

Familial aggregation of IBD was first observed in the early 1930s by Burrill B. Crohn himself [13]. Within families with a positive family history, the prevalence has been recorded to vary between 5.5% and 22.5% [15]. Many studies have shown a high degree of concordance for disease type within families, whereas cross-occurrence of CD and UC was estimated within 25% of cases [13]. The probability of developing CD in a relative of a UC patient is increased by 2–fold, and there is a 4–fold risk of UC in a relative of a CD patient [138]. These data support the existence of genetic variants that are common for both diseases and others specific for the disease type, the phenotypic expression being influenced by environmental factors. The relative risk to siblings of affected individuals in developing IBD is estimated to be 30–40 fold for CD and 10–20 fold for UC [139]. The risk of IBD increases if more than one first-degree relative has the disease [13]. Data concerning the phenotypic similarities of IBD within families indicated the concordance for disease type, disease pattern, and presence of extraintestinal disease manifestations [13, 15]. However, these findings are controversial and there are no strong arguments for phenotypic differences between the familial and sporadic forms of IBD.

The most compelling evidence for the role of genetic factors comes from studies in twins. In large European studies conducted in Sweden [140, 141], Denmark [142], UK [143], and Germany [144] the concordance rate for CD in monozygotic twins was estimated at between 20% and 58%, whereas the concordance rate in dizygotic twins brought up in the same environment was less than 10%. The reported concordance rates of UC in monozygotic and

dizygotic twins ranged from 6–17% and 0–5%, respectively. In the discordant IBD twin pairs the first-born twin had the higher preponderance of being affected, *i.e.*, 70% of CD and 81% of UC monozygotic, and 59% of CD and 64.5% of UC dizygotic discordant twins were affected by IBD [144]. The relative risk to develop IBD for unaffected identical twins compared with that for probands of non-identical twins was 3.49 (P = 0.03) [143]. The calculated heritability of liability based on monozygotic twin pairs was 0.53 for UC and 1.0 for CD, suggesting a much stronger genetic influence in CD [140]. Mixed pairs of monozygotic twins are extremely rare [145], suggesting that the net genetic susceptibility factors causing CD and UC are different.

Closer examination of the phenotypic characteristics of the twin cohort provided some evidence that monozygotic twins concordant for CD status when compared to non-identical twin patients had significantly greater similarity in age of onset, disease location, and disease behavior at diagnosis and 10 years post-diagnosis; whereas UC twin pairs were concordant only for age at diagnosis and symptomatic onset but not for extent of disease at diagnosis or after 10 years [146].

Collectively, these observations strongly support the importance of genetic factors for susceptibility to IBD. However, it also indicates that IBD is not inherited as a classical Mendelian trait, but rather has a complex polygenic mode of inheritance.

1.4.2. Inflammatory bowel disease genetic studies

International teams have been searching for IBD susceptibility genes over the past 15 years. The initial IBD genetic research consisted of candidate gene studies analyzing the association of polymorphisms in functionally plausible genes. The associations of human leukocyte antigen (*HLA*) region (*HLA-DR2* (genotypes *DRB1*1501* and *1502*), *HLA-DR103* (genotype *DRB1*0103*)) with both CD and UC in different populations were mostly notable findings of candidate gene approach [147].

Since 1996, the analysis of the genetic basis of IBD was pursued using hypothesis-free scanning for loci of association using linkage studies. Using this model, a total of nine IBD susceptibility loci (designated IBD1–9) were identified and replicated to a varying extent (Fig. 1.4.2.1) [13].

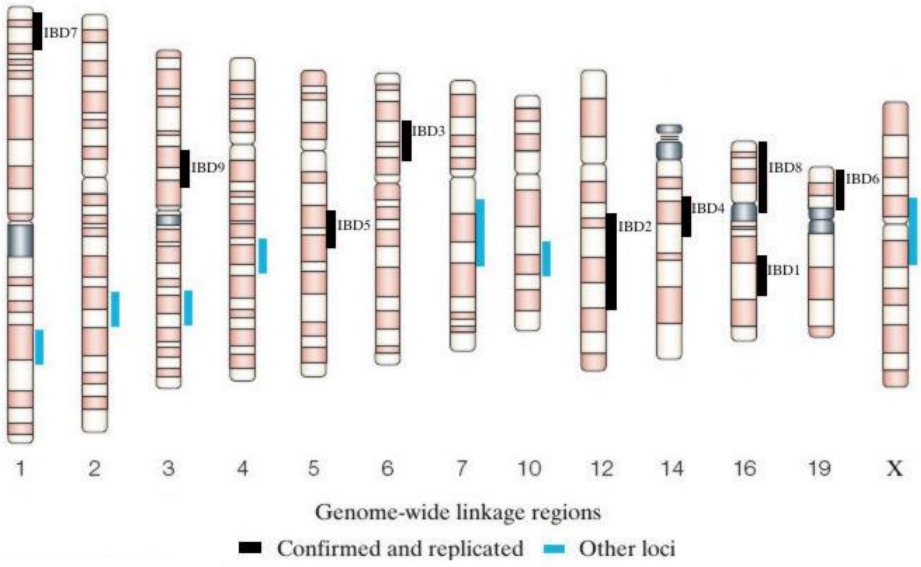

Fig. 1.4.2.1. *Inflammatory bowel disease susceptibility loci and subchromosomal regions identified by nonparametric linkage analysis*
[modified from 14]

Some of these loci appeared to be relatively specific for CD (*e.g.*, IBD1 on 16q) and UC (*e.g.*, IBD2 on 12q), whereas others were associated with IBD as a whole (*e.g.*, IBD3 on 6p) [148]. The closer analysis of region on chromosome 16 (the IBD1 locus) resulted in identification of the first CD susceptibility gene. To be precise, in 2001, two groups simultaneously identified *NOD2* gene, also known as caspase recruitment domain 15 (*CARD15*), as a CD susceptibility gene, using positional cloning and candidate gene approaches [54, 55]. Three variants in this gene (Leu1007insC, Gly908Arg, and Arg702Trp) were found to be associated with CD but not with UC. Patients carrying one of the *NOD2* mutations have a 2–4 fold increased risk of developing CD, while those carrying two mutations have a 20–40 fold increased risk of developing CD [147, 149]. These three *NOD2* variants, however, are carried by only 20–30% of all CD patients, suggesting that other genes are involved in the development of this condition [147, 149]. The association of the three *NOD2* variants with the development of CD has been replicated in a significant number of studies in populations of Caucasian origin from Europe and North America [150], as well as by the number of GWAS [57–59, 67, 151, 152]. The significant impact of

the *NOD2* variants for the CD development has been also confirmed in the first Lithuanian IBD genetic study [20].

In addition, linkage studies facilitated the discovery of suggestive linkage locus on chromosome 5q (IBD5). This association has been confirmed by the number of GWAS [58, 67, 153]. Detailed analysis of this locus demonstrated an association between CD and a common haplotype spanning the chromosome 5q cytokine gene cluster [154–156]. Subsequently, functional polymorphisms in the *SLC22A4* and *SLC22A5* genes in this region were identified, but it remains unclear whether these are the causal genes [69]. IBD5 haplotype has been associated with perianal CD and in some studies, an earlier age of disease onset [65].

Despite much initial promise from these genome-wide linkage studies subsequent progress was frustratingly slow and discovered associations were notoriously difficult to consistently replicate. However, the advent of GWAS in the past 5 years has completely changed the landscape and unparalleled insights into disease pathogenesis have followed. In this time period, there has been a number of high-profile GWAS in CD and, later, UC (Table 1.4.2.1) that have, to date, yielded over 99 IBD disease genes/loci, of which 47 are specific to UC and 71 to CD [17–19]. It is currently estimated that known genetic associations account for approx. 20% of the genetic variance determining individual susceptibility to CD [67], and 16% – UC [170].

Table 1.4.2.1. *Genome-wide association studies performed in IBD*

Population (Reference)	Trait	Genotyping platform	Discovery cohort (case/controls)	Replication cohort (case/controls)
Japanese [157]	CD	Custom array (80k SNPs)	94/752	484/345
N. American (European) [59, 151]	CD	Illumina HumanHap 300	946/977	353/207 and 530 trios
German [57]	CD	Custom SNPlex panel of non-synonymous SNPs (20k)	735/368	498/1032 and 380 trios
German [152]	CD	Affymetrix 100k	393/399	942/1082 and 375 trios
Quebec/German [158]	CD	Perlegen 165k array	382 trios	752/828 and 521 trios
Belgian/French [66]	CD	Illumina HumanHap 300	547/928	1266/559

Table 1.4.2.1 continued

Population (Reference)	Trait	Genotyping platform	Discovery cohort (case/controls)	Replication cohort (case/controls)
British [58, 159]	CD	Affymetrix 500k	1748/2938	1182/2024
German [160]	CD, SA	Affymetrix 100k	382 (CD)/398 (SA)/ 394 (CO)	1549 (CD)/924 (SA)/ 3361 (CO)
European descent/ N.American [67]	CD	Meta-analysis	3230/4829	2325/1809 and 1339 trios
British [68]	UC	Custom Infinium array (11k SNPs)	905/1465	2028/3029
German [161]	UC	Affymetrix SNP 5.0	1167/777	1855/3091
N.American/Italian [162]	IBD	Illumina HumanHap 550	1011/4250	173/3481
N.American [163]	UC	Illumina HumanHap 300 and Human-Hap550	1052/2571	1405/1115
Japanese [164]	UC	Illumina Human-Hap550 and Affemetrix custom 10k	749/2031	635/1026
British [165]	UC	Affymetrix SNP 6.0	2361/5417	2321/4818
N.American/ European descent [166]	UC	Illumina HumanHap 550	1636 (CD)/724 (UC)/ 6158 (CO)	829 (CD)/120 (UC)/5805 (CO)
European descent [167]	UC	Affymetrix SNP 6.0	1043/1703	2539/5428
European descent [168]	IBD, T1D	Illumina HumanHap 550	1689 (CD)/ 777(UC)/989 (T1D)/ 6197 (CO)	
European descent [169]	UC	Meta-analysis	2693/6791	2009/1580
N.American [153]	CD	Illumina Human 610Quad and Illumina 370Duo	896/3204	1174/357
European descent [19]	CD	Meta-analysis	6333/15056	15694/14026 and 414 trios
European descent [170]	UC	Meta-analysis	6687/19718	9628/12917

Table modified from [135]. IBD – inflammatory bowel disease, CD – Crohn's disease, UC – ulcerative colitis, SA – sarcoidosis, T1D – type 1 diabetes, CO – control.

CD has always been at the forefront of the GWAS era. It was the subject of one of the very first GWAS performed in a Japanese population in 2005 [157]. This relatively small study produced strong evidence of association for a SNP in an intron of the *TNFSF15* gene. The association of this gene with CD was replicated in a Caucasian (British) cohort in the same study [157] and in multiple independent European and non-European studies that also showed the association of this gene with UC [168, 171–173].

A North American study group was the first to perform the GWAS of CD with the broad genome coverage (308,332 autosomal SNPs) [151]. Initial study showed a very strong association of multiple SNPs in the interleukin 23 receptor gene (*IL23R*) on chromosome 1p with CD, which was replicated in two independent cohorts (case-control and family-based). Interestingly, the strongest association exhibiting rare, non-synonymous *IL23R* SNP rs11209026 had the reduced frequency in CD cases (OR = 0.26, 95% CI (confidence interval): 0.15–0.43), suggesting the protective effect of this polymorphism. The evidence for the modest association of the reported *IL23R* SNPs with UC has also been shown [151]. This association has been confirmed in the number of different European studies for both IBD subphenotypes [153, 169, 171, 173–176].

An extension of the North American scan [59] replicated previously described *NOD2, IL23R, ATG16L1* variants and identified four novel loci of interest, including a SNP in the promoter of paired-like homeobox 2B (*PHOX2B*) on chromosome 4p13 (rs16853571; $P = 7.7\times10^{-7}$), an intergenic region on chromosome 10q21.1 (rs224136; $P = 7.9\times10^{-6}$), SNPs within an intron of a predicted gene (family with sequence similarity 92, member B (*FAM92B*)) on chromosome 16q (rs8050910; GWA $P = 3.3\times10^{-5}$), and SNPs in an intron of the neutrophil cytosolic factor 4 (*NCF4*) gene on chromosome 22q13 (rs4821544; $P = 2.9\times10^{-5}$) [59]. The findings were replicated in several independent genetic studies [171].

The third CD GWA scan done in the German population revealed an association with *ATG16L1* (rs2241880, $P = 2.9\times10^{-8}$) [57]. The association was replicated in a further German and British panel. This SNP appeared to account for all of the risk at this locus, and might therefore be the causal variant [57]. The contribution of this variant to CD and UC susceptibility has been repeated in the number of association studies [153, 171, 175–177].

A scan of Belgian/French CD cases [66] found an association in a 1.2–Mb region of chromosome 5p13.1 (peak SNP rs1373692; $P = 4.1\times10^{-8}$) that

contains no known genes. CD associated SNPs in this region were correlated with increased expression of the adjacent *PTGER4* gene. This association has been replicated in the independent studies [171, 173]. Other findings from this study [66] included confirmations of the association of the *ATG16L1* variant with CD, and *IL23R* Arg381Gln variant with UC.

A possible functional connection of the *PTGER4* locus was provided by a German case control GWA scan [152] including cases only with a severe phenotype (age of onset ≤25 years and family history of IBD). This scan detected a novel association in the nel-like 1 precursor encoding gene (*NELL1*); and it is known that *PTGER4* is downregulated in *NELL1*-deficient mice. This finding was replicated in independent cohorts, including an UC cohort. In addition, associations for *NOD2, IBD5, IL23R, ATG16L1*, and the 5p13.1 locus were confirmed [152].

The largest CD scan reported to date came from the Wellcome Trust Case Control Consortium (WTCCC) in the UK, which involved the study of seven complex diseases including CD [58]. It identified nine associations with genome-wide significance ($P < 5\times10^{-7}$), including five previously ascribed loci (*NOD2, IL23R, ATG16L1*, 5p13.1 and 10q21.1) and four novel associations: (1) a gene-rich region on chromosome 3p21 (rs9858542; $P = 7.7\times10^{-7}$), a plausible candidate gene, macrophage stimulating 1 (*MST1*), (2) the *IRGM* gene (rs1000013; $P = 5.1\times10^{-8}$), an autophagy-related gene, (3) the NK2 transcription factor-related locus 3 (*NKX2-3*) (rs10883365; $P = 1.4\times10^{-8}$), and (4) the T-cell protein tyrosine phosphatase non-receptor type 2 (*PTPN2*) (rs2542151; $P = 4.6\times10^{-8}$) (of note, this region was also found to be associated with type 1 diabetes and rheumatoid arthritis in this study) [58]. The new findings from the WTCCC were replicated in an independent UK case-control study [159], which also evaluated 31 loci associated at a significance level of $P < 10^{-5}$ from the initial study. The study replicated the associations at four other loci, including two gene deserts on chromosome 1q24 (rs12035082; combined $P = 2.07\times10^{-7}$) and 1q31 (rs10801047; combined $P = 2.83\times10^{-8}$), and *IL12B* (rs6887695; combined $P = 9.21\times10^{-6}$) [159]. Moreover, the WTCCC results were replicated in the number of independent association studies, which indicated association of the several variants with UC, too [153, 164, 167, 169, 171, 178–180].

The scan conducted in a Quebec founder population of French origin involved haplotype-based association analyses within parent-parent-child trios [158]. Replications of the previously described *NOD2, IBD5, IL23R*, and 3p21

loci were reported. Novel findings from this study included a region near the Janus kinase and microtubule interacting protein 1 gene (*JAKMIP1*) gene on chromosome 4p16.1 (rs10003892; $P = 3.7\times10^{-6}$) and regions on chromosome 17q11 (peak SNP, rs4435306; $P = 5.2\times10^{-5}$) and 17q23 (peak SNP, rs6504016; $P = 5.4\times10^{-4}$) [158]. These loci were replicated in a German cohort. This suggests that they are not specific to the Quebec population.

The combined analysis of CD and sarcoidosis GWAS data identified a common susceptibility locus on chromosome 10p12.2 (rs1398024; $P = 4.24\times10^{-6}$) for both chronic inflammatory barrier diseases [160]. The results were verificated and validated in the independent CD and sarcoidosis case-control cohorts. Extensive fine mapping of the 10p12.2 locus pointed to yet unidentified variants in the chromosome 10 open reading frame 67 (*C10orf67*) gene region as the most likely underlying risk factors [160].

Motivated by the need of larger datasets and improved power a genome-wide meta-analysis of the three CD scans [67], combining British [58], North American [59], French and Belgian populations [66], was performed. The study strongly confirmed 11 previously reported loci and provided genome-wide significant evidence for 21 additional loci, including the regions containing *STAT3, JAK2*, inducible T-cell co-stimulator ligand (*ICOSLG*), CDK5 regulatory subunit associated protein 1-like 1 (*CDKAL1*), and intelectin 1 (*ITLN1*) genes [67]. The results had been replicated in an equivalently sized independent panel. The number of independent studies has also witnessed the validity of associations in the panels of CD and UC patients [165, 168, 169, 171, 173, 174, 181–183].

The first combined detailed analysis of the two common IBD was performed in the British population using nonsynonymous SNP array [68]. The initial results were confirmed in two independent case-control panels. A previously unknown susceptibility locus at *ECM1* (rs3737240; $P = 1.3\times10^{-4}$) was identified as determinant of UC. This locus had been replicated in the independent UC association studies [169, 174]. The study also revieled, that UC and CD have several common risk loci (*IL23R, IL12B, HLA, NKX2-3* and *MST1*); whereas, autophagy genes *ATG16L1* and *IRGM*, along with *NOD2*, are specific for CD [68].

The genome wide dissection of UC started from the study in the German population [161]. The initial study identified 20 significant associations that were further genotyped in three independent European case-control replication panels. Significant results across all three replication cohorts were obtained for:

rs3024505 ($P = 1.35\times10^{-12}$) near the 3'untranslated region (UTR) of the *IL10* gene at 1q32; rs12612347 ($P = 2\times10^{-4}$) near the *ARPC2* locus at 2q35; rs9268480 ($P = 6.48\times10^{-18}$), rs9268858 ($P = 2.58\times10^{-12}$) and rs9268877 ($P = 3.15\times10^{-9}$) at the class II/class III junction in the HLA complex at 6p21. The evidence for the modest association of the reported *IL10* SNP with CD has also been shown. Other loci identified included *JAK2*, *IL23R* and 5q13.3 [161]. The associations of newly discovered loci had been confirmed by multiple independent studies [165, 167–171, 174, 183, 184].

A large study analysing only patients with pediatric-onset of IBD had been performed by the North American and Italian study groups [162]. Twelve markers were identified, three of which were previously unreported: two markers on chromosome 20q13, rs2315008 ($P = 6.30\times10^{-8}$) and rs4809330 ($P = 6.95\times10^{-8}$), and one marker on chromosome 21q22, rs2836878 ($P = 6.01\times10^{-8}$). These results were replicated in the indendent cohort collected according to the same definitions as the discovery cohort and in the IBD cohort from the WTCCC study [58]. The study also reported association with previously identified loci *IL23R, NOD2, HLA,* and *TNFSF15* [162]. The study results were also replicated in independent case-control studies [165, 185]. An extension of this study was published in 2009 [166]. Five new regions associated with early-onset IBD susceptibility were identified including 16p11 near the cytokine gene *IL27* (rs8049439; $P = 2.41\times10^{-9}$), 22q12 (rs2412973; $P = 1.55\times10^{-9}$), 10q22 (rs1250550; $P = 5.63\times10^{-9}$), 2q37 (rs4676410; $P = 3.64\times10^{-8}$) and 19q13.11 (rs10500264; $P = 4.26\times10^{-10}$) [166]. The results were replicated in a recent GWA scan [168]. The scan [166] also detected associations at 23 of 32 loci previously implicated in the adult-onset CD (orosomucoid1-like 3 gene (*ORMDL3*), *ICOSLG1*, *etc.*) and at 8 of 17 loci implicated in the adult-onset UC (*IL10, IL20,* etc.), highlighting the close pathogenetic relationship between early- and adult-onset IBD.

In the year 2009 three IBD GWAS studies were published. First, a study in UC patients from North America [163], that identified and replicated in two independent populations signals on chromosomes 1p36 (rs6426833; combined $P = 5.1\times10^{-13}$) and 12q15 (rs1558744; combined $P = 2.5\times10^{-12}$). The possible candidate genes involved in inflammation and immunity in the reported regions include phospholipase A2, group IIE gene (*PLA2G2E*), *IFN-γ, IL26* and *IL22*. In addition, combined genome-wide significant evidence for association was found in a region spanning butyrophilin-like 2 (*BTNL2*) to *HLA-DQB1* and at

the *IL23R* locus [163]. The associations of newly discovered loci had been confirmed in the recent UC meta-analysis [170].

Second, the UC GWAS performed in the Japanese population [164] identified and replicated three new susceptibility loci: a locus on chromosome 13q12 (rs17085007; $P = 6.64\times10^{-8}$), the glycoprotein gene *SLC26A3* (rs2108225; $P = 9.50\times10^{-8}$) and the immunoglobulin receptor gene *FCGR2A* (low affinity immunoglobulin gamma Fc region receptor II-a) (rs1801274; $P = 1.56\times10^{-12}$). *FCGR2A* was reported to be associated with other autoimmune diseases. The previous GWAS associations of chromosome 1p36 and *JAK2* with UC has also been replicated in the frames of this study [164].

Third, the UC GWAS in British cohort was performed as part of the WTCCC2 study of 15 complex disorders and traits [165]. The study showed evidence of association at three new loci, each containing at least one biologically relevant candidate gene, on chromosomes 20q13 (hepatocyte nuclear factor 4 alpha gene (*HNF4A*); $P = 3.2\times10^{-17}$), 16q22 (cadherin 1 (*CDH1*) and *CDH3* genes; $P = 2.8\times10^{-8}$) and 7q31 (laminin, beta 1 (*LAMB1*); $P = 3.0\times10^{-8}$). Of note, *CDH1* has been associated with susceptibility to colorectal cancer, an established complication of longstanding UC. Moreover, the study replicated number of loci previously reported to be associated with UC (*IL23R, IL10,* main histocompatibilty complex (*MHC*)*, IL26, ECM1, ARPC2, MST1, IL12B, JAK2, CARD9, NKX2-3, STAT3, PTPN2, etc*.) [165]. Associations have been confirmed in the recent UC meta-analysis [170].

In April 2010 two new UC GWAS and one CD GWAS were published. The first was GWAS [163] with a subsequent meta-analysis of the current and previously published scan performed by the North American study group [169]. The study population consisted of the North American, Swedish, Italian, and Netherlandish UC patients.The study identified and replicated 13 loci that were significantly associated with UC ($P < 5 \times 10^{-8}$), including the immunoglobulin receptor gene *FCGR2A*, 5p15, 2p16 and *ORMDL3* (orosomucoid1-like 3). The study also confirmed association with previously identified UC susceptibility loci and previously reported CD risk loci [163]. The associations of newly discovered loci had been confirmed in the recent UC meta-analysis [170].

The second published scan performed by the German study group [167] discovered new associations at chromosome 7q22 between karyopherin alpha 7 (*KPNA7*) and SMAD specific E3 ubiquitin protein ligase 1 (*SMURF1*) genes (rs7809799; $P = 2.68\times10^{-5}$) and at chromosome 22q13 in interleukin 17 receptor E-like (*IL17REL*) gene (rs5771069; $P = 4.37\times10^{-5}$) and confirmed in

six replication panels from different regions of Europe (German, British, Belgian, Norvegian, Greek, and Baltic countries (including the Lithuanian/ Latvian patients)) [167]. The validity of associations has been witnessed in the recent UC meta-analysis [170].

The third GWAS was conducted in CD patients by the North American study group [153]. The study identified new associations with genes involved in tight junctions/epithelial integrity (*ARPC1A*), innate immunity (exocyst complex component 2 (*EXOC2*)), dendritic cell biology (cell adhesion molecule (*CADM1*)), macrophage development (monocyte to macrophage differentiation-associated 2 gene (*MMD2*)), transforming growth factor beta (TGF-β) signalling (mitogen-activated protein kinase kinase kinase 7-interacting protein 1 gene (*MAP3K7IP1*)) and galactoside 2-alpha-L-fucosyltransferase 2 gene (*FUT2*) (a physiological trait that regulates gastrointestinal mucosal expression of blood group A and B antigens) (rs602662; $P = 3.4\times10^{-5}$) [153]. Moreover, the study found supportive evidence for 21 out of 40 CD loci identified in in the first CD GWAS meta-analysis [67].

In the end of November 2010 the second meta-analysis of CD has been published [19]. The authors undertook a meta-analysis of six CD GWAS [58, 66, 151, 153, 166]. More than 30 new susceptibility loci meeting genome-wide significance ($P < 5\times10^{-8}$) were identified. Moreover, *in silico* analyses highlighted particular genes within these loci and implicated functionally interesting candidate genes including *SMAD3*, *ERAP2*, *IL10*, *IL2RA*, *TYK2*, *FUT2*, *DNMT3A*, *DENND1B*, *BACH2* and *TAGAP*. Combined with previously confirmed loci, this study mounted the number of the identified distinct CD loci to 71 [19].

The first meta-analysis of UC GWAS datasets has been reported recently [170]. The datasets for meta-analysis were derived from six index GWA scans from Cedars-Sinai [169], Germany [161, 167], Sweden [169], the Early onset IBD consortium [162, 166], the NIDDK IBD Genetics Consortium [163] and the WTCCC2 [165]. The study identified 29 new risk loci ($P < 5\times10^{-8}$), increasing the number of UC-associated loci to 47. Afterwards, the potentially plausible functional annotations for the associated regions had been determined using GRAIL, expression quantitative trait loci data and correlations with non-synonymous SNPs (*e.g.*, *IL1R2* (2q11), *PRDM1* (6q21), *IRF5* (7q32), *LSP1* (11q15), *GNA12* (7p22), *IL8RA-IL8RB* (2q35), *TNFRSF9* (1p36), *DAP* (5p15), *IL7R* (5p13), *IL12B* (5q33), *IRF5* (7q32), *JAK2* (9p24)). This study had more

than doubled the number of the confirmed UC risk loci (*i.e.*, 47 genes/loci) [170].

Although GWA studies have allowed an unprecedented rapid unraveling of the genetic basis of IBD, however there will be much more follow-up work needed in this field. First, to distinguish true positive associations from spurious ones, replication studies are essential, preferably in independent populations using large sample sizes with matched controls and disease phenotypes comparable with those used in the initial studies. Second, as most of the variants identified so far are tagging SNPs that only highlight a certain region, it will be essential to investigate the functional consequences (through deep sequencing and thorough functional studies) of polymorphisms in these loci. Finally, to characterize the allelic architecture of IBD it will be necessary to improve the genotyping technology and methodology in order to assess copy number variations, rare variants, structural variants, SNP-SNP/gene-gene/gene-environment interactions, epigenetic modifications, population specific variants and even individual genetic profiles and explain the missing heritability of IBD.

2. MATERIALS AND METHODS

2.1. Patients

The study included 131 unrelated patients with CD, 447 with UC and 1154 ethnically, age and sex-matched healthy control individuals. All study participants were of Caucasian ethnicity. The recruitment of the study individuals was performed at six Lithuanian hospitals: Hospital of Lithuanian University of Health Sciences Kaunas Clinics, Vilnius University Hospital at Santariskes, M. Marcinkevicius Hospital, Klaipeda University Hospital, Klaipeda Seamen Hospital, Panevezys District Hospital, Siauliai District Hospital, and three Latvian hospitals in Riga: P. Stradin University Hospital, Riga Seamen Medical Center, "Linezers" hospital, during the period 2003 till 2009. The Lithuanian control individuals were recruited from the National Blood Center, Blood Donor Center and Latvian – from hospitals participating in the patient recruitment during the period 2008 to 2009. The study participants were well characterized: all relevant demographic and clinical characteristics had been surveyed using standard questionnaires (Appendix, Patient's questionnaire). Written informed consent from all participants and approval of the Lithuanian Bioethics Committee (Protocol No. 2/2008) and Riga Biomedical Research Ethics Committee (Protocol No. 290910-8L) was obtained. The diagnosis of either CD or UC was based on standard clinical, endoscopic, radiological and histological criteria [186]. Patients' demographic and phenotypic details are summarized in Table 2.1.1. The clinical characteristics provided in the table are given according to the Montreal classification [27].

Table 2.1.1. *Summary of clinical and demographic characteristics of the IBD patients*

Characteristics	**CD (n=131)**	**UC (n=447)**	**Controls (n=1154)**
Gender (male/female)	66/65	222/225	564/590
Age (years ± SD)	38.4±16.6	44.4±16.5	40.2±12.7
Age at diagnosis (years ± SD)	34.9±16.1	38.4±15.8	
Familial IBD	3 (2.3%)	15 (3.6%)	
Surgery treatment	25 (19.1%)	28 (6.3%)	

Table 2.1.1 continued

Characteristics	CD (n=131)	UC (n=447)	Controls (n=1154)
Disease extension in UC			
Proctitis, E1	–	89 (19.9%)	
Left-sided colitis, E2	–	241 (53.9%)	
Extensive colitis, E3	–	117 (26.2%)	
Disease localization in CD			
Terminal ileum, L1	45 (34.3%)	–	
Colon, L2	36 (27.5%)	–	
Ileocolon, L3	49 (37.4%)	–	
Upper GI, L4	1 (0.8%)	–	
Terminal ileum + Upper GI, L1+L4	3 (2.3%)	–	
Colon + Upper GI, L2+L4	1 (0.8%)	–	
Ileocolon + Upper GI, L3+L4	2 (1.5%)	–	
Disease Behavior in CD			
Non-stricturing, non-penetrating, B1	105 (80.1%)	–	
Stricturing, B2	11 (8.4%)	–	
Penetrating, B3	15 (11.5%)	–	
Perianal disease (isolated), B4	–	–	
Non-stricturing, non-penetrating + Perianal, B1p	15 (11.5%)	–	
Stricturing + Perianal, B2p	2 (1.5%)	–	
Penetrating + Perianal, B3p	5 (3.8%)	–	
Extraintestinal manifestations			
Joints	32 (24.4%)	44 (9.8%)	
Cutaneus	11 (8.4%)	19 (4.3%)	
Ocular	3 (2.3%)	2 (0.4%)	
Hepatobiliary	2 (1.5%)	17 (3.8%)	

IBD – inflammatory bowel disease; CD – Crohn's disease; UC – ulcerative colitis;
SD – standart deviation.

2.2. Sample preparation

After recruitment of individuals, genomic DNA (gDNA) was isolated from the donated blood samples. The layout of samples on the 96 well microtiterplates was designed. The whole genome amplification (WGA) was performed. DNA samples were quality checked on an agarose gel (1.5%) after DNA extraction and WGA procedures. Amplified products were arrayed on 96 well microtiterplates, afterwards merged into a 384 well microtiterplates, *i.e.*, 4 x 96 well microtiterplates in order to increase the throughput for downstream processes.

2.2.1. DNA extraction from blood

After recruitment, donated blood samples were stored at –80°C until procedure. gDNA was extracted from EDTA whole blood samples, using the commercially available Invisorb® Blood Universal Kit for DNA isolation from whole blood, and automated DNA purification system Autopure LS®, according to the manufacturer's protocol/instructions.

Reagents:
Invisorb® Blood Universal Kit (Invitek, Berlin, Germany);
100% Ethanol (Merck, Darmstadt, Germany);
Tris (Merck, Darmstad, Germany);
EDTA (Sigma, Munchen, Germany);
Autopure RBC Lysis Solution (Qiagen, Hilden, Germany);
Autopure Cell Lysis Solution (Qiagen, Hilden, Germany);
Autopure Precipitation Solution (Qiagen, Hilden, Germany);
Autopure DNA Hydration Solution (Qiagen, Hilden, Germany);
Autopure 100% Isopropanol (Qiagen, Hilden, Germany);
Autopure 70% Ethanol (Qiagen, Hilden, Germany);
Autopure Qubes® E or D (Qiagen, Hilden, Germany).

Equipment:
10/100/1000 µl single-channel pipetes (Eppendorf, Hamburg, Germany);
GFL 1086 shaking waterbath (GFL, Burgwedel, Germany);
Heraeus Multifuge 3S-R (Kendro, Hanau, Germany);
Heraeus Multifuge 3S+ (Kendro, Hanau, Germany);
Heraeus Biofuge fresco (Kendro, Hanau, Germany);

Vortex-GENIE 2 G-560E (Scientific Industries, Bohemia, NY, USA);
Autopure LS® (Qiagen, Hilden, Germany);
Ultra Clear TWF (SGWasseraufbereitung und Regenerierstation GmbH, Barsbuttel, Germany).

I. Workflow of DNA extraction from compromised blood samples using "Invisorb® Blood Universal Kit":

A. Lysis of Erythrocytes:

1. 9 ml of blood were incubated with 25 ml of cold elution buffer (Buffer EL, provided in the kit) for 10 min on ice.
2. The suspension was centrifuged for 5 min at 2,000 g and the supernatant was carefully discarded.
3. The washing step was repeated with the same volume of Buffer EL and centrifugation (5 min at 2,000 g) until the leucocyte containing pellet was free of haem.
4. After the last centrifugation the supernatant was carefully discarded. The residual fluid was removed by dabbing the tube on a paper tissue (one had to be careful not to decant the cell pellet).

B. Lysis of lymphocytes:

1. The pellet was resuspended in 5 ml of lysis buffer HL (provided in the kit) and 50 µl of Proteinase K (provided in the kit).
2. The tubes were incubated at 60°C for 15 min in a water bath under continuous shaking (95 turns/min) to increase the lysis efficiency. This step leads to the lysis of the leukocytes nuclei and release of DNA into the suspension.

C. DNA precipitation:

1. 5 ml of precipitation solution (provided in the kit) were added to the lysate.
2. The tubes were carefully inverted several times, *i.e.*, until the DNA flakes became visible.
3. The precipitated DNA was transferred to 2 ml reaction tubes containing 1 ml of 70% ethanol, rinsed by vortexing and subsequently centrifuged for 2 min at 13000 rpm.
4. The ethanol was removed by carefully inverting the tube and dabbing it on paper tissue.

5. The pellet was dried at room temperature until all traces of ethanol have evaporated. The time for drying had to be as short as possible (approx. 10 min).

D. Dissolution of gDNA:

1. The purified gDNA was resuspended in 1.2 ml of 1× TE (tris-EDTA) buffer and incubated at 60°C for at least 1 h in a water bath under continuous shaking or at room temperature overnight.
2. The gDNA samples were stored at +4°C for short periods or at –20°C for long periods.

II. Workflow of automated purification of DNA from compromised blood samples on the Autopure LS®:

The Autopure LS® is the DNA purification robotic system allowing the extraction of DNA from different quantities and different types of biological material. The system has a user friendly instrument software and barcode system protecting from the possible sample mix up. The system simultaneously can run upto 16 samples.

Steps performed by the Autopure LS®:

A. RBC lysis

1. The input and output cap bar codes were scanned and verified. The tubes were weighed to check that input tubes contained samples and that output tubes were empty.
2. 30–35 ml of Autopure RBC Lysis Solution (Reagent 1) was dispensed into each input tube. As the system uses Reagent 1 to balance the tubes before centrifugation, the amount dispensed into each tube varied depending on the initial sample volume. The total volume of sample and Reagent 1 was 40 ml.
3. The samples in Autopure RBC Lysis Solution were incubated for 5 min 30 s to lyse the red blood cells. The samples were rotated gently to mix during incubation.
4. The samples were centrifuged at 3,000 g for 2 min to pellet the white blood cells.
5. After centrifugation, the supernatant from step 4 were poured into the waste tray.

B. Cell lysis and protein precipitation

1. 4 ml of Autopure Precipitation Solution (Reagent 3) was dispensed into each input tube.
2. 10 ml Autopure Cell Lysis Solution (Reagent 2) was dispensed into each input tube.
3. The samples were mixed vigorously for 2 min to lyse the cells and precipitate the proteins.
4. The samples were centrifuged at 3,000 x g for 5 min. The precipitated proteins formed a tight pellet at the bottom of the input tube.
5. During the centrifugation in step 4, the instrument dispensed 12 ml Autopure 100% Isopropanol (Reagent 4) into output tubes in Row D (if running 16 samples).
6. The DNA-containing supernatant from step 4 was poured into the output tubes that contain Autopure 100% Isopropanol.

C. DNA precipitation

1. The output tubes were gently rotated for 50 times to precipitate the DNA.
2. The samples were centrifuges at 3,000 x g for 10 min to pellet the DNA.
3. The isopropanol supernatant was poured into the waste tray. The output tubes were inverted for 1 min to evaporate any remaining alcohol.

D. DNA wash

1. 12 ml Autopure 70% Ethanol (Reagent 5) were dispensed into the output tubes.
2. The samples were centrifuged at 3,000 x g for 10 min to pellet the DNA.
3. The ethanol supernatant was poured into the waste tray. The output tubes were inverted for 1 min to evaporate any remaining alcohol.

E. DNA hydration

1. 1.2 ml volume of Autopure DNA Hydration Solution (Reagent 6) was dispensed into the output tubes to rehydrate the DNA.
2. Message informing the user that the protocol run had finished was displayed.
3. After removal of the purified DNA from the instrument, it was incubated at 65°C for 1–2 h to dissolve the DNA.
4. Afterwards, DNA was incubated at room temperature (15–25°C) overnight with gentle shaking.

5. Samples were transferred to new 2 ml storage tubes. The gDNA samples were stored at +4°C for short periods or at –20°C for long periods.

2.2.2. Plate design

For genotyping, 89 DNA samples were arranged in a 96 well format according to a pre-defined plate layout (Fig. 2.2.2.1). Individuals with the same diagnosis were kept on the same plate. Seven wells were used for internal controls and quality control, such as three empty wells (no template controls), and four positive controls, so called CEPH controls (Fondation Jean Dausset Cenre d'Etude du Polymorphisme Humain, Paris, France). No template 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 plate was labeled with a unique plate name for database storage to allow unmistakable identification. SNPlex™ plates received the prefix "X", *e.g.*, XG01.

row	1	2	3	4	5	6	7	8	9	10	11	12
A	CEPH	BALT1006	BALT1014	BALT1020	BALT1028	BALT1036	BALT1044	BALT1052	BALT1060	BALT1068	BALT1075	BALT1083
B	CEPH	BALT1007	BALT1015	BALT1021	BALT1029	BALT1037	BALT1045	BALT1053	BALT1061	BALT1069	BALT1076	BALT1084
C	CEPH	BALT1008	CEPH	BALT1022	BALT1030	BALT1038	BALT1046	BALT1054	BALT1062	BALT1070	BALT1077	BALT1085
D	BALT1001	BALT1009	EMPTY	BALT1023	BALT1031	BALT1039	BALT1047	BALT1055	BALT1063	EMPTY	BALT1078	BALT1086
E	BALT1002	BALT1010	BALT1016	BALT1024	BALT1032	BALT1040	BALT1048	BALT1056	BALT1064	BALT1071	BALT1079	BALT1087
F	BALT1003	BALT1011	BALT1017	BALT1025	BALT1033	BALT1041	BALT1049	BALT1057	BALT1065	BALT1072	BALT1080	BALT1088
G	BALT1004	BALT1012	BALT1018	BALT1026	BALT1034	BALT1042	BALT1050	BALT1058	BALT1066	BALT1073	BALT1081	BALT1089
H	BALT1005	BALT1013	BALT1019	BALT1027	BALT1035	BALT1043	BALT1051	BALT1059	BALT1067	BALT1074	BALT1082	EMPTY

Fig. 2.2.2.1. *Plate layout*
Wells D3 and D10 were used as negative controls for TaqMan® genotyping and for allelic ladder in case of SNPlex™.

2.2.3. Whole genome amplification

WGA is an *in vitro* method that is used to amplify gDNA samples and generate amplified DNA for further molecular genetic analyses [187–189]. In many large genetic studies the amount of available high-quality DNA can be one of the limiting criteria for selecting samples for study. WGA is a useful method for production of sufficient DNA quantity from samples with limited

DNA content [190]. The Genomiphi™ V2 DNA Amphification Kit, based on multiple displacement amplification (MDA) method (Fig. 2.2.3.1) [187], was used for amplification of DNA in this study.

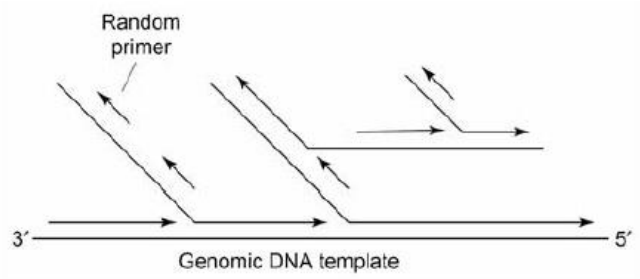

Fig. 2.2.3.1. *Multiple displacement amplification reaction* [191]
DNA synthesis is primed by random hexamers. Exponential amplification occurs by a 'hyperbranching' mechanism. Unlike PCR, which requires thermal cycling to repeatedly melt template and anneal primers, the Phi29 DNA polymerase acts at +30°C to concurrently extend primers as it displaces downstream DNA products.

Reagents:
Genomiphi™ V2 DNA Amphification Kit (GE Healthcare UK Limited, Buckinghamshire, UK);
Tris (Merck, Darmstad, Germany);
EDTA (Sigma, Munchen, Germany).

Equipment:
10/100/1000 µl single-channel pipetes (Eppendorf, Hamburg, Germany);
GeneAmp® PCR System 9700 (Applied Biosystems Inc., Foster City, CA, USA);
Heraeus Multifuge 3S-R (Kendro, Hanau, Germany);
Heraeus Biofuge fresco (Kendro, Hanau, Germany);
Vortex-GENIE 2 G-560E (Scientific Industries, Bohemia, NY, USA);
Hydra 96 Robbins Scientific (Dunn Labortechnik, Asbach, Germany);
Hydra 384 Robbins Scientific (Dunn Labortechnik, Asbach, Germany);
Te-MO (Tecan, Deutschland GmbH, Crailsheim, Germany);
Tecan Genesis RSP 150 (Tecan, Deutschland GmbH, Crailsheim, Germany);
PCR chamber (Ba-RO® Technology, Leichlingen, Germany);
Heraeus 3 incubator (Kendro, Hanau, Germany);
Platesealer ALPS-300 (ABgene, Epsom, UK);

Ultra Clear TWF (SG Wasseraufbereitung und Regenerierstation GmbH, Barsbuttel, Germany).

All steps were carried out according to the manufacturer instructions/kit's protocol. *In brief:*

1. 9 µl Sample Buffer (provided in the kit) were mixed with 1 µl of 10 ng template DNA in the 96 well paltes.
2. In order to denature template DNA the samples were heated to 95°C for 3 min then cooled to 4°C on ice.
3. The master mix, *i.e.*, combination of 9 µl of Reaction Buffer and 1 µl of Enzyme Mix, was prepared on ice.
4. 10 µl of prepared master mix (Step 3) were transferred to the cooled sample (Step 2). The procedure was performed on ice.
5. The samples were incubated at 30°C for 1.5 hour for DNA amplification. The generated fragments ranged between 10 and 100 kb.
6. In order to inactivate the Phi29 DNA polymerase enzyme the samples were heated to 65°C for 10 min then cooled to 4°C.

After the WGA the 20 µl (~5 µg) of reaction volume was diluted in the following way:

1. 1:5 with 1× TE-buffer, final volume of 100 µl (~50 ng/µl).
2. 100 µl were split (10 µl and 90 µl) into two fresh 96 well plates. One plate was used for SNPlex™ (90 µl) and the other for TaqMan® (10 µl) plate production.
3. In case of SNPlex™, the WGA-DNA was fragmented for 5 minutes according to the SNPlex™ protocol and then diluted 1:2 with 1× TE-buffer to a final volume of 180 µl (~25 ng/µl).
4. For TaqMan®, the 10 µl WGA-DNA were further diluted with 1× TE-buffer 1:80 to a final volume of 800 µl (~0.625 ng/µl).

Four 96 deepwell microtiter plates were then merged to one 384 deepwell plate using 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 into fresh 384 polymerase chain reaction (PCR) plates. For TaqMan®, the plates were dried down at 60°C for one hour and subsequently sealed. In case of SNPlex™, the plates were left to dry overnight. Dried plates (ready-to-use for genotyping) were sealed. Each plate received a unique barcode label for database tracking.

2.2.4. Agarose gel electrophoresis

Agarose gel electrophoresis was used for quality control of gDNA after DNA purification and WGA steps.

Reagents:
HyperLadder I (Bioline, Luckenwalde, Germany);
Agarose (Biozym Scientific GmbH, Hessisch Oldendorf, Germany);
Bromphenol blue (Sigma, Munchen, Germany);
Xylene Cyanol FF (Sigma, München, Germany);
Glycerol (Sigma, München, Germany);
Ethidium Bromide solution (10 mg/ml) (Invitrogen, Karlsruhe, Germany);
Rotiphorese® 10× TBE (Tris-borate-EDTA) Buffer (ROTH, Karlsruhe, Germany).

Equipment:
10/100/1000 µl single-channel pipetes (Eppendorf, Hamburg, Germany);
Gel Doc XR (Bio-Rad, München, 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);
Heraeus Multifuge 3S-R (Kendro, Hanau, Germany);
Heraeus Biofuge fresco (Kendro, Hanau, Germany);
Vortex-GENIE 2 G-560E (Scientific Industries, Bohemia, NY, USA);

Preparation and running of agarose gel:

1. For quality control of gDNA 1.5 % agarose gel was used.
2. 1× TBE buffer was used as a running buffer.
3. The buffer and agarose mixture was boiled in a microwave until it became a clear solution.
4. The gel was cooled down to approx. 60°C. After the addition of ethidiumbromide (10 mg/ml), the gel was poured into the casting device, and left for approx. 30 min until polymerization had finished.
5. 9 µl of 2× DNA-loading buffer (0.25% bromphenol blue + 0.25% xylene cyanol FF + 30% glycerol in water) was added to the 1 µl of sample.
6. For size and approximate quantity of the DNA molecule determination 100 bp DNA ladder was added on the same gel.

7. Fragments were separated in horizontal gel chambers electrophoretically at 110 V, 400 mA for 60 min until complete band seperation.
8. The gDNA was evaluated under UV-light illumination with the Bio-Rad Gel Doc XR gel documentation system.

2.3. Genotyping

Genotyping refers to the process of determining the genotype of an individual with a biological assay. Two genotyping methods were used in this study: TaqMan®, SNPlex™.

2.3.1. SNP selection

Selection of 79 SNPs (Appendix , Table 1) was made based on the data from the original GWAS: (1) CD – (A) Franke *et al.*, 2007 [152], (B) WTCCC/Parkes *et al.*, 2007 [58, 159], and (C) Barrett *et al.*, 2008 [67]; (2) UC – (D) Franke *et al.*, 2008 [161], (E) Silverberg *et al.*, 2009 [163], and (F) Franke *et al.*, 2010 [167] that presented new associations, as well as replicated SNPs from previous studies *(e.g., NOD2, DLG5, IL23R, SLC22A4, SLC22A5, TNFSF15, PTGER4, MST1,* and *ATG16L1).*

2.3.2. SNPlex™

The SNPlex™ Genotyping System enables the simultaneous genotyping of up to 48 SNPs against a single biological sample. The SNPlex Genotyping System is based on the oligonucleotide ligation/PCR assay (OLA/PCR) with a universal ZipChute™ probe detection for high throughput SNP genotyping [192]. Fluorescently labeled ZipChute™ probes are hybridized to complementary ZipCode™ sequences that are part of genotype specific amplicons. These ZipChute™ probes are eluted and detected by electrophoretic separation on Applied Biosystems 3730 or 3730x*l*DNA Analyzers.

Assays for the SNPlex™ Genotyping System were designed by Applied Biosystem's automated high-throughput pipeline (assays used in this study are presented in the Appendix Table 2). The pipeline combines SNP-specific assays into compatible multiplex pools.

These steps include [192]:

1. Screening the SNP context sequences against the target genome to avoid designing assays for SNPs in repetitive or duplicated genomic regions that would lead to low specificity (this step can be omitted for organisms that do not have an assembled genome).
2. Selection and design of the SNP-specific ligation probes by applying assay and probe manufacturing rules to select the more suitable strand and probe sequence.
3. Assignment of ZipCode sequences to each allele-specific oligo (ASO) probe of an assay.
4. 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:

1. Three SNP-specific ligation probes:
   - Two of the probes are 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.
2. Three linkers for each SNP:
   - Two of the linkers anneal to the two ASOs. These linkers contain: (1) a PCR primer sequence corresponding to the universal forward primer (UA sequence); (2) a partial ZipCode™ sequence
   - The third linker anneals to the LSO and has a universal sequence that is compatible with all LSOs. The sequence includes a partial binding site for a universal reverse primer.

In a 48-plex reaction, there are: 96 ASOs and 48 LSOs (for a total of 144 SNP-specific oligos), 96 ASO linkers and a single LSO linker (for a total of 97 linkers). This pool confers genotyping specificity to the SNPlex™ System assay. All other reagents are universal and not SNP specific.

Reagents:

SNPlex™ System Core Kit (Applied Biosystems, Foster City, USA)

Equipment:

10/100/1000 µl single-channel pipetes (Eppendorf, Hamburg, Germany);
Heraeus Multifuge 3S-R (Kendro, Hanau, Germany);
Heraeus Biofuge fresco (Kendro, Hanau, Germany);
Micro Centrifuge (Roth, Karlsruhe, Germany);
TiMix Control incl. TH15 hood (Edmund Buhler Labortechnik, Germany);
Vortex-GENIE 2 G-560E (Scientific Industries, Bohemia, USA);
GeneAmp® PCR System 9700 (Applied Biosystems Inc., Foster City, USA);
Te-MO (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 Workstation 150 (Tecan, Deutschland GmbH, Crailsheim, Germany);
Tecan Genesis Workstation 200 (Tecan, Deutschland GmbH, Crailsheim, Germany);
3730xl DNA Analyzer (Applied Biosystems Inc., Foster City, USA).

The assay workflow for the SNPlex Genotyping System involves eight steps (Fig. 2.3.2.1) (note: all necessary master mixes were prepared manually, while all other pipetting steps were carried out on four different TECAN multipipetting robots):

Day 1 – OLA laboratory.

A. Phosphorylation and ligation of the probes (OLA)

During the "OLA reaction", which is the allele-discriminating step, the genotype information was encoded by highly specific ligation of the ASO probes to the LSO probes using fragmented WGA amplified gDNA (100–150 ng per well, *i.e.*, 2–3 ng per assay) as the target. ASO and LSO linkers connect to the corresponding ASO and LSO probes. Only 384 well plates were used throughout the process.

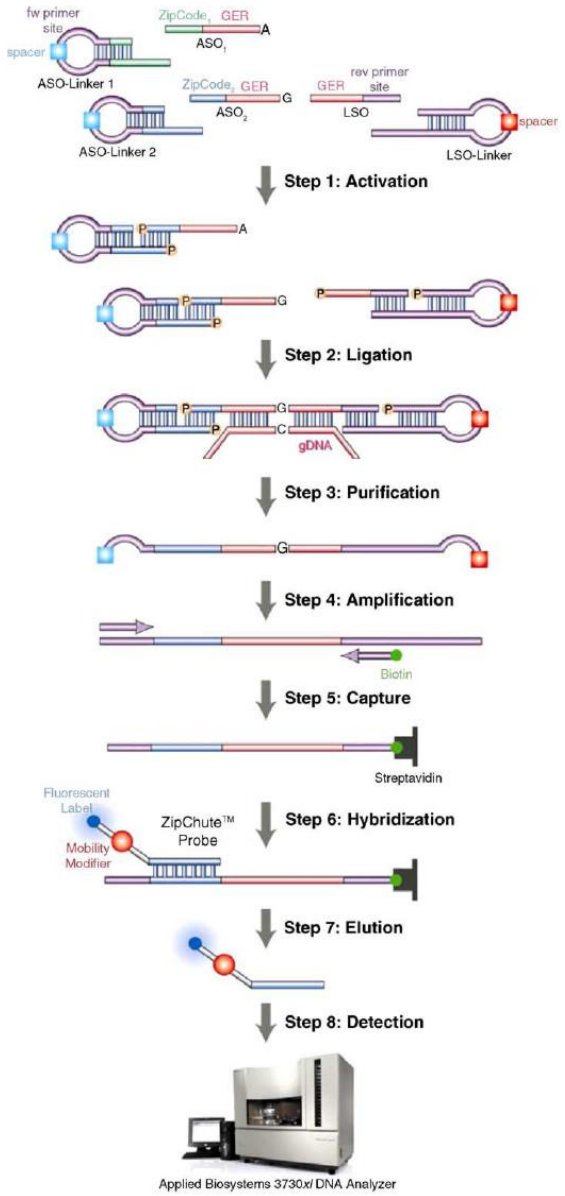

Fig. 2.3.2.1 *SNPlex Genotyping System workflow* [192].
The key step is the oligo ligation assay, which is the allele-discrimnating step. For a description see text below and for more details see the protocol from Applied Biosystems.

Procedure:

1. An OLA reaction mix was prepared by scaling the volumes (Table 2.3.2.1) to the desired number of OLA reactions.

Table 2.3.2.1. *OLA reaction mix*

Reagent	**Volume (µl) for 1 reaction**
OLA Master Mix (2×) SNPlex System	2.5
SNPlex Universal linkers 48-plex	0.05
dATP (10×)	0.05
SNPlex Ligation Probes	0.05
Nuclease free H_2O	2.35
Total	**5**

2. 5 µl of OLA reaction mix was added into each well by a TECAN multipipetting robot. The mix was not added to the allelic ladder wells (see Fig. 2.3.2.2).
3. The reaction was facilitated under temperature controlled conditions (Table 2.3.2.2) in the thermocycler. The ligation of the linkers and the ASO and LSO probes that have annealed to the gDNA target was performed.

Table 2.3.2.2. *Thermal-cycling program of the OLA reactions*

Step	**Temperature**	**Time**	**Cycle(s)**
1	48ºC	30 min	1
2	90ºC	20 min	1
3	94ºC	15 s	25
4	60ºC	30 s	
5	51ºC	30 s	
6	99ºC	10 min	1
7	4ºC	∞	

The OLA Master Mix contains enzymes to promote phosphorylation of probes and linkers, uracil-N-glycosylase to degrade contaminating amplicons, and ligase. All steps were carried out sequentially during the thermocycling protocol.

Day 2 – OLA laboratory.

B. Purification of the ligation product by exonuclease digestion.

After the ligation reaction, unligated and incompletely ligated oligonucleotides, as well as the gDNA templates, had to be removed ("purification step"). This reduces the background noise of the signal.

Procedure:

1. 2× Exonuclease master mix was prepared on ice (directly before use) by scaling the volumes (Table 2.3.2.3) to the desired number of OLA reactions.

Table 2.3.2.3. *Exonuclease master mix*

Reagent	**Volume (µl) for 1 reaction**
SNPlex™ exonuclease buffer	0.5
SNPlex™ λ exonuclease	0.2
SNPlex™ exonuclease I	0.1
Nuclease free H_2O	4.2
Total	**5**

2. 5 µl of 2× Exonuclease master mix was pipetted into each well of the OLA reaction by a TECAN multipipetting robot. The plate was covered, then vortexed and spined briefly.
3. The enzymes, under temperature controlled conditions (Table 2.3.2.4) in the thermocycler, facilitated the purification of the ligated OLA reaction products.

Table 2.3.2.4. *Thermal-cycling program of the purification reaction*

Step	**Temperature**	**Time**	**Cycle(s)**
1	37ºC	90 min	1
2	80ºC	10 min	1
3	4ºC	∞	

C. PCR amplification of the ligated and exonuclease digested products.

Following dilution of the digested material, an aliquot was subjected to a PCR reaction with two universal primers, one of which is biotinylated ("PCR setup").

Procedure:

1. The 10 µl Exonuclease-treated ligation reactions were diluted with 15 µl of nuclease-free water to each well and mixed to combine.
2. A PCR master mix was prepared by scaling the volumes (Table 2.3.2.5) to the desired number of PCR reactions.

Table 2.3.2.5. *PCR master mix*

Reagent	**Volume (µl) for 1 reaction**
SNPlex amplification master mix (2×)	5
SNPlex amplification primers (20×)	0.5
Nuclease free H_2O	2.5
Total	**8.00**

3. The following reaction mix was dispensed into each well: 8 µl PCR master mix and 2 µl diluted OLA reaction product by a TECAN multipipetting robot. The plates were sealed, vortexed briefly, and then spined down.
4. The diluted, exonucelase digested OLA reaction products were amplified (program Table 2.3.2.6) using the universal primers. The resulting product was a double-stranded amplicon with one biotinylated strand.

Table 2.3.2.6. *Thermal-cycling program of the PCR reaction*

Step	**Temperature**	**Time**	**Cycle(s)**
1	95ºC	10 min	1
2	95ºC	15 s	30
3	70ºC	1 min	
4	4ºC	∞	

Day 3 – PCR laboratory and capillary electrophoresis (CE)

By the manufacturer it is recommended to perform all subsequent steps in a different laboratory in order to avoid amplicon contamination.

D. Capture of biotinylated amplicons on streptavidin-coated plates, and removal of the unbound strand.

Procedure:

1. The wells of the SNPlex Hybridization Plate were washed once with 100 µl of Wash Buffer diluted 1:10 with deionized water.
2. 0.009 µl positive hybridization control to 17.491 µl Binding Buffer (containing streptavidin) was added.

3. 17.5 µl of the Binding Buffer containing positive hybridization control to the SNPlex Hybridization Plate was added by a TECAN multipipetting robot.
4. 3 µl of PCR product (section C) was transfered to the SNPlex Hybridization Plate by a TECAN multipipetting robot and then mixed. The plate was covered and incubated at room temperature for 60 min on a rotary shaker.
5. The plate was briefly spun and then 50 µl of 0.1 N NaOH was added. The plate was incubated for 30 min at room temperature on a rotary shaker.
6. The plate was briefly spun, the supernatant was removed, and then washed three times with 100 µl Wash Buffer diluted 1:10 with deionized water.

E. Hybridization of the universal set of ZipChute™ probes to the complementary ZipCode™ product sequences on the captured PCR strand.

Upon removal of the non-biotinylated amplicon strands, a mixture of 102 pre-optimized, universal ZipChute™ probes was 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 assay. 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 mobility modifiers). The ZipChute™ probes are eluted after stringent washing and detected by electrophoretic separation on Applied Biosystems 3730x*l*DNA Analyzers.

Procedure:

1. A hybridization master mix was prepared by scaling the volumes (Table 2.3.2.7) to the desired number of samples.

Table 2.3.2.7. *ZipChute™ Hybridisation master mix*

Reagent	**Volume (µl) for 1 reaction**
ZipChute mix	0.05
SNPlex denaturant	11.25
SNPlex ZipChute dilution buffer	13.7
Total	**25**

2. 25 µl of the hybridization master mix to was added into each well of the SNPlex Hybridization Plate by a TECAN multipipetting robot, and then covered.

3. The plate was incubated for 60–75 min at 37°C on a rotary shaker. During incubation the plate was protected from the bright light.

F. Release of specifically hybridized ZipChute™ probes.

To establish a sizing calibration curve that is used to identify ZipChute™ probes, a fluorescently labeled SNPlex Size Standard was pipetted into each well. Eleven size-standard (orange) peaks appear in each lane of the electropherogram. The plate was incubated at 37°C to release the ZipChute™ probes from the biotinylated strand.

Procedure:

1. A Sample Loading Mix was prepared by scaling the volumes (Table 2.3.2.8) to the desired number of samples.

Table 2.3.2.8. *Sample loading master mix*

Reagent	**Volume (µl) for 1 reaction**
SNPlex size standart	0.54
SNPlex sample loading reagent	16.96
Total	**17.5**

2. The plate was briefly spun, the supernatant was removed, and then washed three times with 100 µl Wash Buffer diluted 1:10 with deionized water. After the final wash, the plate was spun upside down at 1000 rpm (rotations per minute) for 60 sec on a stack of paper towels.
3. 17.5 µl of Sample Loading Mix was added into each well and mix by a TECAN multipipetting robot.
4. The plate was covered and incubated at 37°C for 30 min on a rotary shaker.

G. Preparation of samples for electrophoresis.

Procedure:

1. The hybridization plates were removed from the oven (37°C) and briefly spun to collect the fluid at the bottom of the wells.
2. New reaction plates were labeled.
3. 7.5 µl of fluid from each well of the hybridization plate was transferred into the wells of the new 384-well microtiterplate by a TECAN multi-pipetting robot.

4. Into 8 wells of each plate an allelic ladder of the ZipChute™ probes labeled with FAM and NED dyes was dispensed (Fig. 2.3.2.2).

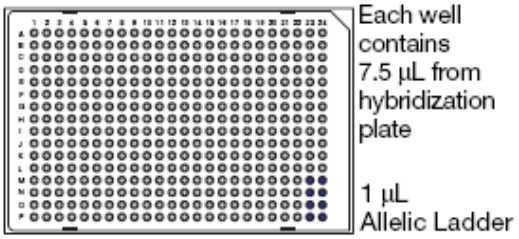

Fig. 2.3.2.2. *The final sample and alleles ladder (in dark blue) layout in the 384-well microtiterplate*

5. The plates were briefly spun to remove air bubbles trapped at the bottom of the wells.

H. Detection of fluorescent ZipChute™ probes by CE.

The plates were loaded onto the 96-capillary 3730*xl* analyzer to generate sample files. The data analysis was conducted using GeneMapper® Analysis Software v3.5.1. The automated allele calling of all plates has been used. Auto-calls were manually inspected for faulty genotype assignments before the data was exported from GeneMapper® and then imported into the in-house database (Fig. 2.3.2.3).

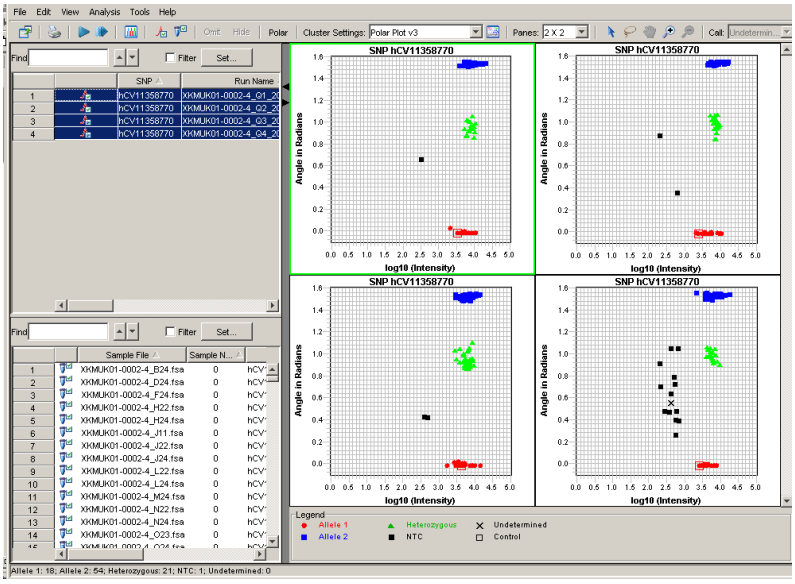

Fig. 2.3.2.3. *SNPlex™ genotype cluster plots*
Homozygotes for allele 1 are shown in red, heterozygotes in green (both dyes are measured), and homozygotes for allele 2 in blue. The black squares close to the origin are the negative controls, which control for potential contamination, and black crosses are undefined genotypes that were excluded from subsequent analyses.

2.3.3. TaqMan®

For genotyping SNPs that did not work with SNPlex™, the robust genotyping method TaqMan® was chosen. TaqMan® is a single-tube PCR assay [192–196] that exploits the 5'exonuclease activity of DNA polymerase.

The TaqMan® SNP Genotyping 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 with a minor groove binder at the 3' end [197] (Fig. 2.3.3.1). 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® and with the quencher TAMRA™ (6-carboxytetramethylrhodamine, succinimidyl ester). The passive reference dye

ROX (6-carboxy-X-rhodamine, succinimidyl ester) was included in every well for normalization.

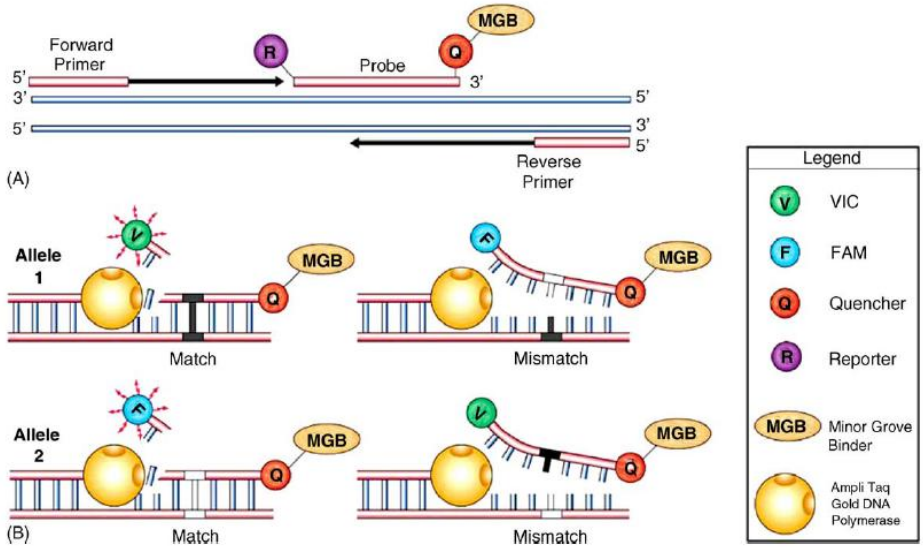

Fig. 2.3.3.1. *TaqMan® assay overview* [192]
(A) Probe binding and primer extension in a TaqMan® SNP Genotyping Assay. (B) Allelic discrimination is achieved by the selective annealing of matching probe and template sequences, which generates an allele-specific (fluorescent dye-specific) signal.

The genotyping assays used in this study were Assays-on-Demand (AoD), a pre-designed and validated assay format offered by the manufacturer (see Appendix Table 3).

Reagents:
AmpliTaq Gold® with GeneAmp 10× PCR Buffer II & $MgCl_2$ solution (Applied Biosystems, Foster City, CA, USA);
TaqMan® Universal PCR Master Mix (Applied Biosystems, Foster City, CA, USA);
20x SNP genotyping assay mix (Applied Biosystems, Foster City, CA, USA).

Equipment:
10/100/1000 µl single-channel pipetes (Eppendorf, Hamburg, Germany);
Heraeus Multifuge 3S-R (Kendro, Hanau, Germany);
Heraeus Biofuge fresco (Kendro, Hanau, Germany);

Micro Centrifuge (Roth, Karlsruhe, Germany);
Vortex-GENIE 2 G-560E (Scientific Industries, Bohemia, NY, USA);
ABI Prism™ 7900HT Sequence Detection System (Applied Biosystems Inc., Foster City, CA, USA);
GeneAmp® PCR System 9700 (Applied Biosystems Inc., Foster City, CA, USA);
Tecan Genesis RSP 150 (Tecan, Deutschland GmbH, Crailsheim, Germany);
Tecan Genesis Workstation 150 (Tecan, Deutschland GmbH, Crailsheim, Germany).

In brief the genotyping procedure is as follows:

1. 5 µl of the reaction mix (Table 2.3.3.1) was added to the 384 well plates with the dried gDNA by a TECAN multipipetting robot. In performing SNP genotyping assays, AmpliTaq Gold DNA polymerase from the TaqMan Universal PCR Master Mix amplifies target DNA, using sequence-specific primers and TaqMan minor groove binder probes from the SNP Genotyping Assay Mix.

Table 2.3.3.1. *Sample loading master mix*

Reagent	**Volume (µl) for 1 reaction**
TaqMan® PCR master mix	2.50
Ready-to-use SNP genotyping assay mix	0.25
Nuclease free H_2O	2.25
Total	**5**

2. The plates were sealed, and briefly spun to remove air bubbles trapped at the bottom of the wells.
3. As the data acquired during PCR amplification is not necessary for analysis, the GeneAmp® PCR System 9700 thermalcycler has been used for PCR amplification. Two-step PCR protocol was used (Table 2.3.3.2).

Table 2.3.3.2. *TaqMan® PCR protocol*

Step	**Temperature**	**Time**	**Cycle(s)**	**Function**
1	95ºC	10 min	1	activation of Ampli Taq Gold®
2	95ºC	15 s	45	denaturation
3	60ºC	1 min		annealing, elongation, nucleolytic cleavage of hybridized probes
4	4ºC	∞	1	storage

4. After PCR amplification, an endpoint plate read was carried out with the ABI Prism® 7900 Sequence Detection System. The SDS software calculated the fluorescence measurements made during the plate read and ploted Rn values based on the signals from each well. Allele calling for each plate was done manually to ensure data quality (Fig. 2.3.3.2). Using the software, one can determine which alleles are present in each sample.

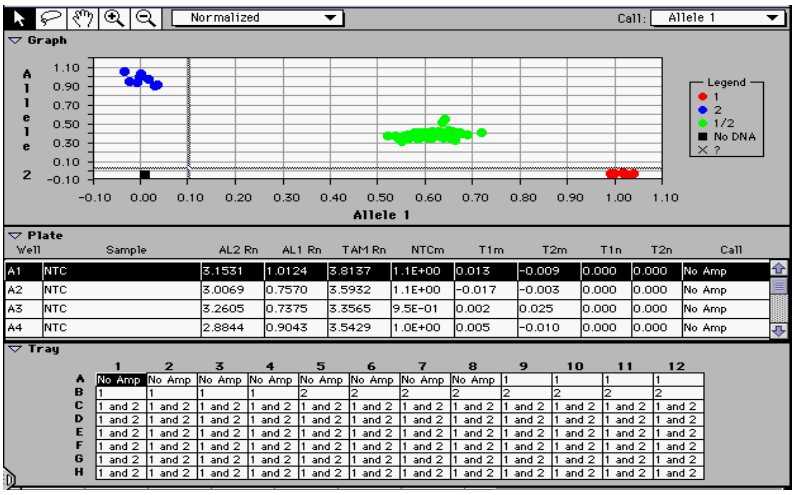

Fig. 2.3.3.2. *TaqMan® cluster plots*

Homozygotes 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.

2.4. Statistical analysis

2.4.1. *Pre-hoc* analysis

The statistical power is one of the major limiting factors for the detecting the true associations in the genetic studies. The statistical power within a given sample size depends on the effect size (OR) of variant, sample size, and allele frequency of the risk variant in controls. The lower the minor allele frequency (MAF) of the variant, the lower the power (or probability) to detect a variant with a certain OR (effect strength). Statistical power rises with increasing sample size.

Statistical power of this study to detect a given allelic disease association (calculations performed for carriership of the rarer SNP allele) in screening panel (578 cases, 1154 controls) is illustrated in Fig. 2.4.1.1. Screening panel had 80% power to detect a variant with an OR of 1.4 or higher at the 5% significance level, assuming a frequency of the disease-associated allele of at least 20% in controls. The detectable OR was 1.8 for CD and 1.44 for UC. Calculations were performed for different allele frequencies using PS Power and Sample Size v2.130 [198].

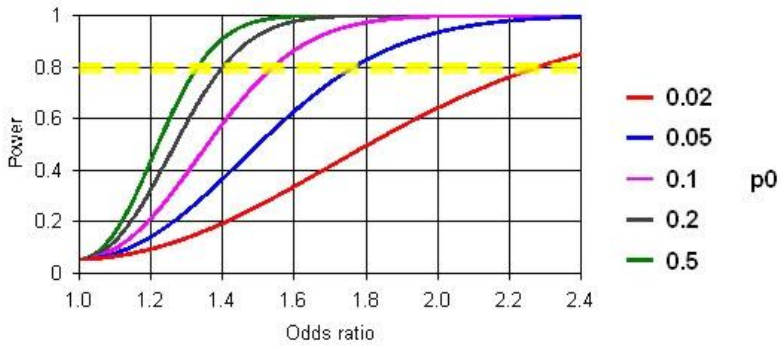

Fig. 2.4.1.1. *Power calculations*
p0 denotes different allele frequencies. Yellow dotted line shows the threshold of 80% power.

2.4.2. Quality control measures

Genotyping errors (*i.e.*, the observed genotype of an individual does not correspond to the true genotype [199]) markedly decrease the power for detecting associations [200–203]. Several levels of quality checks were applied to the data as suggested by Pompanon *et al.* (2005) [204] to ensure reliability of the results:

1. ***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 of the processes. This included the use of barcodes and scanners for plates and samples.

2. ***Low quantity or quality of DNA.*** Only a few or low-quality target DNA molecules favour allelic dropouts and false alleles [205]. Furthermore, the risk of contamination is increased. Therefore, each DNA sample was quality checked on a gel (chapter 2.2.4), normalized to a specific concentration using WGA (chapter 2.2.3).
3. ***Call rate, missingness, minor allele frequency and Hardy-Weinberg equilibrium*** [206, 207]. An indirect measure to control for genotyping quality is to control for deviation from Hardy-Weinberg equilibrium (HWE). The Hardy Weinberg distribution of genotype frequencies (F) for alleles A and B (frequencies f_A and f_B) in a population equals $F_{AA} = f_A^2$, $F_{AB} = f_A \times f_B$ and $F_{BB} = f_B^2$ [208]. Significant deviations from this measure either hint to genotyping errors (unacceptable high type I error (false positives) rates [116, 209, 210]) or to selective evolutionary pressure on this allele [207]. It is recommended to exclude variants that deviate from HWE in healthy controls. All markers were tested for HWE in the control group using a χ^2 test before inclusion in the association statistics ($P>0.01$ threshold). Only biallelic markers with minor allele frequency of >0.01 passing a call rate (CR) of >95% in cases and controls were used in the further analyses. Poor quality samples (genotype CR<95%) were also excluded from downstream analysis.
4. ***Positive and negative controls.*** Another quality control for the assays and the genotyping process was the inclusion of positive and negative controls as described in 2.2.2. In addition, each genotyping method had its own internal controls. For a description, see the corresponding chapters: SNPlex™ (2.3.2) and TaqMan® (2.3.3).

2.4.3. Association analysis

Association studies can be family- or population-based resulting in two different analysis methods, the transmission disequilibrium test for family-based studies [211] and the case-control analysis for population-based studies. In frames of this thesis case-control association analysis has been performed using the diagnostic disease categories CD, UC and controls.

Quality assessments (chapter 2.4.2) and further statistical analysis were performed using the PLINK v1.07 [212], which is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analysis in a computationally efficient manner.

The genotype and allele-based case-control tests were used for association analysis. Observed alleles and genotypes frequencies were compared to the expected frequencies, under the null hypothesis (H_0) that no differences exist between cases and controls. The statistical significance assessment of associations between cases and controls were calculated: (1) genotypes – using χ^2 or Fisher's exact test for 2 × 3 contingency tables; (2) alleles – using χ^2 or Fisher's exact test for 2 × 2 contingency tables. The significance level of the tests for considering P-values as significant was set to < 0.05. To take the different geographic origin of the study panels (Lithuanian and Latvian UC patients and controls) into account, the Cochran-Mantel-Haenzsel test (P_{CMH}) and a Breslow-Day test for OR at disease-associated regions heterogeneity (P_{BD}) were used in the combined two study groups analysis [213].

False-positive associations in genetic studies, occurring due to multiple hypotheses testing, are equally undesirable as false negative results. The Bonferroni correction is the most widely accepted way to correct for multiple testing [214]. Basically, a Bonferroni correction consists of multiplying the obtained P-values by the number of independent tests performed; however, the true mathematical background is much more complicated [214]. The number of independent SNPs was determined using de Bakker's tagging algorithm (pair-wise r^2 between tagged SNPs > 0.8, P_{HWE} > 0.01, SNPs less than 250 kb apart) [215] as implemented in the software program Haploview 4.1 [216] (78 SNPs were captured 100% using 69 tag SNPs; Appendix Table 4). A signicant threshold for corrected P values was set at $P_{CORR} < 0.05$.

The OR is a measure of the effect size, thus the strength of the effect. ORs were calculated and the 95% CIs were approximated using Fisher's exact method. The ratio of the odds of an event in the experimental (cases) group to the odds of an event in the control group is defined as the OR (Table 2.4.3.1).

Table 2.4.3.1. *Relative risk and exposure* [217]

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

Genotype–phenotype associations were calculated with χ^2-tests combining SNPs with phenotype subgroups of interest in 2 × 2 tables. A Bonferroni

(multiple testing) correction was applied for the number of complementary subgroups of patients.

2.4.4. SNP-SNP epistasis analysis

Epistasis is a phenomenon whereby the effects of a given gene on a biological trait are masked or enhanced by one or more other genes [129, 130]. In addition to increasing the power to detect associations, the interaction analysis will alow to elucidate the biological and biochemical pathways that underpin complex diseases [130].

SNP-SNP epistasis for case/control population-based sample was calculated using the PLINK implemented in the --fast-epistasis command. PLINK provides a logistic regression test for interaction that assumes an allelic model for both the main effects and the interactions. This test is based on a Z-score for the difference in SNP-SNP association (OR) between cases and controls. To follow the procedure for constructing an allelic test of a single locus, three genotype categories are twice collapsed into two allele categories. Specifically, the 4N independent alleles observed at two loci are counted in a sample of N individuals into a 2 × 2 table, following the logic below, so the allele (not the individual or haplotype) is the unit of analysis.

	BB	Bb	bb
AA	a	b	c
Aa	d	e	f
aa	g	h	i

First alleles at one locus are counted, *e.g.*, B, conditional on the genotype at A, which can be represented as a 3 × 2 table, which represents 2N alleles, not N individuals.

	B	b
AA	2a+b	2c+b
Aa	2d+e	2f+e
aa	2g+h	2i+h

Again this 3 × 2 table is collapsed into a 2 × 2 table, as follows:

		B	b
A		4a+2b+2d+e	4c+2b+2f+e
a		4g+2h+2d+e	4i+2h+2f+e

Based on this 2 × 2 table, the OR between loci A and B and its standard error are calculated in the standard manner. When cases and controls are present, the above procedure is performed separately in cases and controls, and the test for epistasis is the difference of the two ORs, where R and S are the ORs in cases and controls respectively, estimated as ab/cd with variance 1/a+1/b+1/c+1/d and a, b, c, d are the four cells of the 2 × 2 table above:

$$Z = (\log(R) - \log(S))/\mathrm{sqrt}(SE(R) + SE(S))$$

This test follows a standard normal distribution under the multiplicative model of no interaction. All pairwise combinations of SNPs can be tested. OR for interaction, χ^2 statistics and asymptotic P-value ($P < 0.01$) were provided in the output file. Nominal P-values were adjusted for multiple testing using Bonferroni correction by multiplying P-values by the number of effective tests performed.

2.4.5. *In sicilo* prediction of gene interactive network

In silico prediction of the possible genes association network and genes functions was performed using the GeneMANIA v2.7.12 (Gene Function Prediction using a Multiple Association Network Integration Algorithm; www.genemania.org) [218] – an integrated interaction network program that predicts gene functions and possible interaction networks using many large publicly available datasets including protein-protein and genetic interaction networks, gene expression data, protein domain information, pathways and biochemical reaction databases, or subcellular localization.

GeneMANIA makes gene function predictions based on query-dependent weighting (query list consists of >6 genes), equal weighting, and gene ontology (GO) annotations (query list consists of <6 genes) patterns. The GeneMANIA algorithm consists of two parts: 1. a linear regression-based algorithm that

calculates a single composite functional association network from multiple data sources; 2. a label propagation algorithm for predicting gene function given the composite functional association network. Each functional association network derived from the data sources is assigned a positive weight, reflecting the data sources' usefulness in predicting the function. The weighted average of the association networks is constructed into a function-specific association network. GeneMANIA predicts gene function from the composite network using a variation of the Gaussian field label propagation algorithm that is appropriate for gene function predictions in which there are typically relatively few positive examples. Label propagation algorithms assign a score (Q-value) to each node in the network. This score reflects the computed strength of association that the node has to the seed list defining the given function. This value can be thresholded to enable predictions of a given gene function.

2.4.6. Genetic risk profile analysis

Genetic risk profiles were constructed for CD cases and healthy controls, and UC cases and healthy controls using: (1) the SNPs that were associated with CD and UC in the current study after correction for multiple testing and (2) previously CD or UC associated SNPs in the current study exhibiting nominal significance, which were also validated using the area under the receiver operator characteristic (ROC) curve (Appendix Tables 10 and 11) – measure for determining the efficacy of clinical diagnostic and prognostic tests in correctly classifying diseased and non-diseased individuals. ROC curve has been used in the context of genomic profiling [219–222]. A risk assessment model was constructed by including only SNPs that were independently associated with disease. Standardized coefficients r^2 and D' were computed with the software program Haploview 4.1 [216] for pair-wise LD estimation between markers.

Two different scoring models were used to generate the scores in the genetic panel:

1. In the first model, the cumulative number of risk alleles per individual was calculated and summed up. The categories of similarly sized groups of individuals with a specific number of risk alleles were made and then ORs for CD and UC were calculated in binary logistic regression analysis for each category with a reference group: CD reference group consisted of all individuals with zero or one risk allele; UC reference group – individuals

with ≤ 8 SNPs. These reference groups were chosen because the groups containing zero risk alleles were too small to be used as a reference. Student's t-test was used to see whether patients had more risk alleles than controls. The significance level of the tests for considering two-sided P-values as significant was set to <0.05. The corresponding 2 × 2 tables were used to determine sensitivities, specificities, positive (PPV) and negative predictive values (NPV), and likelihood ratios (LR).

2. In the second model, a weighted score for the number of risk alleles calculated per individual, as different genetic variants have different effect sizes on disease susceptibility. The β-coeficients per genotype were calculated from separate logistic regression analyses for each CD or UC associated SNP, assuming a multiplicative disease model (Appendix Tables 12 and 13). The β-coefficient of regression is a relative measure of how much the individual result contributes to the model in developing specific phenotype. Thus, the β-coeficient was added once for a heterozygote and twice for a homozygote. In this way, the score was weighted for the strength of association for each individual SNP. For example, the logistic regression coefficient for the *NOD2* SNP rs2066847 was 0.91 and heterozygous individuals were therefore awarded a score of 0.91 and homozygotes a score of 1.82. The β-coeficients were summed up per individual to obtain a weighted genetic load. In this way, the sum of risk alleles was adjusted for the strength of association for each genetic variant. All patients and controls were then categorized on the basis of weighted score, and ORs for each category were determined by logistic regression analysis, with the category with the lowest weighted score being used as a reference group (UC ≤8; CD ≤3). A Student's t-test was used to analyze whether CD or UC patients had a higher score than controls. The significance level of the tests for considering two-sided P-values as significant was set to <0.05. The corresponding 2 × 2 tables were used to determine sensitivities, specificities, PPV, NPV, and LR.

Data was evaluated using the web interface SISA [223], Microsoft Excel 2010, and SPSS v.13.0.

3. RESULTS

3.1. Genotyping success rate and heterogeneity

In total, 99.6% of all cases and controls were successfully genotyped. One SNP (rs2289310) failed genotyping and six SNPs (rs2925757, rs10077785, rs2076756, rs7868736, rs10974944, rs2066847) showed heterogeneity between Lithuanian and Latvian UC study samples ($P_{BD} < 0.05$) and therefore were excluded from the further data analysis. None of the SNPs showed deviation from HWE ($P_{HWE} > 0.05$) (Appendix Tables 5 and 6). In the Lithuanian study group of 128 CD patients and 1097 control individuals the remaining 76 SNPs were analyzed (UC specific SNPs rs7809799 and rs5771069 were not analysed in the CD sample set); whereas in the two study populations comprising UC study sample (444 UC patients and 1154 control individuals) 72 SNPs were analyzed.

3.2. Single marker case-control association analysis

3.2.1. Association analysis in Crohn's disease

Of the 76 SNPs included in the case-control single marker analysis, 14 showed significant association with CD (see Table 3.2.1.1; Appendix Table 5), and these comprise seven independent loci.

The candidate region showing the strongest association was in the *NOD2* gene. Both lead SNPs genotyped within this locus showed robust evidence for association (rs2066847, $P_{CORR} = 1.62\times10^{-13}$; rs2076756, $P_{CORR} = 5.56\times10^{-7}$). The MAFs of rs2066847 and rs2076756 were increased in cases (15.6% and 35%, respectively) *vs* controls (3.9% and 19.3%, respectively) resulting in increased risk of CD (rs2066847: $OR_{allele} = 4.52$ (95% CI: 3.02–6.78), $OR_{carriership} = 4.37$ (95% CI: 2.77–6.91); rs2076756: $OR_{allele} = 2.24$ (95% CI: 1.69–2.97), $OR_{carriership} = 2.60$ (95% CI: 1.79–3.78)). Homozygous variant allele carriers of rs2066847 were at 21.74–fold (95% CI: 5.35–88.36; $P_{HOM} = 9.91\times10^{-10}$) and rs2076756 – 4.18–fold (95% CI: 2.23–7.82; $P_{HOM} = 1.67\times10^{-6}$) increased risk; whereas heterozygous variant allele carriers of rs2066847 were at 3.72–fold (95% CI: 2.28–6.04; $P_{HET} = 2.12\times10^{-8}$) and rs2076756 – 2.34–fold (95% CI: 1.58–3.49; $P_{HET} = 2\times10^{-5}$) increased risk of CD. Thus, the risk of disease in rs2066847 CC and rs2076756 GG homozygotes was substantially

Table 3.2.1.1. *Summary of results for SNPs significantly associated with Crohn's disease*

Gene marker	Gene	A1	MAFco	MAFca	P_{CCA}	OR (95% CI)	Power
rs2066847	*NOD2*	C	0.039	0.156	**2.46×10^{-15}**	4.52 (3.02–6.78)	84.52%
rs2076756	*NOD2*	G	0.193	0.349	**8.43×10^{-9}**	2.24 (1.69–2.97)	96.76%
rs10521209	*NOD2*	G	0.443	0.332	7.71×10^{-4}	0.62 (0.47–0.82)	44.46%
rs2066845	*NOD2*	C	0.008	0.028	3.83×10^{-3}	3.41 (1.41–8.25)	69.50%
rs3024505	*IL10*	A	0.129	0.188	0.010	1.56 (1.11–2.19)	43.84%
rs9268858	6p21.32 (*HLA)*	C	0.238	0.311	0.011	1.44 (1.09–1.92)	[a]49.78%
rs2395185	6p21.32 (*HLA*)	T	0.241	0.31	0.016	1.42 (1.07–1.88)	[a]34.35%
rs13361189	*IRGM*	C	0.042	0.075	0.017	1.85 (1.11–3.09)	21.12%
rs9268877	6p21.32 (*HLA*)	G	0.491	0.569	0.021	1.37 (1.05–1.78)	[a]61.37%
rs7712957	*S100Z*	C	0.058	0.094	0.021	1.70 (1.08–2.69)	37.10%
rs1736135	21q21.1	C	0.426	0.356	0.033	0.74 (0.57–0.98)	13.95%
rs2301436	*CCR6*	A	0.484	0.413	0.033	0.75 (0.58–0.98)	[a]17.84%
rs11747270	*IRGM*	G	0.043	0.073	0.034	1.75 (1.04–2.95)	30.25%

A1 – minor allele; MAFco – minor allele frequency in the controls sample set (n=1097); MAFca – minor allele frequency in the CD group (n=128); P_{CCA} – P-values from an allele-based case-control comparison with 1 degree of freedom; P-values that withstood correction for multiple testing (corrected for 67 independent tests; $P_{CORR} < 0.05$) are highlighted in **bold**; OR (95% CI) – odds ratio for carriership of the rarer allele (95% confidence interval of OR); Power – the power of this study to replicate the association at 0.05 significance level (MAF and OR presented in the original studies were used for calculations); a – SNPs that displayed the opposite risk-increasing alleles as reported in the previous studies.

higher than the risk in carriers of the rs2066847 C allele and rs2076756 G allele, suggesting a dosage effect.

The previously reported CD associated SNPs in *NOD2* (rs10521209, rs2066845), *IRGM* (rs13361189, rs11747270), and *CCR6* (rs2301436); and UC associated SNPs in the *IL10* (rs3024505), S100 calcium binding protein Z gene *(S100Z;* rs7712957), and *HLA-DRA* (rs9268858, rs2395185, rs9268877) loci showed only nominal evidence for association in our study sample and failed to withstand Bonferroni correction for multiple testing (Table 3.2.1.1, Appendix Table 5). Our power to detect association (assuming the same effect size as documented in original studies) with each of the SNP is given in Table 3.2.1.1. Estimated power varied widely between SNPs. However, the highest power had the two lead SNPs in the *NOD2*. Moreover, for several markers in *HLA* locus and *CCR6* we observed the opposite risk-increasing alleles as reported in the previous studies, therefore for these SNPs analyses in larger cohorts will be required to elucidate their role in CD. Furthermore, we could not confirm the previously described CD associations with *IL23R, ATG16L1, IL12B, NKX2-3, STAT3, NELL1, 5p13, PTPN22, etc.* (Appendix Table 5).

Moreover, it has been previously reported that *NOD2* mutations rs2066844, rs2066845, and rs2066847 have a dose dependent effect as mutated homozygotes and compound heterozygotes are found more frequently in CD than expected [224]. As expected, none of the studied individuals were carriers of all three *NOD2* risk alleles. However, four CD patients were determined as compound heterozygotes. The combined allele carriership in the group of patients with CD was much higher than in controls (33.59% *vs* 11.12%, respectively) and resulted in significant association ($P = 2.70\times10^{-12}$; OR = 4.00, 95% CI: 2.65–6.03).

3.2.2. Association analysis in ulcerative colitis

The results of the single marker case-control association analysis of UC are presented in Table 3.2.2.1 and Appendix Table 6. Twenty SNPs from 15 independent loci showed evidence for association in the allele and genotype-based comparisons. After Bonferroni correction five lead SNPs, tagging five loci, remained significantly associated. The genetic loci showing robust evidence for association included 21q21.1 (rs1736135, $P_{CORR} = 4.89\times10^{-4}$), 6q21 (rs7746082, $P_{CORR} = 3.91\times10^{-3}$), *JAK2* (rs10758669, $P_{CORR} = 4.93\times10^{-3}$), *ORMDL3* (rs2872507, $P_{CORR} = 7.59\times10^{-3}$), and ring finger protein 186 gene (*RNF186;* rs3806308, $P_{CORR} = 0.015$). Importantly, for each marker showing

association, we observed the same risk-increasing allele as reported previously in IBD studies.

The MAF of the lead SNP rs1736135 was decreased in cases (33.8%) *vs* controls (42.6%) resulting in a protective effect against UC (OR_{allele} = 0.69 (95% CI: 0.59–0.81), $OR_{carriership}$ = 0.66 (95% CI: 0.53–0.83)). The disease risk in CC homozygotes (OR = 0.44 (95% CI: 0.31–0.64); $P_{HOM} = 9\times10^{-6}$) was twice lower than the risk conferred by CT heterozygosity (OR = 0.75 (95% CI: 0.59–0.95); P_{HET} = 0.015), suggesting a dose effect.

Table 3.2.2.1. *Summary of results for SNPs significantly associated with ulcerative colitis*

Gene marker	Gene	A1	MAFco	MAFca	P_{CMH}	OR (95% CI)	Power
rs1736135	21q21.1	C	0.426	0.338	$\mathbf{8.01\times10^{-6}}$	0.69 (0.59–0.81)	29.68%
rs7746082	6q21	C	0.266	0.339	$\mathbf{6.41\times10^{-5}}$	1.41 (1.19–1.67)	26.02%
rs10758669	*JAK2*	C	0.355	0.434	$\mathbf{8.08\times10^{-5}}$	1.38 (1.17–1.62)	16.71%
rs2872507	*ORMDL3*	A	0.411	0.485	$\mathbf{1.24\times10^{-4}}$	1.36 (1.16–1.59)	17.42%
rs3806308	*RNF186*	T	0.446	0.376	$\mathbf{2.40\times10^{-4}}$	0.74 (0.63–0.87)	84.93%
rs3024505	*IL10*	A	0.128	0.173	1.04×10^{-3}	1.43 (1.16–1.77)	80.83%
rs11209026	*IL23R*	A	0.066	0.037	2.16×10^{-3}	0.55 (0.38–0.81)	63.14%
rs3197999	*MST1*	T	0.228	0.279	3.21×10^{-3}	1.315 (1.10–1.58)	50.56%
rs9268877	6p21.32 (*HLA*)	T	0.508	0.564	4.34×10^{-3}	1.25 (1.07–1.47)	96.13%
rs6426833	1p36.13	A	0.486	0.541	6.01×10^{-3}	1.25 (1.07–1.46)	92.48%
rs2395185	6p21.32 (*HLA*)	T	0.238	0.191	6.44×10^{-3}	0.76 (0.63–0.93)	76.01%
rs11190140	*NKX2-3*	T	0.464	0.519	7.27×10^{-3}	1.25 (1.06–1.48)	82.85%
rs4263839	*TNFSF15*	A	0.327	0.282	0.01	0.80 (0.67–0.95)	36.27%
rs9268858	6p21.32 (*HLA*)	C	0.236	0.193	0.011	0.78 (0.64–0.94)	91.93%

Table 3.2.2.1 continued

Gene marker	Gene	A1	MAFco	MAFca	P_{CMH}	OR (95% CI)	Power
rs11465804	*IL23R*	G	0.05	0.03	0.012	0.58 (0.38–0.89)	51.74%
rs10883365	*NKX2-3*	G	0.472	0.517	0.021	1.20 (1.03–1.41)	99.08%
rs762421	*ICOSLG*	G	0.36	0.406	0.023	1.21 (1.03–1.42)	19.15%
rs9268480	*BTNL2*	T	0.211	0.173	0.025	0.79 (0.65–0.97)	47.05%
rs9858542	*BSN*	A	0.229	0.266	0.026	1.23 (1.03–1.48)	99.94%
rs1992660	5p13.1	G	0.387	0.35	0.046	0.85 (0.72–0.99)	96.79%

A1 – minor allele; MAFco – minor allele frequency in the controls sample set (n=1154); MAFca – minor allele frequency in the UC sample set (n=444); P_{CMH} – p-values from Cochran-Mantel-Haenzsel test; OR (95% CI) – odds ratio for carriership of the rarer allele (95% confidence interval of OR); P-values that withstood correction for multiple testing (corrected for 63 independent tests; $P_{CORR} < 0.05$) are highlighted in **bold**; Power – the power of this study to replicate the association at 0.05 significance level (MAF and OR presented in the original studies were used for calculations).

The *RNF186* associated SNP rs3806308 also exhibited the protective effect against UC (MAFca = 37.6% *vs* MAFco = 44.6%; OR_{allele} = 0.74 (95% CI: 0.63–0.87), $OR_{carriership}$ = 0.68 (95% CI: 0.54–0.85)). The risk conferred by homozygous TT genotype (OR = 0.57 (95% CI: 0.41–0.79); $P_{HOM} = 8.5\times10^{-4}$) was 21% lower than the risk conferred by TC heterozygosity (OR = 0.72 (95% CI: 0.57–0.92); $P_{HET} = 8.47\times10^{-3}$).

The MAFs of rs7746082, rs10758669 and rs2872507 were increased in cases (33.9%, 43.4%, and 48.5%, respectively) *vs* controls (26.6%, 35.5%, and 41.1%, respectively) resulting in increased risk of UC (rs7746082: OR_{allele} = 1.41 (95% CI: 1.91–1.67), $OR_{carriership}$ = 1.54 (95% CI:1.24–1.93); rs10758669: OR_{allele} = 1.38 (95% CI: 1.17–1.62), $OR_{carriership}$ = 1.55 (95% CI: 1.23–1.96); rs2872507: OR_{allele} = 1.36 (95% CI: 1.16–1.59), $OR_{carriership}$ = 1.53 (95% CI: 1.19–1.96)). Homozygous variants alleles carriers of the three SNPs had approximately 25% higher UC risk (rs7746082: OR = 1.88 (95% CI: 1.27–2.79), $P_{HOM} = 1.37\times10^{-3}$; rs10758669: OR = 1.98 (95% CI: 1.41–2.80), $P_{HOM} = 7\times10^{-5}$; rs2872507: OR = 1.85 (95% CI: 1.33–2.55); $P_{HOM} = 2\times10^{-4}$) compared to heterozygotes (rs7746082: OR = 1.48 (95% CI: 1.17–1.87), $P_{HET} = 1.01\times10^{-3}$;

rs10758669: OR = 1.44 (95% CI: 1.13–1.85), $P_{HET} = 3.40\times10^{-3}$; rs2872507: OR = 1.43 (95% CI: 1.10–1.85); $P_{HET} = 7.72\times10^{-3}$).

The previously reported UC associated SNPs in the *IL10* (rs3024505), *HLA* (rs9268858, rs2395185, rs9268877, rs9268480), and 1p36.13 (rs6426833) loci; UC and CD associated variants in *IL23R* (rs11209026, rs11465804), *MST1* (rs3197999), *NKX2-3* (rs11190140, rs10883365), and *BSN* (rs9858542) genes; and previously reported CD risk loci: *TNFSF15* (rs4263839), *ICOSLG* (rs762421), 5p13.1 (rs1992660) were only moderately associated in our cohort and all 15 failed to withstand correction for multiple testing (Appendix Table 6). Our power to detect association (assuming the same effect size as documented in original studies) with each of the SNP is given in Table 3.2.2.1. Estimated power varied widely between SNPs. The highest power exhibited previously UC associated loci: *HLA*, *IL10*; and both IBD predisposing variants in *NKX2-3, BSN*. For those markers for which we had little power, analyses in larger cohorts will be required to elucidate their role in UC.

Moreover, the previously reported UC associations with *ARP2C, S100Z, IL12B, STAT3, NELL1, IL17REL, etc.* (Appendix Table 6) were not confirmed in our study sample.

3.3. Genetic association with disease phenotype

The detailed genotype-phenotype analyses of all SNPs that showed at least marginal association with CD and UC are presented in Tables 3.3.1.1 and 3.3.2.1. We tested genetic association with phenotypic characteristics such as: disease localization, disease behavior, extraintestinal manifestations, family history of IBD, surgery treatment, and treatment using biological therapy. As isolated upper gastrointestinal (GI) involvement had only one patient, patients having involvement in intestines and upper GI (L1+L4=3; L2+L4=2; L3+L4=1) were combined for futher genetic analysis. Moreover, none of the patients had isolated perianal behavior of disease; therefore patients having combined behavior with perianal (*i.e.*, B1p, B2p, or B3p) were compared with patients having isolated disease behavior form (*i.e.*, B1, B2, or B3). In addition, the extraintestinal manifestations subgroup of IBD patients with joint involvement included patients with diagnosis of peripheral arthritis, spondyloarthropathies (sacroiliitis), and ankylosing spondylitis.

3.3.1. Association with Crohn's disease phenotype

The CD disease location subgroup analysis showed significant associations with the two disease associated loci (Table 3.3.1.1). The robust evidence for association showed only the upper GI involvement risk increasing association with *NOD2* genetic locus (rs2066847: OR = 6.38 (95% CI: 2.10–19.38); $P_{CORR} = 1.04\times10^{-3}$). Homozygous CC genotype carriers were at 22.50–fold ($P_{HOM} = 1.90\times10^{-4}$) increased risk for phenotype occurance; whereas heterozygous variant allele carriers were at 5.63–fold ($P_{HET} = 0.042$) increased risk of CD upper GI involvement, suggesting a dosage effect. The *HLA* locus associated SNP was only moderately associated with CD intestinal involvement phenotype (*i.e.*, ileal involvement (combining terminal ileum and ileocolon) *vs* colon) in our study sample and failed to withstand correction for multiple testing (rs9268877: $P_{CORR} = 0.092$).

Table 3.3.1.1. *Significant associations between SNPs and Crohn's disease phenotype*

Phenotype	SNP	Allele	GT_{PH+} (11/12/22)	GT_{PH-} (11/12/22)	P_{CCA}	OR (95%CI)
Localization						
Ileal/non-ileal	rs9268877	G	34/43/13	6/18/9	0.023	1.93 (1.09–3.41)
Upper GI/ non-upper GI	rs2066847	C	2/3/2	4/24/90	**2.60×10^{-4}**	6.38 (2.10–19.38)
Disease behavior						
Non-stricturing/ Stricturing	rs13361189	C	0/11/90	1/3/8	**5.46×10^{-3}**	0.22 (0.07–0.70)
	rs11747270	G	0/11/87	1/2/8	0.027	0.27 (0.08–0.93)
Non-stricturing/ Stricturing	rs9268858	C	13/40/48	0/3/9	0.042	3.40 (0.97–11.80)
	rs2395185	T	13/40/48	0/3/9	0.042	3.40 (0.97–11.80)
	rs2066847	C	4/19/77	1/5/5	0.023	0.33 (0.12–0.89)
Perianal/ non-perianal	rs9268858	C	0/4/17	14/47/45	**9.44×10^{-4}**	0.19 (0.07–0.56)
	rs9268877	G	7/10/1	16/51/39	**2.06×10^{-3}**	0.32 (0.15–0.68)

Table 3.3.1.1 continued

Phenotype	SNP	Allele	GT_{PH+} (11/12/22)	GT_{PH-} (11/12/22)	P_{CCA}	OR (95%CI)
Perianal/ non-perianal	rs2395185	T	0/4/17	14/46/45	**9.99×10^{-4}**	0.19 (0.067–0.56)
Extraintestinal manifestations						
Joints/ no extraintestinal	rs2066847	C	0/4/28	6/20/57	0.015	0.28 (0.09–0.83)
All extraintestinal/ no extraintestinal	rs2066847	C	0/7/35	6/20/57	0.015	0.38 (0.16–0.90)
	rs2076756	G	1/20/21	15/36/33	0.04	0.55 (0.31–0.98)
Biological therapy/ no biological therapy	rs13361189	C	1/7/18	1/8/92	**2.52×10^{-3}**	4.02 (1.54–10.49)
	rs11747270	G	1/6/17	1/8/90	**5.56×10^{-3}**	3.76 (1.40–10.12)
Combined analysis*	rs13361189	C	2/7/32	0/9/81	**0.013**	3.18 (1.23–8.23)
	rs11747270	G	2/6/31	0/9/79	0.024	2.94 (1.11–7.77)

Allele – allele associated with CD in our initial case-control study (Table 3.2.1.1); GT_{PH+} – genotype count of cases positive for the phenotype under study; GT_{PH-} – genotype count of cases negative for the phenotype under study; 11 = homozygous for minor allele; 12 = heterozygous; 22 = homozygous for common allele; P_{CCA} – values from an allele-based case-control comparison with 1 degree of freedom; P-values that withstood correction for multiple testing ($P_{CORR} < 0.05$) are highlighted in **bold**; OR (95% CI) – odds ratio for carriership of the rarer allele (95% confidence interval of OR); Combined analysis* – surgery + biological therapy/ no surgery + no biological therapy.

The *HLA* locus in the perianal *vs* non-perianal analysis and *IRGM* gene in the stricturing *vs* non-stricturing comparison were substantially associated with CD behavior (*HLA:* rs9268858, $P_{CORR} = 1.34\times10^{-3}$; rs2395185, $P_{CORR} = 1.48\times10^{-3}$; rs9268877, $P_{CORR} = 2.03\times10^{-3}$; *IRGM:* rs13361189, $P_{CORR} = 0.022$). Homozygous carriers of the rs13361189 C allele had a 31.94–fold increased risk of stricturing form of CD ($P_{HOM} = 1.48\times10^{-3}$); whereas allele heterozygosity increased the risk 9.86–fold, but the association was not significant ($P_{HET} = 0.09$). Thus, the risk of stricturing disease behavior in rs13361189 CC homozygotes was substantially higher than the risk in carriers of the rs13361189 C allele, suggesting a dosage effect. Homozygous variant allele carriers of the three *HLA* locus SNPs had approximatelly 80% lower risk

of CD perianal behavior (rs9268858: OR = 0.09 (95% CI: 0.005–1.59), P_{HOM} = 0.026; rs2395185: OR = 0.09 (95% CI: 0.005–1.59), P_{HOM} = 0.026; rs9268877: OR = 0.06 (95%CI: 0.007–0.52), $P_{HOM} = 1.35\times10^{-3}$) compared to heterozygotes (rs9268858: OR = 0.23 (95% CI: 0.07–0.72), $P_{HET} = 7.76\times10^{-3}$; rs2395185: OR = 0.23 (95% CI: 0.07–0.74), $P_{HET} = 8.86\times10^{-3}$; rs9268877: OR = 0.45 (95% CI: 0.15–1.37), P_{HET} = 0.15). The association with the *NOD2* and *HLA* locus in the stricturing *vs* non-stricturing comparison was marginal and did not withstand correction for multiple testing ($P_{CORR} > 0.05$).

The carriers of *IRGM* disease associated alleles had an increased risk of the need for biological therapy (rs13361189: OR = 4.02 (95% CI: 1.54–10.49), $P_{CORR} = 5.04\times10^{-3}$; rs11747270: OR = 3.76 (95% CI: 1.40–10.12), P_{CORR} = 0.011). Homozygous SNPs allele carriers were at increased risk for acquiring phenotype, however the association was not significant (rs13361189: OR = 5.11, P_{HOM} = 0.21; rs11747270: OR = 5.29, P_{HOM} = 0.19); whereas the heterozygous rs13361189 CT and rs11747270 AG genotypes significantly increased the risk (rs13361189, $P_{HET} = 5.9\times10^{-3}$, OR = 4.47 (95% CI: 1.44–13.89); rs11747270 P_{HET} = 0.016, OR = 3.97 (95% CI: 1.22–12.90).

Moreover, the genetic analysis of the severe disease cases (*i.e.*, combining patients that required surgery treatment and/or biological therapy during the course of the disease) had revieled the significant association with the disease associated *IRGM* gene SNP carriers (rs13361189: OR = 3.18 (95% CI: 1.23–8.23), $P_{CORR} = 3.91\times10^{-2}$). Homozygous carriers of the rs13361189 C allele had an increased risk of severe form of CD (OR = 12.54, P_{HOM} = 0.028); whereas allele heterozygosity increased the risk, but not significantly (OR = 1.97, P_{HET} = 0.21). Thus, the risk of disease in rs13361189 CC homozygotes was substantially higher than the risk in carriers of the rs13361189 C allele, suggesting a dosage effect.

Finally, the genetic analysis of the extraintestinal manifestation subgroups revealed nominal association with the *NOD2* SNPs, however after subjection for Bonferroni correction none of the associations remained significant (rs2066847: P_{CORR} = 0.075; rs2076756: P_{CORR} = 0.205). Moreover, there was no association between the disease associated SNPs and family history of CD, colonic form of CD, and extraintestinal manifestations affecting eyes, skin, and hepatobiliary system.

3.3.2. Association with ulcerative colitis phenotype

The UC disease extension subgroup comparison revealed significant associations with the two disease associated loci (Table 3.3.2.1). The robust evidence for association showed extensive colitis risk increasing association with *HLA* genetic locus (rs9268480: OR = 1.79 (95% CI: 1.14–2.81); $P_{CORR} = 0.033$). The homozygous carriers of the common rs9268480 G allele were at approximately 20% increased risk for phenotype occurance ($OR_{HOM} = 2.16$) compared to carriers of G allele, however the association was not significant ($P_{HOM} = 0.33$); whereas the carriage of at least one risk allele (*i.e.,* GG and GT) increased the phenotype risk significantly (OR = 1.99 (95% CI: 1.20–3.30), $P = 7.11\times10^{-3}$) compared to carriers of TT genotype. The *IL10* locus associated SNP was only moderately associated with both left-sided colitis and extensive colitis in our study population and failed to withstand correction for multiple testing (rs3024505, $P_{CORR} = 0.063$ (proctitis *vs* left-sided colitis), $P_{CORR} = 0.11$ (proctitis *vs* extensive colitis). However, in the combined analysis of both more severe disease representing extensions (*i.e.,* left-sided colitis and extensive colitis) the association with *IL10* locus was robustly significant (OR = 1.83 (95% CI: 1.11–3.03); $P_{CORR} = 0.048$). The genotype analysis revealed that carriage of at least one risk allele (*i.e.,* AA and AG) increased the phenotype risk approximately 20% (OR = 2.18 (95% CI: 1.23–3.87), $P = 6.61\times10^{-3}$) compared to carriers of GG genotype.

The disease associated alleles carriers of rs2872507 in 1p36.13 locus had an increased risk of the joints involvement in the UC patients ($P_{CORR} = 0.047$). Homozygous carriers of the rs2872507 A allele had a 3.45–fold (95% CI: 1.29–9.24) increased risk of phenotype ($P_{HOM} = 9.96\times10^{-3}$), compared to the carriers of the A allele (OR = 1.84).

The 1p36.13 locus was robustly associated with an increased risk of the need for biological therapy (rs6426833: OR = 2.02 (95% CI: 1.05–3.87), $P_{CORR} = 2.9\times10^{-3}$). The risk of disease phenotype in AA homozygotes was substantially higher (OR = 4.36 (95% CI: 1.34–14.18), $P_{HOM} = 8.50\times10^{-3}$) than the risk in carriers of the A allele (OR = 2.02). The 6q21 locus SNP rs7746082 showed marginal association with the need for biological therapy in UC patients and did not withstand correction for multiple testing ($P_{CORR} = 0.064$).

Moreover, the genetic analysis of the severe UC cases (*i.e.,* combined analysis of patients that required colectomy or other UC-related surgery treatment and/or biological therapy during the course of the disease) revieled

Table 3.3.2.1. *Significant associations between SNPs and ulcerative colitis phenotype*

Phenotype	SNP	Allele	GT_{PH+} (11/12/22)	GT_{PH-} (11/12/22)	P_{CCA}	OR (95%CI)
Disease extension						
Proctitis/left-sided colitis+ extensive colitis	rs3024505	A	3/14/72	13/108/235	**0.016**	0.55 (0.31–0.90)
Left-sided colitis/ extensive colitis	rs9268480	G	144/80/7	89/25/2	**0.011**	0.56 (0.36–0.88)
Extraintestinal manifestations						
Joints/ no extraintestinal	rs2872507	A	16/20/6	75/190/97	$\mathbf{9.49\times10^{-3}}$	1.84 (1.15–2.92)
All extraintestinal/ no extraintestinal	rs2872507	A	22/39/13	75/190/97	0.043	1.44 (1.01–2.05)
Biological therapy/no-biological therapy	rs6426833	A	10/7/2	78/217/122	$\mathbf{1.45\times10^{-3}}$	3.04 (1.49–6.21)
	rs7746082	C	6/7/6	42/193/183	0.032	2.02 (1.05–3.89)
Combined analysis*	rs6426833	A	13/21/6	75/203/118	**0.015**	1.77 (1.11–2.83)
	rs7746082	C	11/15/14	37/185/175	**0.014**	1.78 (1.12–2.83)
	rs2872507	A	16/16/8	81/213/102	0.031	1.67 (1.04–2.67)

Allele – allele associated with CD in our initial case-control study (Table 3.2.2.1); GT_{PH+}– genotype count of cases positive for the phenotype under study; GT_{PH-} – genotype count of cases negative for the phenotype under study; 11 = homozygous for minor allele; 12 = heterozygous; 22 = homozygous for common allele; P_{CCA} – values from an allele-based case-control comparison with 1 degree of freedom; P-values that withstood correction for multiple testing ($P_{CORR} < 0.05$) are highlighted in **bold**; OR (95% CI) – odds ratio for carriership of the rarer allele (95% confidence interval of OR); Combined analysis* – colectomy + other surgery + biological therapy/ no surgery + no biological therapy.

the significant association with 1p36.13 locus (rs6426833: OR = 1.77; P_{CORR} = 0.046) and 6q21 locus (rs7746082: OR = 1.78, P_{CORR} = 0.042) disease associated alleles. Homozygous carriers of the rs6426833 A allele or rs7746082 C allele had an increased risk of severe form of UC (rs6426833: OR = 3.41, P_{HOM} = 0.013; rs7746082: OR = 3.72, $P_{HOM} = 1.79\times10^{-3}$); whereas allele heterozygosity increased the risk, however the association was not significant

(rs6426833: OR = 2.03, P_{HET} = 0.13; rs7746082: OR = 1.01, P_{HET} = 0.97). Thus, the risk of disease phenotype in rs6426833 AA and rs7746082 CC homozygotes was substantially higher than the risk in carriers of the rs6426833 A allele and and rs7746082 C allele, suggesting a dosage effect.

Finally, there was no association between the disease associated SNPs and family history of UC and extraintestinal manifestations: cutaineous, ocular, hepatobiliary.

3.4. SNP-SNP epistasis

The SNP-SNP interaction (epistasis) has been investigated among all candidate SNPs (that passed quality criteria (chapter 3.1)) using a logistic regression test. As a result, the statistically significant interactions were found between 31 pair of SNPs in the UC group (Appendix Table 7) and 17 pairs of SNPs in the CD group (Appendix Table 8). However, after subjection for Bonferroni correction only one SNP pair: rs2476601 and rs3764147, in the UC group remained significant ($P_{CORR} = 3.93\times10^{-3}$, OR = 2.44) assuming an additive genetic model. Interacting SNPs were in genes *PTPN22* (rs2476601) and *C13orf31* (rs3764147). The interaction pattern for the most significant SNP pair is reported in Table 3.4.1.

Table 3.4.1. *Genotype counts for the SNPs pair in ulcerative colitis (rs2476601, rs3764147) and odds ratio relative to the most common double homozygous genotype: (rs3764147, rs2476601) = (AA,GG)*

		rs3764147		
	rs2476601	AA	AG	GG
Controls	GG	407	357	63
	GA	161	92	16
	AA	17	8	0
Ulcerative colitis	GG	170	119	19
	GA	45	57	15
	AA	3	7	1
OR relative to AA/GG (95% CI)	GG	1	0.80(0.61–1.05)	0.72(0.42–1.24)
	GA	**0.67(0.46–0.97)**	**1.48 (1.02–2.16)**	**2.24(1.09–4.64)**
	AA	0.42 (0.12–1.46)	2.10 (0.75–5.87)	NA

The ORs for the genotypes reaching the level of significance are presented in **bold** (P < 0.05).

These SNPs had a minor allele frequency (MAF) of 14.3% (rs2476601, A allele) and 27.5% (rs3764147, G allele) in the control group, and affected individuals showed an excess of genotype pairs (AA, AG), (GA, GG) and (GA, AG), corresponding to epistatic model M11 proposed by Evans *et al.* (2006) [225]. Risks, relative to the most common homozygous genotype (GG, AA), are reported in Figure 3.4.1. For genotypes (GA, GG) and (GA, AG), the relative risks was significantly higher than 1: OR = 2.24 (95% CI: 1.09–4.64) and OR = 1.48 (95% CI: 1.02–2.16). Although the risk for the genotype (AA, AG) did not reach the level of significance (OR = 2.10 (95% CI: 0.75–5.87), P = 0.15) possibly because of its low frequency, its value was larger than 1. The joint OR, that combined the three at-risk genotypes, was 1.63 (95% CI: 1.16–2.29) and this was significantly larger than 1 ($P = 4.32\times10^{-3}$). These results have confirmed that carrying at least three minor alleles combining rs2476601 and rs3764147 elevates the risk for UC in the Lithuanian/Latvian sample set.

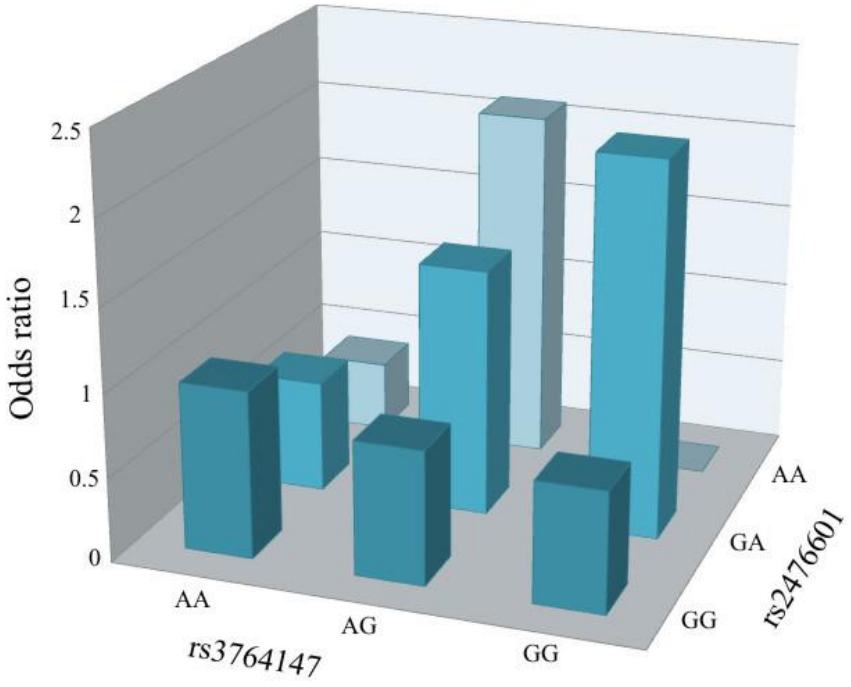

Fig. 3.4.1. *Odds ratio for the SNPs pair in ulcerative colitis (rs3764147, rs2476601) relative to the most common double homozygote genotype: (rs3764147, rs2476601) = (AA,GG)*

3.4.1. *In sicilo* prediction of *PTPN22* and *C13orf31* interactive network

The two genes showing possible interaction in the SNP-SNP epistasis analysis, *i.e. PTPN22* and *C13orf31* were entered into the GeneMANIA program. In the prediction process the co-expression, co-localization, genetic interactions, pathway, physical interactions and predicted networks were included. For the estimation of the networks weight the default network weighting method "Gene-Ontology (GO) based weighting, Molecular Process based" was chosen. This method assumes that the input gene list is related through GO molecular processes.

In the results generated by GeneMANIA 10 related genes, including the two input genes, were displayed (Fig. 3.4.1.1). The constructed composite network is a weighted sum of individual data sources; each edge (link) in the composite network is weighted by the corresponding individual data source.

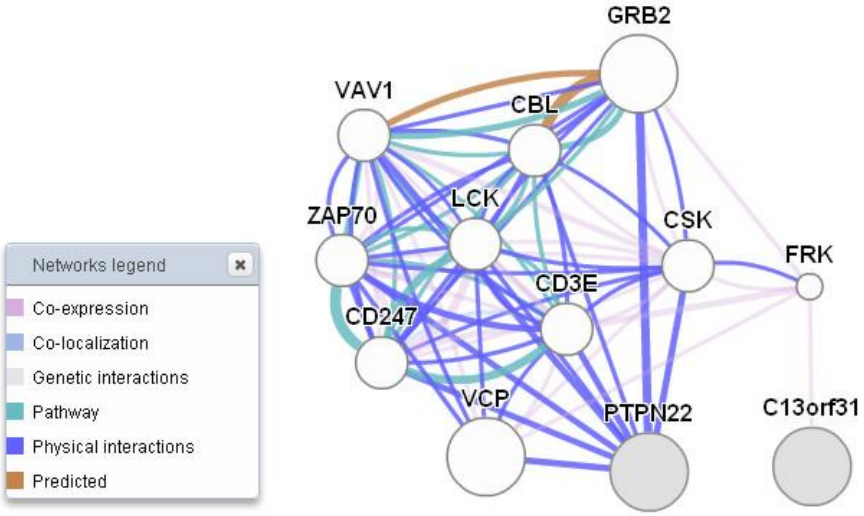

Fig. 3.4.1.1. *In silico* prediction of the possible *PTPN22* and *C13orf31* genes interactions.

C13orf31 – Chromosome 13 open reading frame 31, PTPN22 – protein tyrosine phosphatase non-receptor 22, FRK – fyn related kinase, CD3E – CD3e molecule, epsilon (CD3-TCR complex), VCP – valosin-containing protein, GRB2 – growth factor receptor-bound protein 2, CBL – Cas-Br-M (murine) ecotropic retroviral transforming sequence, VAV1 – vav 1 guanine nucleotide exchange factor, CSK – c-src tyrosine kinase, ZAP70 – zeta-chain (TCR) associated protein kinase 70kDa, LCK – lymphocyte-specific protein tyrosine kinase, CD247 – CD247 molecule.

The program assigned that the association network members were linked through these networks: physical interactions 90.22%, co-expression 3.50%, predicted 2.08%, pathway 2.13%, co-localization 1.86%, genetic interactions 0.21% (Appendix Table 14). The network weights sum to 100% and reflect the relevance of each data source for predicting membership in the query list. These scores were used to rank the genes (Appendix Table 15). The score assigned to each gene pair reflects how often paths that start at a given gene node end up in one of the query nodes and how long and heavily weighted those paths were. It has been determined that query genes were linked through co-expression (*FRK-C13orf31* (weight: 0.034), *FRK-CSK* (weight: 0.0069), *FRK-GRB2* (weight: 0.014), *FRK-VCP* (weight: 0.0046), *FRK-CD3E* (weight: 0.047)) and protein-protein interaction (*i.e.*, physical interaction; FRK-CSK (weight: 0.016), CSK-PTPN22 (weight: 0.057), GRB2-PTPN22 (weight: 0.159), CD3E-PTPN22 (weight: 0.068), VCP-PTPN22 (weight: 0.072)) pathways. Moreover, the program delivered the list of GO functional terms, in which the interactive network members were involved, ranked based on which function was the most statistically significant (Q-value) (Appendix Table 16). The most significant assigned functions include: T cell receptor complex ($Q = 6.4\times10^{-4}$), positive regulation of immune system processes ($Q = 1.1\times10^{-2}$), and positive regulation of T cell activation ($Q = 1.1\times10^{-2}$). These processes are mainly linked to the query gene *PTPN22*, as there are no existing functional annotations to the *C13orf31* gene.

3.5. Genetic risk profile

3.5.1. Genetic risk profile for ulcerative colitis

The SNPs that remained associated after Bonferroni's correction in 21q21.1 (rs1736135), 6q21 (rs7746082), *JAK2* (rs10758669), *ORMDL3* (rs2872507), *RNF186* (rs3806308) and markers from seven nominally associated loci (*HLA*, *IL23R*, *IL10*, *MST1*, 1p36.13, *NKX2-3*, *BSN*), that were previously associated with UC, were used to construct genetic dose-response risk models. The nominally associated SNPs in the *HLA* locus ($r^2>0.6$), *IL23R* ($r^2 = 0.72$) and *NKX2-3* ($r^2 = 0.97$) were in strong LD (Appendix Table 9). Therefore, only the most strongly associated SNPs were chosen for further analysis (*HLA*: rs9268877, rs9268858; *IL23R:* rs11209026; *NKX2-3*: rs11190140), which

resulted in 13 SNPs to be included. Two genetic risk profiles as explained in the chapter 2.4.6 were constructed.

The distribution of the number of risk alleles in the UC cases and healthy controls is shown in Figure 3.5.1.1. Independent samples t-test on the number of risk alleles in UC patients and controls showed a significant difference in the mean number of risk alleles carried by UC patients (mean±SD = 11.40±2.77) and controls (mean±SD = 9.93±2.57) ($P = 4.68\times10^{-23}$) and weighted score in UC patients (mean±SD = 11.35±2.58) and healthy controls (mean±SD = 10.03±2.42) ($P = 2.71\times10^{-21}$). This difference in the mean number of risk alleles was caused by a shift in the distribution of risk alleles between the two groups (Fig. 3.5.1.1).

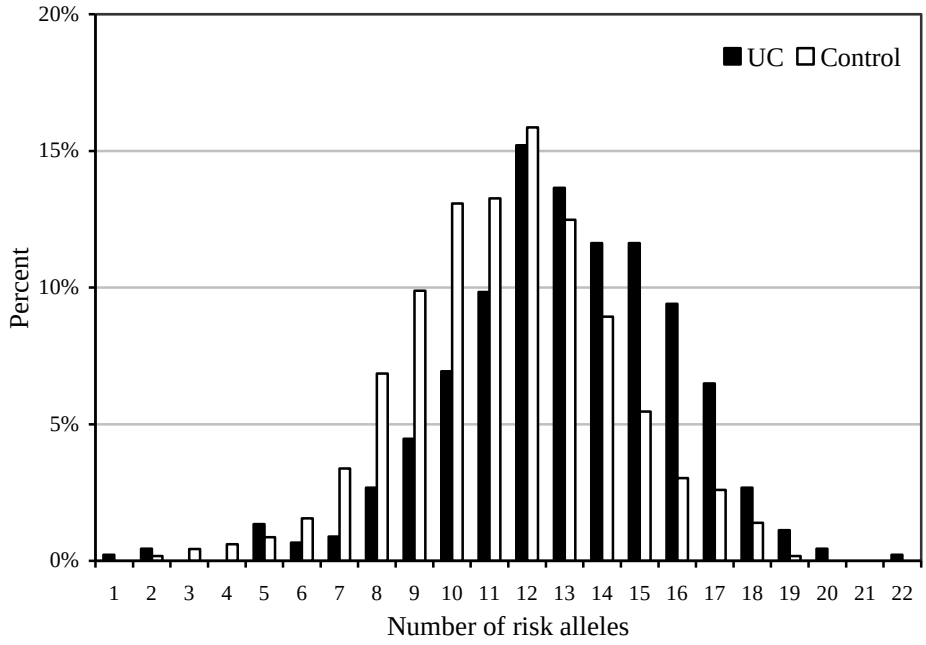

Fig. 3.5.1.1. *The distrubution of the number of risk alleles per individual for controls and UC cases*

Both in cases and controls, the number of risk alleles per individual follows a normal distribution, but in cases this normal distribution is shifted to the right.

Binary logistic regression showed that individuals with more than 11 risk alleles were at statistically significantly higher risk for UC compared with individuals carrying less than 8 risk alleles (Table 3.5.1.1). As an increasing

number of risk alleles were required to meet the criteria for a positive test, the sensitivity and negative predictive values decreased while the specificity and positive predictive values increased. For example, individuals carrying 19 or more risk alleles had an OR of 14.29 (95% CI: 12.39–16.49), high specificity (>98%), PPV (>70%) for UC susceptibility compared with the reference group.

Table 3.5.1.1. *Genetic risk profile of UC based on the number of risk alleles in 21q21.1, 6q21, JAK2, ORMDL3, RNF186, HLA, IL23R, IL10, MST1, OTUD3, NKX2-3, and BSN*

No. risk alleles	UC number (frequency)	Control number (frequency)	OR (95% CI)	P-value
≥9	20 (0.42)	114 (0.42)	1.00 (0.90–1.12)	4.97×10^{-1}
≥10	31 (0.53)	151 (0.49)	1.17 (1.02–1.36)	2.87×10^{-1}
≥11	44 (0.62)	153 (0.49)	1.64 (1.43–1.90)	$\mathbf{3.06\times10^{-2}}$
≥12	68 (0.71)	183 (0.54)	2.12 (1.85–2.45)	$\mathbf{1.11\times10^{-3}}$
≥13	61 (0.69)	144 (0.48)	2.42 (2.10–2.80)	$\mathbf{2.19\times10^{-4}}$
≥14	52 (0.65)	103 (0.4)	2.88 (2.51–3.33)	$\mathbf{2.39\times10^{-5}}$
≥15	52 (0.65)	63 (0.29)	4.72 (4.09–5.45)	$\mathbf{3.11\times10^{-9}}$
≥16	42 (0.6)	35 (0.18)	6.86 (5.95–7.92)	$\mathbf{1.49\times10^{-11}}$
≥17	29 (0.51)	30 (0.16)	5.52 (4.79–6.38)	$\mathbf{2.53\times10^{-8}}$
≥18	12 (0.3)	16 (0.1)	4.29 (3.72–4.95)	$\mathbf{1.90\times10^{-4}}$
≥19	5 (0.16)	2 (0.02)	14.29 (12.39–16.49)	$\mathbf{4.48\times10^{-5}}$

Table 3.5.1.1 continued

No. risk alleles	Sensitivity (%)	Specificity (%)	PPV	NPV	LR+	LR–
≥9	41.67	58.39	14.93	41.61	1.00	1.00
≥10	52.54	51.45	17.03	48.55	1.08	0.92
≥11	61.11	51.12	22.34	48.88	1.25	0.76
≥12	70.83	46.65	27.09	53.35	1.33	0.63
≥13	68.54	52.63	29.76	47.37	1.45	0.60
≥14	65.00	60.84	33.55	39.16	1.66	0.58
≥15	65.00	71.75	45.22	28.25	2.30	0.49
≥16	60.00	82.05	54.55	17.95	3.34	0.49
≥17	50.88	84.21	49.15	15.79	3.22	0.58
≥18	30.00	90.91	42.86	9.09	3.30	0.77
≥19	15.15	98.77	71.43	1.23	12.27	0.86

UC – ulcerative colitis, OR – odds ratio, 95% CI – 95% confidence interval, $P < 0.05$ are highlighted in **bold**; PPV – positive predictive value, NPV – negative predictive value, LR+ – positive likehood ratio, LR– – negative predictive value.

The sensitivity, however, dropped to less than 15%. In contrast, when the cutoff was set at 12 risk alleles the OR = 2.12 (95% CI: 1.85–2.45), the sensitivity was more than 70%, whereas specificity dropopped to less than 50%, PPV was less than 30%. This indicates that it is highly unlikely that an individual with fewer than 12 risk alleles has UC. The likelihood ratio of a positive test (LR+) is the ratio between the chance of a positive test in cases and controls, and gives the

Table 3.5.1.2. *Genetic risk profile of UC based on a weighted score for the strength of association, using the coefficients of regression analyses to attribute scores to each risk allele for 21q21.1, 6q21, JAK2, ORMDL3, RNF186, HLA, IL23R, IL10, MST1, OTUD3, NKX2-3, and BSN*

Score	UC number (frequency)	Control number (frequency)	OR (95% CI)	P-value
≥9	31 (0.74)	163 (0.72)	1.03 (0.92–1.17)	4.51×10^{-1}
≥10	49 (1.17)	176 (0.78)	1.51 (1.35–1.71)	$\mathbf{3.76\times10^{-2}}$
≥11	74 (1.77)	197 (0.87)	2.04 (1.81–2.30)	$\mathbf{4.35\times10^{-4}}$
≥12	62 (1.48)	157 (0.69)	2.14 (1.91–2.42)	$\mathbf{3.04\times10^{-4}}$
≥13	61 (1.46)	104 (0.46)	3.18 (2.83–3.59)	$\mathbf{1.72\times10^{-7}}$
≥14	61 (1.46)	62 (0.28)	5.34 (4.75–6.02)	$\mathbf{5.59\times10^{-13}}$
≥15	45 (1.08)	42 (0.19)	5.82 (5.17–6.56)	$\mathbf{4.13\times10^{-12}}$
≥16	13 (0.31)	23 (0.11)	3.07 (2.73–3.46)	$\mathbf{1.28\times10^{-3}}$
≥17	7 (0.17)	2 (0.01)	19.00 (16.87–21.41)	$\mathbf{6.96\times10^{-7}}$

Table 3.5.1.2 continued

Score	Sensitivity (%)	Specificity (%)	PPV	NPV	LR+	LR–
≥9	42.47	58.31	15.98	71.49	1.02	0.99
≥10	53.85	56.44	21.78	77.19	1.24	0.82
≥11	63.79	53.65	27.31	86.40	1.38	0.67
≥12	59.62	59.22	28.31	68.86	1.46	0.68
≥13	59.22	68.67	36.97	45.61	1.89	0.59
≥14	59.22	78.62	49.59	27.19	2.77	0.52
≥15	51.72	84.44	51.72	18.42	3.33	0.57
≥16	23.64	90.84	36.11	10.09	2.58	0.84
≥17	14.29	99.13	77.78	0.88	16.43	0.86

UC – ulcerative colitis, OR – odds ratio, 95% CI – confidence interval, $P < 0.05$ are highlighted in **bold**; PPV – positive predictive value, NPV – negative predictive value, LR+ – positive likehood ratio, LR– – negative predictive value.

likelihood that a patient has UC when a diagnostic test (in this case a genetic risk profile) is positive. In our cohort, patients with 19 or more risk alleles had a LR+ of 12.27. This gives a moderate increase in the likelihood of disease.

As different genetic variants have different effects on disease susceptibility, we have decided to perform the same analysis on the basis of the β-coefficients. β-coefficients were calculated from separate binary logistic regression for each UC-associated SNP, as explained in the Methods section (chapter 2.4.6). This model showed an even larger increase of the OR for disease susceptibility with an increase of the weighted score in risk alleles (Table 3.5.1.2). As reference, we used a group of individuals with a weighted score in risk alleles of 8 or less. Individuals with a weighted score in risk alleles over 17 had an OR for UC susceptibility of 19 (95% CI: 16.87–21.41) compared with this reference group. Moreover, having a score above 17.0 had a high specificity (>99 %), PPV (>77%) for UC, and LR+ of 16.43. The sensitivity, however, dropped to less than 15%. On the contrary, when the cutoff was set at a weighted score of less than 11, the sensitivity was more than 60% and specificity approx. 50%, however, NPV was more than 80%, PPV was less than 30% and LR+ was less than 1.40.

3.5.2. Genetic risk profile for Crohn's disease

The SNPs that remained associated with CD after a correction for multiple testing in *NOD2* (rs2066847, rs2076756) and nominally associated SNPs in loci: 21q21.1, *NOD2* and *IRGM*, that were associated with CD in previous association studies, were used to construct two genetic risk models. The two nominally associated SNPs in the *IRGM* locus were in strong LD ($r^2 = 0.97$), therefore, only the most strongly associated SNP, *i.e.*, rs13361189, was taken; whereas the four SNPs in the *NOD2* locus were associated with the disease independently ($r^2<0.2$) (Appendix Table 9). In the result, six SNPs were used for the genetic risk models construction.

The distribution of the number of risk alleles in the CD cases and healthy controls is presented in Figure 3.5.2.1. Independent samples t-test on the number of risk alleles in CD patients and controls showed that CD patients (mean±SD = 3.76±1.78) had more risk alleles than controls (mean±SD = 2.77±1.45) ($P = 1.78\times10^{-12}$) and CD patients (mean±SD = 6.98±2.95) had a higher weighted score than healthy controls (mean±SD = 5.41±2.66) (P =

5.85×10^{-10}). This difference in the mean number of risk alleles was caused by a shift in the distribution of risk alleles between the two groups (Fig. 3.5.2.1).

Using different cutoffs, measures of the two CD genetic risk scoring models for CD were calculated, *i.e.*, sensitivities, specificities, PPV and NPV, and positive and negative likelihood ratios (LR+ and LR–, respectively), as shown in Tables 3.5.2.1 and 3.5.2.2. Strong associations with CD were seen with "possession" of at least 3, 4, 5, 6, and 7 risk alleles (Table 3.5.2.1) and any score greater than 5 (using the coefficients of regression analyses to attribute scores to carriage of each risk genotype) (Table 3.5.2.2). As an increasing number of risk alleles/points were required to meet the criteria for a positive test, the sensitivity and negative predictive values decreased while the specificity and positive predictive values increased.

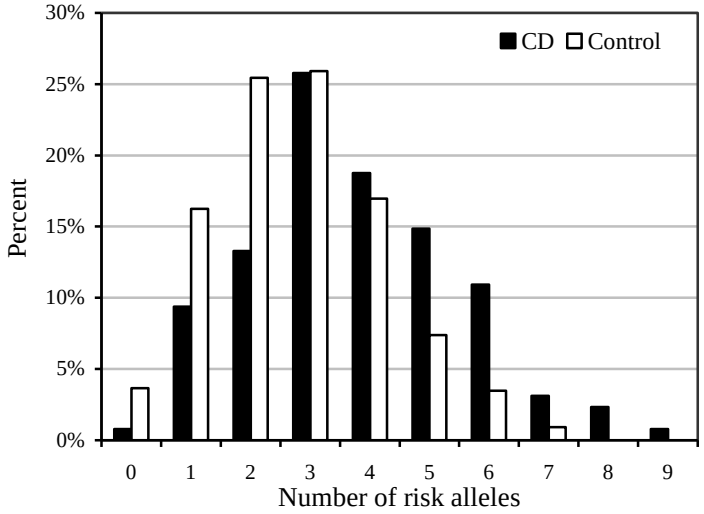

Fig. 3.5.2.1. *The distrubution of the number of risk alleles per individual for controls and CD cases*

Both in cases and controls, the number of risk alleles per individual follows a normal distribution, but in cases this normal distribution is shifted to the right.

In the first model, binary logistic regression showed that individuals with a higher number of risk alleles were at higher risk for CD compared with individuals carrying less than one risk alleles (Table 3.5.2.1). For example, individuals carrying 7 or more risk alleles had an OR of 6.71 (95% CI: 5.90–

7.64), high specificity (>95%), and PPV (approx. 30%) for CD susceptibility compared with the reference group. The sensitivity, however, dropped to less than 25% and NPV <5%. The likelihood ratio of a positive test (LR+) in our cohort of patients with seven or more risk alleles was 5.36. In contrast, a maximum sensitivity (>70%) was achieved with possession of 3 or more risk alleles, although the specificity was at 43.43%, OR was 1.95 (95% CI: 1.72–2.22), PPV was at 10.41%, and NPV – 56.57%. This indicates that it is highly unlikely that an individual with fewer than three risk alleles has CD.

Table 3.5.2.1. *Genetic risk profile of CD based on the number of risk alleles in 21q21.1, NOD2, and IRGM*

No. risk alleles	CD number (frequency)	Control number (frequency)	OR (95% CI)	P-value
≥2	17 (0.57)	279 (0.57)	1.02 (0.90–1.17)	4.77×10^{-1}
≥3	33 (0.72)	284 (0.57)	1.95 (1.72–2.22)	$\mathbf{2.31\times10^{-2}}$
≥4	24 (0.65)	186 (0.47)	2.16 (1.91–2.47)	$\mathbf{1.41\times10^{-2}}$
≥5	19 (0.6)	81 (0.28)	3.93 (3.46–4.48)	7.83×10^{-5}
≥6	14 (0.52)	38 (0.15)	6.18 (5.44–7.03)	1.16×10^{-6}
≥7	4 (0.24)	10 (0.05)	6.71 (5.90–7.64)	5.18×10^{-4}

Table 3.5.2.1 continued

No. risk alleles	Sensitivity (%)	Specificity (%)	PPV	NPV	LR+	LR–
≥2	56.67	43.86	5.74	56.14	1.01	0.99
≥3	71.74	43.43	10.41	56.57	1.27	0.65
≥4	64.86	53.96	11.43	46.04	1.41	0.65
≥5	59.38	72.91	19.00	27.09	2.19	0.56
≥6	51.85	85.16	26.92	14.84	3.49	0.57
≥7	23.53	95.61	28.57	4.39	5.36	0.80

CD – Crohn's disease, OR – odds ratio, 95% CI – confidence interval, $P < 0.05$ are highlighted in **bold**; PPV – positive predictive value, NPV – negative predictive value, LR+ – positive likehood ratio, LR– – negative predictive value.

The second model, based on the β-coefficients calculated from separate binary logistic regression for each CD-associated SNP showed similar results as the first model (Table 3.5.2.2). As reference, we used a group of controls with a weighted score of 3 or less. Individuals with a weighted score in risk alleles over 10 had an OR for CD susceptibility of 7.86 (95% CI: 7.00–8.85) compared with this reference group. Moreover, a score above 10 resulted in high

specificity (>90 %), sensitivity (approx. 40%), LR+ of 5.07, and NPV – less than 9%. The PPV of this test, however, dropped to 30% and further decreased with the increase of the weighted score. On the contrary, when the cutoff was set at a weighted score of less than four, the sensitivity and specificity was more than 50%. However, NPV reached even more than 90%, PPV was less than 7% and LR+ was 1.08.

Table 3.5.2.2. *Genetic risk profile of CD based on a weighted score for the strength of association, using the coefficients of regression analyses to attribute scores to each risk allele for 21q21.1, NOD2, and IRGM*

Score	CD number (frequency)	Control number (frequency)	OR (95% CI)	P-value
≥4	17 (1.07)	242 (0.93)	1.15 (1.03–1.30)	3.44×10^{-1}
≥5	10 (0.63)	94 (0.36)	1.75 (1.56–1.97)	$\mathbf{8.97\times10^{-2}}$
≥6	25 (1.57)	204 (0.78)	2.01 (1.80–2.27)	$\mathbf{1.64\times10^{-2}}$
≥7	18 (1.13)	116 (0.45)	2.55 (2.27–2.87)	$\mathbf{3.85\times10^{-3}}$
≥8	11 (0.69)	76 (0.29)	2.38 (2.12–2.68)	$\mathbf{1.57\times10^{-2}}$
≥9	10 (0.63)	55 (0.21)	2.99 (2.66–3.37)	$\mathbf{4.02\times10^{-3}}$
≥10	11 (0.69)	23 (0.09)	7.86 (7.00–8.85)	$\mathbf{8.97\times10^{-8}}$
≥11	5 (0.32)	16 (0.07)	5.14 (4.57–5.78)	$\mathbf{8.72\times10^{-4}}$
≥12	1 (0.07)	7 (0.03)	2.35 (2.09–2.65)	2.12×10^{-1}

Table 3.5.2.2 continued

Score	Sensitivity (%)	Specificity (%)	PPV	NPV	LR+	LR–
≥4	51.52	52.08	6.56	92.02	1.08	0.93
≥5	38.46	73.67	9.62	35.74	1.46	0.84
≥6	60.98	56.32	10.92	77.57	1.40	0.69
≥7	52.94	69.39	13.43	44.11	1.73	0.68
≥8	40.74	77.58	12.64	28.90	1.82	0.76
≥9	38.46	82.70	15.38	20.91	2.22	0.74
≥10	40.74	91.96	32.35	8.75	5.07	0.64
≥11	23.81	94.27	23.81	6.08	4.15	0.81
≥12	5.88	97.41	12.50	2.66	2.27	0.97

CD – Crohn's disease, OR – odds ratio, 95% CI – confidence interval, $P < 0.05$ are highlighted in **bold;** PPV – positive predictive value, NPV – negative predictive value, LR+ – positive likehood ratio, LR– – negative predictive value.

4. DISCUSSION

Genetic association analysis has become a common task in human genetics and human disease studies. The higher statistical power of genetic association studies compared with pedigree linkage analysis [86]; thorough genomic infrastructure (*i.e.*, complete DNA sequence of the human genome [90, 226], the location of SNP genetic markers [97]); high-throughput genotyping technologies [104, 105] providing possibilty to carry out candidate-gene, regional, or whole-genome association studies easier and in more cost-effective manner; and development of statistical methods helping to overcome main genetic analysis issues (*i.e.*, population stratification and spurious association signals) [114, 116–119] are several reasons for this trend.

In the past five years, GWAS in CD and UC have identiffied a number of new susceptibility genes. In the frames of this study we chose six studies that undertook genome wide analysis of CD [58, 67, 152, 159] and UC [161, 163, 167]. As it is essential that such associations are confirmed in independent cohorts, we undertook the current study in a large Lithuanian-Latvian IBD cohort. This is the first comprehensive analysis of the contribution of previously defined multiple genetic risk factors to the onset of CD and UC in low-incidence populations [11, 35] of the North-Eastern Europe – Lithuania and Latvia. Baltic countries still observe low IBD incidence rates, especially for CD in their populations. In Lithuania (2006) – 2.0 per 100 000 inhabitants [11, 35]; and in Estonia (1993–1998) the incidence rate of CD was reported to be 1.4 per 100 000 inhabitants [36]. Therefore, analysis of the genetic contribution to disease susceptibility in this region was of great interest.

Due to the small to moderate effect sizes that characterize susceptibility genes for complex diseases and multi-factorial traits large sample sizes are needed in order to reach the required study power [115]. Collaborations involving sample collection are therefore essential. In the frames of this study we have arranged collaborations with the biggest gastroenterology centers in Lithuania and Latvia, what enabled the recruitment of more than 500 IBD patients and more than 1000 healthy controls. Involving the cohorts from different countries for genetic epidemiological research, the problem of confounding by population stratification has to be addressed [113, 114]. Heterogeneity between studied samples can give false-positive results in association studies, as association with the trait may be the result of the systematic ancestry difference in allele frequncies between groups [116]. The

analysis of the populations‘ genetic differences analysis in Europe [124], investigating the detailed structure of the Baltic countries and other North-Eastern European populations, revealed that the three Baltic countries (Lithuania, Latvia, Estonia), Poland and Western Russia together form a genetic cluster (inflation factor $\lambda = 1.23$), thereby indicating that our two study populations can be combined in association analysis. Moreover, we used Cochran-Mantel-Haenzsel and Breslow-Day tests to assess disease-associated regions heterogeneity between the two study populations.

In the single-marker case-control association analysis we have identified the association of *NOD2* with the Lithuanian CD study sample of 131 cases and 1097 controls. The association reached genome-wide significance ($P < 10^{-7}$). The rs2066847 and rs2076756 variants were responsible for the major contribution of *NOD2* to disease susceptibility in the Lithuanian CD population (rs2066847: MAF = 15.6%, OR = 4.52 (95% CI: 3.02–6.78); rs2076756: MAF = 35%, OR = 2.24 (95% CI: 1.69–2.97)). The risk of disease in homozygous variant allele carriers of the two strongest associations of the *NOD2* was substantially (rs2066847: 80% ($OR_{HOM} = 21.74$); rs2076756: 40% ($OR_{HOM} = 4.18$)) higher than the risk of single variant allele carriers, suggesting a dose effect.

The rs2066847 variant is one of the three initially discovered [54, 55] and therefore, mostly studied variants in the *NOD2* gene. The reported MAFs and the contributable rik of the rs2066847 is consistent with previous reports from Central Europe and North America (MAF = 6.6%–16%) [150]. However, the data contrasts markedly with studies performed in Northern Europe, where carriage rates of rs2066847 and other *NOD2* variants are relatively low, *i.e.*, the carriage of at least one *NOD2* variant varies from 2.8% to 22% [227, 228].

We have also identified the associations with the other two *NOD2* SNPs (rs10521209, rs2066845), but the associations were only nominally significant and did not withstand correction for multiple testing. The rs2066845 also belongs to the trio of the mostly studied variants in the *NOD2* gene. The reported allele frequency of rs2066845 in our study sample is similar to previously reported ones in the Southern and Central European populations (CD: 3.3%–6.1%; controls: 0.6%–3.0%) [150]. However, we were not able to confirm the association between rs2066844 (the third member of the initially discovered trio) and IBD susceptibility in our study sample. The rs2066844 MAFs in both the cases (3.2%) and control (2.6%) groups were lower compared to previously reported ones in Southern and Central European

populations, where a positive association between rs2066844 and CD was detected (CD: 6.7%–12.5%, controls: 3.5%–6.9%) [150].

The *NOD2*, also referred to as caspase-activation recruitment domain containing protein 15 (*CARD15*), is the first susceptibility gene for CD that has been identified [54, 55]. The NOD2 belongs to a Nod1/Apaf-1 superfamily of apoptosis regulators and is primarily expressed in peripheral blood leukocytes [229]. NOD2 is an intracellular protein with a modular tripartite domain structure, characterized by a central nucleotide binding domain, C-terminal leucine-rich repeats (LRR) and two N-terminal CARD domains [229]. The LRR domain functions as a pattern recognition receptor of bacterial components. So far, more than 60 variations in this gene have been identified. The three common CD-associated variants of *NOD2* (R702W (rs2066844), G908R (rs2066845), and 1007fsinsC (rs2066847)) are located in the C-terminal portion [55]. Functional studies indicate that these variants lead to an inappropriate response to bacterial components altering signalling pathways in the innate immune system (lack of appropriate nuclear factor κB activation) and ultimately causing intestinal inflammation [229, 230].

Since 2001, a significant number of studies have replicated the association of the *NOD2* variants with the development of CD in populations of Caucasian origin from Europe and North America [150]. We have also reported strong association of the *NOD2* rs2066847 variant in our first report on the prevalence of the previously defined *NOD2*, *ATG16L1* and *IL23R* disease associated variants in an IBD case-control sample from Lithuania [20]. Moreover, the determined PAR% of *NOD2*, an indication of the contribution of a mutation to the disease in a specific area, was 29.5% in the Lithuanian population and was similar to the the Central European populations and North America reporting PAR% around 30% [55, 150, 231]; whereas the other Northern European populations reported lowest PAR% ranging: 1.88%–11% [227, 228]. However, significant heterogeneity in the frequencies of these variants has been observed not only between ethnically divergent populations [232, 233], but also within Europe [150].

Moreover, it has been previously reported that *NOD2* mutations have a dose dependent effect as mutated homozygotes and compound heterozygotes are found more frequently in CD than expected [224]. The results of this study and our first genetic study [20] confirmed that CD in Lithuania has a strong genetic background that is related partially to *NOD2* susceptibility variants. Interestingly, the relatively high carriership frequency of the *NOD2* alleles in

the healthy controls (11.12%) in our study is in contrast with data of low CD incidence in Lithuania [20, 35]. These data are in concordance with previously reported rates of 30%–50% in CD and 7%–20% in controls from other European regions [150]. Moreover, this indicates the importance of environmental factors (*e.g.*, diet, lifestyle) in disease development.

We also identified the associations with several other CD associated loci (*IRGM* (rs13361189, rs11747270), *CCR6* (rs2301436)), and UC associated SNPs (*IL10* (rs3024505), *HLA* (rs9268858, rs2395185, rs9268877), *S100Z* (rs7712957)) that demonstrated only moderate association with CD in previous studies [161], but the associations were only nominally significant and did not withstand correction for multiple testing. However, we failed to replicate previously described CD associations with *IL23R, ATG16L1, IL12B, NKX2-3, STAT3, NELL1, 5p13, PTPN22, etc.*

It must be noted that our relatively small CD study population was underpowered to demonstrate such weak to moderate disease associations. The panel had a power of 80% to detect an OR of 1.8 or higher at the 5% significance level, assuming a frequency of the disease associated allele of at least 20% in the controls. Therefore, larger-sized CD case-control panels will be needed in order to further evaluate the importance of the herein tested loci.

In the UC case-control study we have identified five SNPs tagging five genetic risk loci as associated with UC in a Lithuanian-Latvian study sample of 447 cases and 1154 controls. We confirmed the association with *RNF186* (rs3806308). This association was first discovered in UC GWA study by Silverberg *et al.* (2009) [163] and only recently replicated also in UC GWAS study by McGovern *et al.* (2010) [169]. Although RNF186 is a protein with unknown function, it contains RING protein domain that have been associated with protein ubiquitination [234]. The study of the McGovern *et al.* (2010) [169] for the first time explored the expression pattern of RNF186. They discovered that it was higher in intestinal tissues, specifically at the basal pole of epithelial cells and lamina propria within colonic tissues, than in immune tissues. This indicates the possible involvement of RNF186 in intestinal barrier functions. However, the exact functions of the protein remain to be unclear.

The other four loci (*JAK2* (rs10758669), *ORMDL3* (rs2872507), 6q21 (rs7746082) and 21q21 (rs1736135)), that provided strong association with UC in our study sample, were previously reported to be strongly associated with CD [67, 153, 173, 235, 236] and other immune-mediated diseases [170]. Our strongest UC associations 21q21 (rs1736135; intergenic region *NRIP1*,

CYCSP42), 6q21 (rs7746082; near *PRDM1*), and *ORMDL3* (rs2872507) for the first time were associated with UC only in the recently performed UC GWAS study by the McGovern *et al.* (2010) [169] and replicated in the UC meta-analysis [170]; whereas the number of the UC replication and GWAS studies failed to confirm these associations [163, 167, 173, 183]. Moreover, the 21q21 locus was also nominally associated with CD in our study population. The frequencies of the *JAK2* C allele reported in our study (35.5% controls and 43.4% UC) were similar to the published studies performed in Germany (41% UC and 35% controls) [167], UK (38% UC and 33.6% controls) [183], and Sweden (34.3% controls) [173]. One study did not confirm this associations in the UC cohort [163].Thus, our study confirms that these loci are involved in the general IBD pathogenesis.

JAK2 is a gene encoding an signaling component up-stream of STAT3. JAK-STAT signaling pathway is important for cytokine and growth factor downstream signalling. This transmission pathway coordinates multiple signaling events in T cells leading to their differentiation into distinct subpopulations as well as regulation of pro- and anti-apoptotic cascades [237]. *JAK2* belongs to a gene network which is typically referred to as the "IL-23 pathway" [238]. Moreover, JAK2 is closely related to STAT3, therefore it also influences IL-17 signaling [239]. Unfortunately, the exact biological implication of polymorphisms in *JAK2* have not been investigated yet.

Rs2872507 was shown to be associated with expression levels of the closely linked *ORMDL3* gene in lymphoblastoid cell lines, which therefore stood out as prime candidate gene [67, 240]. *ORMDL3* has been implicated in the pathogenesis of many diseases involving dysregulated immune responses such as asthma [240, 241], reumathoid arthritis [242], primary biliary cirrhosis [243], and ankylosing spondylitis [244], although the underlying mechanisms of this association remain unclear. *ORMDL3* is expressed ubiquitously, particularly high expression levels are recorded in cells participating in the inflammatory response [240, 245] and immune tissues [169], whereas the expression pattern in the intestinal tissue revealed no difference when comparing CD, UC to healthy controls [169, 244]. The ORMDL3 protein is thought to be involved in protein folding, and growing evidence indicates that there are interactions between the unfolded protein response (UPR) and immune responses [169, 246, 247]. Overexpression of *ORMDL3* decreased both the basal and ER-stress-induced UPR, whereas knockdown of *ORMDL3* expression induced a higher

UPR, thereby indicating that *ORMDL3* expression levels can regulate UPR and that ORMDL3 might be an important factor in ensuring ER homeostasis [169].

The 21q21 locus SNP rs1736135 is located in the intergenic region between *NRIP1* (also known as *RIP140*) and *CYCSP42*. *CYCSP42* is a somatic cytochrome c pseudogene [254]. Processed pseudogenes are disabled copies of functional genes that do not produce a functional, full length protein [254]. RIP140 has been characterized as a nuclear receptor cofactor, interacting with a number of nuclear receptor family members, such as peroxisome proliferator activated receptors, liver X receptor, estrogen receptor-related receptor, and estrogen receptor [255]. A recent study has indicated a coactivating function of the RIP140 in the control of nuclear factor κB dependent proinflammatory gene expression, thereby revealing the important role of this protein in the inflammatory processes [256]. However, further studies are needed to investigate the functional consequences of polymorphism in this locus.

The recent UC meta-analysis provided functional annotations to the number of the UC-associated loci [170]. One of these loci was 6q21 (rs7746082) which lays upstream the *PRDM1* gene. *PRDM1* encodes a transcriptional repressor B lymphocyte-induced maturation protein-1 (BLIMP1). BLIMP1 is expressed in B and T cells, granulocytes, macrophages, epithelial cells, and germ cells [248–250]. This protein is a master transcriptional regulator of plasma cells [251]. It also functions in T cells to attenuate IL2 production upon antigen stimulation [252] and to promote the development of short-lived effector cells and regulate clonal exhaustion in both CD4 and CD8 cells [253].Thereby, BLIMP1 plays an important role in the proliferation, survival and differentiation of B and T cells.

We also showed nominal associations with the previously reported UC risk SNPs in *HLA* (rs9268877, rs2395185, rs9268858, rs9268480), *IL10* (rs3024505), *IL23R* (rs11209026), *NKX2-3* (rs11190140, rs10883365), and *MST1* (rs3197999) loci. Although our study had relatively high power to replicate these associations at nominal significance level (approx. 80%), at the significance level of $P < 8\times10^{-4}$ (*i.e.,* P-value after Bonferroni correction) the replication power of associations dropped to approx. 23%–65%. Furthermore, in our study reported alleles frequencies distribution and contributable risk of the SNPs in aforementioned loci were similar to previous reports in other Caucasian populations [68, 161, 163, 168, 173, 174, 180, 184], thereby indicating that the increased study sample would improve study power and the possibility to replicate the associations. A number of previously reported risk loci, icluding *STAT3, IL12B, PTPN2, NELL1, ECM1,* and *ARP2C* were not

replicated in our Lithuanian-Latvian UC study sample. This does not necessarily mean that these are not truly UC associated genes; it may merely reflect a lack of statistical power in our moderate sized UC study sample.

Taken together, our study results support the previously proposed functional implications of the genetic associations in the resolution of inflammation in the pathogenesis of UC [164–166, 169, 170], *i.e.,* the importance of gene sets that have an important role in alterations of barrier functions, transcriptional regulation, cell-specific innate responses, and regulatory functions in adaptive immunity.

The relationship of genotype to phenotype is a fundamental problem in the genetics of complex disorders. Through these investigations it is hoped that deeper understanding of the phenotypic expression as well as disease susceptibility will be gained. The precise diagnostic classification and collection of complete clinical and demographic data maximizes one‘s ability to identify disease susceptibilty genes or disease modifier genes, which do not alter risk of the disease itself just the expression [257]. It has been hypothesized that IBD is not a single or even two diseases (*e.g.,* CD and UC), but rather is likely to be composed of subsets of disorders presenting within the broad clinical picture of CD or UC, and that these distinct diseases may have different pathogenic mechanisms and may require distinct therapies for successful treatment [258]. The numerous genotype–phenotype studies of IBD have revealed that a number of clinical characteristics (*e.g.*, age of disease onset, disease involving a specific part of the bowel, extraintestinal manifestations) may be inherited and influenced by disease suscpetibility genes [150, 257]. In the frames of this study we also performed the analysis of the possible genotype associations with the IBD phenotypes. The SNPs that showed at least nominal significance in the single marker analysis were included into the genotype-phenotype analysis.

In the CD phenotype analysis we found that *NOD2* polymorphism (rs2066847) was associated with the increased risk for the upper GI involvement in CD sample set (OR = 6.38 (95% CI: 2.10–19.38). Upper GI involvement is uncommon (1.7%–10%) [259], *e.g.,* in our study population we had only one patient having isolated form of the upper GI CD (0.8%). One of the features of CD is segmental involvement and in our study population we had six patients with the combined upper GI and intestine involvements (4.6%). The possible association of the *NOD2* with upper GI involvement has not been stated previously. However, *NOD2* was reported to be associated with the increased risk of ileal involvement, young age of disease onset and complicated

forms of disease: stricturing and penetrating manifestations or need for surgery treatment [150]. In our study, only the trend for *NOD2* association was shown with stricturing disease behavior and extraintestinal manifestations. However, after corrections for multiple testing none of these associations remained significant. As noted above, our CD study population is relatively small and has little power; therefore, an increased CD study population is needed to confirm or reject these associations.

Furthermore, we have shown associations between the SNPs (rs9268858, rs9268877, rs2395185) in the intergenic region of the *HLA* locus (*BTNL2*, *HLA-DRA*, *HLA-DRB1*, *HLA-DRB5*, *HLA-DRB6*) and the increased risk for CD perianal localization; whereas in the UC patient group the association between the SNP in the *BTNL2* gene increased the risk for extensive colitis compared to left-sided colitis. Previous studies have also shown the association of the *HLA* classII alleles (*HLA-DRB1*0103*; *HLA-DRB3*0301 – HLA-DRB1*1302*) with the perianal disease behavior in CD patients and extensive colitis in UC [260, 261]. The perianal manifestations of CD are among the most devastating and mutilating complications [24] and it has been reported in 13% to 43% of patients with CD [262]. Extensive colitis is the most severe form of UC, with the inflammation spread throughout the entire large intestine. This form of UC is being diagnosed in approximately 20% of cases [5]. The *HLA* region located on chromosome 6q21.3 is a highly polymorphic gene dense region with complex pattern of LD. The class I and II *HLA* genes are essential for normal lymphocyte function, and a number of immunoregulatory functions. Considering the central role played by the immune system in mediating tissue damage in IBD, *HLA* class II genes are good candidates for conferring a distinct clinical phenotype to patients with IBD [263]. Since the first report of the *HLA* association with IBD in 1972 [264], a number of linkage and association studies have investigated the role of *HLA* genes in determining susceptibility and phenotype of IBD. The *HLA* locus has also been previously associated with extraintestinal manifestations, colonic and ileal disease locations in CD, disease behavior in CD, age of disease onset, need for surgery treatment, more aggressive clinical course in UC (extensive colitis, colectomy, extraintestinal manifestations), and for failed medical therapy [260, 263, 265]. In our study, only the trend for *HLA* association was shown with stricturing disease behavior and ileal disease location in CD study sample. However, after corrections for multiple testing none of these associations remained statistically significant.

The polymorphism located in an intergenic region proximal to the 3'UTR end of the *IL10* gene (rs3024505) in our study has been found to be associated with the increased risk for left-sided UC and pancolitis. This polymorphism has been previously associated with an increased risk in developing UC [161]; whereas in our study only the nominal association with this SNP has been revieled. However, the possible links between the rs3024505 and the phenotype of UC have not been analysed previously. The biological significance of rs3024505 in IBD remains unclear [161]. The region has a high regulatory potential score (AP-1 binding motif) and may thus regulate *IL10* gene expression [161]. Furthermore, rs3024505 is in perfect linkage with other polymorphisms located within the *IL10* gene [161]. IL10 is a pleiotropic cytokine with potent anti-inflammatory properties that are important for immunoregulation of many of the cell-types in the immune system. *IL10* knock-out mice develop colitis if they are not kept in germ-free environment [266], and the administration of IL10 ameliorates the inflammation in animal and *in vitro* models [267]. A recent study indicated that gnotobiotic IL10-deficient (–/–) mice in the presence of two commensal bacteria (nonpathogenic *Enterococcus faecalis* or a nonpathogenic *Escherichia coli* strains) develop aggressive pancolitis and duodenal inflammation [268]. In patients, an impaired IL10 production has been found in severe cases of CD and UC [269, 270]. These functional findings support our current association of rs3024505 with the extended forms of UC.

Further, we found a novel association between the rs2872507 and increased risk for extraintestinal manifestations in particular - joints involvement in UC patients. As stated above, in the initial studies this SNP has been strongly associated with CD [67, 153, 173, 235, 236] and only moderately – with UC; whereas in our UC analysis it was one of the top candidate genes. As stated above this SNP was shown to influence the expression levels of the closely linked *ORMDL3* gene in lymphoblastoid cell lines, which therefore stood out as prime candidate gene [67, 240]. The association of this SNP has not been stated previously, however, the association with the pathogenesis of diseases involving joints has been published recently (*i.e.*, reumathoid arthritis [242], and ankylosing spondylitis [244]).

The 1p36.13 locus SNP rs6426833 and 6q21 locus SNP rs7746082 were robustly associated with an increased risk of the need for biological therapy and/or colectomy or other type surgery treatment in the UC patients group. These phenotypic associations are novel. However, the functional implications

of the polymorphisms have not been defined. The rs6426833 polymorphism is located within an approximately 100 kb region upstream the *OTUD3* and *PLA2G2E* genes. The OTUD3 is expressed broadly and has homology to an OTU-like cysteine protease [271]. PLA2G2E is a member of the secretory phospholipase A2 family of proteins that release arachidonic acid from membrane phospholipids, which leads to the production of proinflammatory lipid mediators, such as prostaglandins and leukotrienes [272]. Furthermore, PLA2G2E expression in the lung and small intestine is induced with lipopolysaccharide stimulation, suggesting a role in bacterially associated inflammation [272, 273]. Rs7746082 lays upstream the *PRDM1* gene, that plays important functions in T cell proliferation, survival and differention through repression of IL2 (key mediator of T cell proliferation). TNF-α inhibitors have been found to be critical for T-cell viability and activation. The recent study analysing naive human T cells reported that infliximab treatment inhibits proliferation of human T cells during T cell receptor (TCR) directed stimulation and that this inhibitory effect is caused by IL2 deprivation [298]. However, further studies will be required to resolve the functional implications (in detali) of the genes that lay in the vicinity of SNPs for the occurance of the phenotype.

Finally, we revealed a novel association between the SNPs in the autophagy gene *IRGM* (rs13361189, rs11747270) and increased risk for the severe forms of the disease, having poor medical outcome, *i.e.*, stricturing form of the disease, the necessity for biological therapy and/or surgery treatment during the course of the CD. Recently, autophagy has been shown to be a key process in the innate immune response against cytoplasmic constituents, including intracellular pathogens [274]. Autophagy also has been linked to the adaptive immunity. Facilitating endogenous major histocompatibility complex class II antigen presentation has been shown to have a critical role in modulating CD4+ T-cell responses [275]. Previous studies analyzing the *IRGM* phenotypic outcomes indicated the possible associations of this gene with fistulizing behavior and perianal fistulas [276], ileal involvement at diagnosis [277, 278], male sex, time to development of non-perianal fistulas [277], colonic location [259]. The possible association with the biological therapy was investigated only in the recent study by Meggyesi *et al.* (2010) [259]. However, they have investigated the efficacy of the biological therapy and they did not find any significant associations. Therefore, further studies are needed to confirm the reported novel association of the *IRGM* with the severe forms of the disease.

The past decade has witnessed remarkable success in the identification of low-penetrance, high-frequency susceptibility variants in common, complex diseases [81]. However, a large part of the genetic variance in many of these diseases is still unaccounted for. One of the main reasons is that complex human diseases result from the poorly understood systematic epistatic interactions of genetic variants [129, 130]. There is growing evidence that genetic interactions, whether synergistic or antagonistic, are not only possible but are also ubiquitous [21–23]. The inheritance of combinations of functional and disease-linked commonly occurring SNPs may additively or synergistically disturb the system-wide communication of the biological processes, leading to disease [21]. Therefore, the effect might be missed if the gene functioning primarily through a complex mechanism is examined in isolation without allowing for its potential interactions with other genes and, possibly, environmental factors [130].

Our study is one of the first studies investigating the possible SNP-SNP interactions in the association with the inflammatory bowel disease in hypothesis free way, *i.e.*, investigation of the impact of all possible SNP pairs, even those that initially were not associated with IBD. A number of previous studies have investigated the interactions between pathway-related genes [131–133] or genes that were individually associated with IBD [134–136]. Only a recent study analysed SNP-SNP interactions based on the WTCCC genome scale data [58]. The study has identified the association between the SNP pair (rs6496669 and rs434157) that is in LD with adenomatous polyposis coli (*APC*) and IQ motif containing GTPase activating protein 1 (*IQGAP1*) genes and CD [137]. However, study analysing of both forms of inflammatory bowel disease has not been performed, yet. Therefore, the novelty of our study is the demonstration of statistically significant interactions between SNPs (rs2476601 and rs3764147) that did not have an effect on UC risk individually. The interaction pattern between rs2476601 and rs3764147 indicates that carrying at least three minor alleles of SNPs increases the risk for UC by a factor 1.63. Interacting SNPs are in genes *PTPN22* (rs2476601) and *C13orf31* (rs3764147) and both are coding mutations (rs2476601 – synonymous; rs3764147 – missense). The two interacting regions have been related to the development of the autoimmune diseases (*PTPN22:* CD [67], type 1 diabetes, rheumatoid arthritis, systemic lupus erythematosus, Graves' disease, autoimmune thyroid disease, alopecia areata, juvenile idiopathic arthritis and Hashimoto's

thyroiditis [17, 279]; *C13orf31*: CD [67], Leprosy [280]). However, none of the markers independently were previously associated with UC.

PTPN22 encodes a lymphoid-specific protein tyrosine phosphatase (LYP), a member of a family of proteins involved in suppressing spontaneous T-cell activation [281, 282]. *PTPN22* is expressed in many hematopoietic cell types, notably T cells. The rs2476601 autoimmune risk allele is a gain-of-function mutation and it results in a phosphatase with higher catalytic activity and more potent negative regulation of T-cell activation [283, 284]. By contrast, knockout mice (Lyp is the mouse ortholog of *PTPN22*) have an increased T-cell activation in combination with an increased production of antibodies [285]. Biologically, the *C13orf31* functions are not known. However, as *C13orf31* polymorphisms were previously associated with CD and Leprosy it has been suggested that it might be involved in *Mycobacterium* clearance.

Moreover, the performed *in silico* prediction of the possible interactive *PTPN22* and *C13orf31* network indicated, that both of them could be connected through co-expression and protein-protein interaction (*i.e.*, physical interaction) pathways. The most significant molecular processes predicted by the program were regulation of positive T cell activation and segregation of the TCR complex, which mainly affect immune system processes regulation and cell (lymphocyte, T cell) activation. These processes are mainly linked to the query gene *PTPN22.* As noted above, there are no existing functional annotations of the *C13orf31* gene. However, *in silico* analysis refered that through co-expression with *FRK* (fyn related kinase) *C13orf31* could be involved in the above mentioned processes [286, 287]. FRK (originally called RAK) belongs to the family of the Src-related tyrosine kinases. This protein has been implicated in the regulation of epithelial cell differentiation and apoptosis. Originally FRK has been identified in the melanoma, breast cancer cells and normal intestinal epithelium [288]. Using chemical proteomics approach it has been revieled that FRK physically interacts with Csk (c-src tyrosine kinase) [289]. However, the exact interaction mechanism is not known, yet. Csk is known as one of the adaptor molecules of the protein tyrosine phosphatases (PTPs, including LYP), which (as mentioned above) are involved in the negative regulation of T-cell activation [290, 291]. Once the T cell receptor complex (*e.g.*, TCR/CD3) has been activated, adaptor molecules (*i.e.*, Csk, Cbl (Cas-Br-M (murine) ecotropic retroviral transforming sequence), Grb2 (growth factor receptor-bound protein 2)) in physical association with LYP have the challenging task of preventing T-cell activation and maintaining T cells in an inactive mode until co-stimulatory

receptors (*e.g.*, CD28, CTLA-4 (cytotoxic T-lymphocyte antigen 4)) are engaged [290–292]. TCR co-stimulation activates various intracellular signal transduction cascades (*e.g.*, PI3K (phosphatidylinositol 3-kinase) – Akt (serine/threonine protein kinase Akt) – NFκB (nuclear factor of kappa light polypeptide gene enhancer in B-cells)) and cytoskeletal remodeling (*e.g.*, VAV1 (vav 1 guanine nucleotide exchange factor)) resulting in T cell activation; whereas signal abrogation induces the state of functional unresponsiveness known as clonal T-cell anergy, *i.e.*, the expression of anergy associated genes is activated, including cell cycle inhibitors, tyrosine phosphatases, proteinases, transcriptional regulators, diacylglycerol kinases and E3 ligases [290–293]. This signal transduction mechanism is one of the numerious regulatory mechanisms, that ensures *in vivo* the delicate balance between T-cell activation, tolerance, and autoimmunity.

Our study confirmed the proposed epistasis model, *i.e.*, that SNPs without main effects or with main effects, too small to detect, may interact with others and confer an increased risk for disease. However, in the future larger studies will allow a better application of the interaction model, in which more complex interactions could be investigated.

From a clinician's point of view, it might be attractive to create a genetic risk profile that could be used as an accurate, composite, and predictive index in diagnosis and management of the IBD. Genetic risk profiling would be beneficial in prioritizing individuals with the suspicion of the disease and increasing the chance of early disease detection. Moreover, noninvasive genetic testing would be especially valuable in the differential diagnosis of IBD as it often remains difficult to differentiate between CD and UC with current diagnostic methods. A correct diagnosis is essential for the treatment of patients, as many drugs that are effective in one form of IBD have insufficient beneficial therapeutic effect in the other. However, despite the advances in the field of IBD genetics there are currently no genetic tests which are recommended routinely for diagnosis or management of the diseases [294]. This is why further studies are needed to assess collectively all potential genetic predictors in large, phenotypically well-defined cohorts, in order to build an accurate composite predictor index.

In accordance with previously proposed models [295, 296], we created two genetic risk profiles based on the number of alleles and on a weighted score for the strength of association of each SNP using the coefficients of regression analyses. For the models, we used the loci that were associated with CD and

UC after correction for multiple testing and nominally associated SNPs, that in previous studies were associated with CD and UC. Although each of the replicated loci had only a small individual effect on UC and CD risk, our risk models clearly show that individuals with more risk alleles have an increased risk for IBD. Similar increase in risk with each additional risk allele in CD and UC has been reported in previously published studies [174, 296]. Although each individual risk allele only conveys an OR of about 1.3 for UC, an individual with 19 or more risk alleles already has a strongly increased risk for developing UC, with an OR of 14.29. If we sum the weight of the individual effect of each risk locus, the risk for UC increases even more with an OR of 19.00 for the weighted score of 17 or more. The individual risk allele impact on CD development is varrying from 4.52 (rs2066847) to 1.35 (rs1736135). However, in the individuals carrying more than seven risk alleles the risk for developing CD is strongly increased, with an OR of 6.71. The evaluation of the individual SNP weighted score showed similar results as the first model, *i.e.*, the risk for CD increases even more with an OR of 7.86 for the weighted score of 10 or more.

Moreover, although we showed that it is possible to create a genetic risk profile that is clinically useful, however the combinations of SNPs used in the study are not sufficient for accurate and sensitive diagnosis. The highest sensitivities, which can help the clinician to rule out CD or UC were achieved with less than three risk alleles for CD (71.74%) and less than 12 risk alleles for UC (70.83%) or a weighted score of less than six for CD (60.98%) and more than 11 for UC (63.79%). Vice versa, high specificities (at the expense of low sensitivities) were achieved with more than seven risk alleles for CD (95.61%) and more than 19 risk alleles for UC (98.77%) or a weighted score of 12 for CD (97.41%) and more than 17 for UC (99.13%), which makes the presence of disease very likely. However, it should be noted that most patients will not have these very low or high scores, which makes its clinical usefulness more difficult. It has been possible to yield likelihood ratios of a positive test of up to 12.27 for UC and 5.36 for CD that gives a moderate increase of the probability of disease. Generally, a likelihood ratio of 10 or more is conceived as being conclusive for disease. Moreover, the minimal achieved likelihood ratio of a negative test was 0.56 for CD and 0.49 for UC, reflecting a moderate decrease in likelihood of disease.

Our findings strengthen the concept that a genetic risk profile can be constructed to aid the clinician in making decisions. However, predictive

testing is not yet feasible as the difference in absolute number and weight of risk alleles between IBD cases and control individuals is significant but small. This small difference is due to the fact that many disease associated variants are common, *i.e.*, highly prevalent in the general population. In the future, these models will need to be expanded including novel identified risk loci and should be combined with other diagnostic tests or risk factors (*e.g.*, smoking).

CONCLUSIONS

1. The association between Crohn's disease and single nucleotide polymorphisms in *NOD2* gene (rs2066847, rs2076756) was determined.
2. The association between ulcerative colitis and single nucleotide polymorphisms in 21q21.1 (rs1736135), 6q21 (rs7746082), *JAK2* (rs10758669), *RNF186* (rs3806308), *ORMDL3* (rs2872507) loci was determined.
3. The association between single nucleotide polymorphisms in *NOD2*, *IRGM* and *HLA* genes and Crohn's disease phenotype was determined. *NOD2* gene's single nucleotide polymorphism (rs2066847) was associated with Crohn's disease affecting the upper gastrointestinal tract. *IRGM* gene's single nucleotide polymorphisms (rs13361189, rs11747270) were associated with severe course of Crohn's disease (stricturing Crohn's disease behavior, the need for biological therapy and/or surgery treatment). The *HLA* locus polymorphisms (rs9268858, rs9268877, rs2395185) protected from the perianal Crohn's disease form.
4. The association between single nucleotide polymorphisms in *IL10*, *BTNL2*, *ORMDL3* and 1p36.13 region and ulcerative colitis phenotype was determined. *IL10* gene's single nucleotide polymorphism (rs3024505) was associated with an increased risk for left sided ulcerative colitis and pancolitis. *BTNL2* gene's single nucleotide polymorphism (rs9268480) protected from pancolitis. *ORMDL3* gene's single nucleotide polymorphism (rs2872507) was associated with joint manifestations. Single nucleotide polymorphism in the *OTUD3* gene region (rs6426833) was associated with severe forms of ulcerative colitis that required treatment and/or biological therapy.
5. The combination of the single nucleotide polymorphisms in the *PTPN22* (rs2476601) and *C13orf31* (rs3764147) genes increased the risk for ulcerative colitis.
6. A higher number of disease associated alleles increase the risk of developing inflammatory bowel disease. As the number of single nucleotide polymorphisms increase the high genetic test specificities are achieved at the expense of low sensitivities. The highest specificity in the ulcerative colitis group was achieved with the combination of 19 alleles or β-coefficient equal 17, the highest sensitivity – combination of

12 alleles or β-coefficient equal 11. The highest specificity in the Crohn‘s disease group was achieved with the combination of seven alleles or β-coefficient equal 12, the highest sensitivity – combination of three alleles or β-coefficient equal six. Therefore, the combinations of the used genetic markers are not sufficiently accurate for the routine clinical diagnostics of inflammatory bowel disease.

REFERENCES

1. Artis D. Epithelial–cell recognition of commensal bacteria and maintenance of immune homeostasis in the gut. Nat Rev Immunol 2008;8:411–20.
2. Hill DA, Artis D. Intestinal Bacteria and the Regulation of Immune Cell Homeostasis. Annu Rev Immunol 2010;28:623–67.
3. Budarf ML, Labbe C, David G, Rioux JD. GWA studies: rewriting the story of IBD. Trends Genet 2009;25(3):137–46.
4. Salzman NH, Bevins CL. Negative interactions with the microbiota: IBD. Adv Exp Med Biol 2008;635:67–78.
5. Baumgart DC, Sandborn WJ. Inflammatory bowel disease: clinical aspects and established and evolving therapies. Lancet 2007;369:1641–57.
6. Schirbel A, Reichert A, Roll S, Baumgart DC, Buning C, Wittig B, et al. Impact of pain on health–related quality of life in patients with inflammatory bowel disease. World J Gastroenterol. 2010;16(25):3168–77.
7. Greenley RN, Hommel KA, Nebel J, Raboin T, Li SH, Simpson P et al. A meta–analytic review of the psychosocial adjustment of youth with inflammatory bowel disease. J Pediatr Psychol 2010;35(8):857–69.
8. Kulaylat MN, Dayton MT. Ulcerative colitis and cancer. J Surg Oncol 2010;101(8):706–12.
9. Binder V. Epidemiology of IBD during the twentieth century: an integrated view. Best Pract Res Clin Gastroenterol 2004;18:463–79.
10. Loftus EV. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. Gastroenterology 2004;126:1504–17.
11. Lakatos L, Lakatos PL. Is the incidence and prevalence of inflammatory bowel diseases increasing in Eastern Europe? Postgrad Med J 2006;82:332–7.
12. Lakatos PL, Fischer S, Lakatos L. Is the epidemiology of inflammatory bowel disease changing in Eastern Europe? Scand J Gastroenterol 2006;41:870–2.
13. Cooney R, Jewell D. The Genetic Basis of Inflammatory Bowel Disease. Dig Dis 2009;27:428–42.
14. van Limbergen J, Wilson DC, Satsangi J. The genetics of Crohn's disease. Annu Rev Genomics Hum Genet 2009;10:89–116.
15. Baumgart DC, Carding SR. Inflammatory bowel disease: cause and immunobiology. Lancet 2007;369:1627–40.
16. Achkar JP, Duerr R. The expanding universe of inflammatory bowel disease genetics. Curr Opin Gastroenterol 2008;24:429–34.
17. Lees CW, Satsangi J. Genetics of inflammatory bowel disease: implications for disease pathogenesis and natural history. Expert Rev Gastroenterol Hepatol 2009;3(5):513–34.

18. Lees CW, Barrett JC, Parkes M, Satsangi J. New IBD genetics: common pathways with other diseases. Gut 2011.
19. Franke A, McGovern DP, Barrett JC, Wang K, Radford–Smith GL, Ahmad T, et al. Genome–wide meta–analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42:1118–25.
20. Sventoraityte J, Zvirbliene A, Franke A, Kwiatkowski R, Kiudelis G, Kupcinskas L, et al. NOD2, IL23R and ATG16L1 polymorphisms in Lithuanian patients with inflammatory bowel disease. World J Gastroenterol 2010;16(3):358–63.
21. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 2003;56:73–82.
22. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, et al. Global mapping of the yeast genetic interaction network. Science 2004;303(5659):808–13.
23. Segre D, DeLuna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. Nat Genet 2005;37:77–83.
24. Kumar V, Abbas AK, Fausto N, Aster JC, editors. Robbins and Cotran Pathologic Basis of Disease, Eigth Edition. Philadelphia: Elsevier Inc.; 2010. p. 807–13.
25. Heresbach D, Alexandre JL, Branger B, Bretagne JF, Cruchant E, Dabadie A, et al. Frequency and significance of granulomas in a cohort of incident cases of Crohn's disease. Gut 2005;54:215–22.
26. Louis E, Collard A, Oger AF, Degroote E, Aboul Nasr El Yafi FA, Belaiche J. Behaviour of Crohn's disease according to the Vienna classification: changing pattern over the course of the disease. Gut 2001;49:777–82.
27. Silverberg MS, Satsangi J, Ahmad T, Arnott ID, Bernstein CN, Brant SR, et al. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. Can J Gastroenterol 2005;19:5–36.
28. Danese S, Semeraro S, Papa A, Roberto I, Scaldaferri F, Fedeli G, et al. Extraintestinal manifestations in inflammatory bowel disease. World J Gastroenterol 2005;11:7227–36.
29. Katz JA. Management of inflammatory bowel disease in adults. J Dig Dis 2007;8(2):65–71.
30. Kirsner JB. The Historical Basis of the Idiopathic Inflammatory Bowel Diseases. Inflamm Bowel Dis 1995;1(1):2–26.
31. Rutgeerts P, Sandborn WJ, Feagan BG, Reinisch W, Olson A, Johanns J, et al. Infliximab for induction and maintenance therapy for ulcerative colitis. N Engl J Med 2005; 353:2462–76.

32. Cohen JL, Strong SA, Hyman NH, Buie WD, Dunn GD, Ko CY, et al. Practice parameters for the surgical treatment of ulcerative colitis. Dis Colon Rectum 2005;48:1997–2009.
33. Bernstein CN. Epidemiologic clues to inflammatory bowel disease. Curr Gastroenterol Rep 2010;12(6):495–501.
34. Shivananda S, Lennard–Jones J, Logan R, Fear N, Price A, Carpenter L, 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 1996; 39:690–7.
35. Kiudelis G, Zvirbliene A, Sventoraityte J, Kupcinskas L. Referral epidemiology of ulcerative colitis and Crohn's disease in Kaunas region, Lithuania: one year prospective study. J Crohn's Colitis Suppl 2008;2:70.
36. Salupere R. Inflammatory bowel disease in Estonia: a prospective epidemiologic study 1993–1998. World J Gastroenterol 2001;7(3):387–8.
37. Loftus EV, Jr, Sandborn WJ. Epidemiology of inflammatory bowel disease. Gastroenterol Clin North Am 2002;31:1–20.
38. van Limbergen J, Russell RK, Drummond HE, Aldhous MC, Round NK, Nimmo ER, et al. Definition of phenotypic characteristics of childhood–onset inflammatory bowel disease. Gastroenterology 2008;135(4):1114–22.
39. Hanauer SB. Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities. Inflamm Bowel Dis 2006;12 Suppl 1:S3–9.
40. Li X, Sundquist J, Hemminki K, Sundquist K. Risk of inflammatory bowel disease in first– and second–generation immigrants in Sweden: A nationwide follow–up study. Inflamm Bowel Dis 2010 Nov 4.
41. Jess T, Loftus EV, Harmsen WS, Zinsmeister AR, Tremaine WJ, Melton LJ, et al. Survival and cause specific mortality in patients with inflammatory bowel disease: a long term outcome study in Olmsted County, Minnesota, 1940–2004. Gut 2006b;55:1248–54.
42. Jess T, Loftus EV, Velayos FS, Harmsen WS, Zinsmeister AR, Smyrk TC, et al. Risk of intestinal cancer in inflammatory bowel disease: a population–based study from Olmsted county, Minnesota. Gastroenterology 2006a;130:1039–46.
43. Wolters FL, Russel MG, Sijbrandij J, Schouten LJ, Odes S, Riis L, et al. Crohn's disease: increased mortality 10 years after diagnosis in a Europe–wide population based cohort. Gut 2006;55:510–8.
44. Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. Nature. 2007;448(7152):427–34.
45. Maynard CL, Weaver CT. Intestinal effector T cells in health and disease. Immunity. 2009;31(3):389–400.
46. Barnes MJ, Powrie F. Regulatory T cells reinforce intestinal homeostasis. Immunity. 2009;31(3):401–11.

47. Hooper LV, Macpherson AJ. Immune adaptations that maintain homeostasis with the intestinal microbiota. Nat Rev Immunol. 2010;10(3):159–69.
48. Yamamoto–Furusho JK. Innovative therapeutics for inflammatory bowel disease. World J Gastroenterol 2007;13(13):1893–6.
49. Brand S. Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. Gut 2009;58(8):1152–67.
50. Zenewicz LA, Antov A, Flavell RA. CD4 T–cell differentiation and inflammatory bowel disease.Trends Mol Med 2009;15(5):199–207.
51. Boden EK, Snapper SB. Regulatory T cells in inflammatory bowel disease. Curr Opin Gastroenterol 2008;24:733–41.
52. Turner JR. Molecular basis of epithelial barrier regulation: from basic mechanisms to clinical application. Am J Pathol 2006;169:1901–9.
53. Kaser A, Zeissig S, Blumberg RS. Inflammatory Bowel Disease. Annu Rev Immunol 2010;28:573–621.
54. Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 2001;411(6837):603–6.
55. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, et al. Association of NOD2 leucine–rich repeat variants with susceptibility to Crohn's disease. Nature 2001;411(6837):599–603.
56. De Jager PL, Franchimont D, Waliszewska A, Bitton A, Cohen A, Langelier D, et al. The role of the Toll receptor pathway in susceptibility to inflammatory bowel diseases. Genes Immun 2007;8(5):387–97.
57. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, et al. A genome–wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 2007;39(2):207–11.
58. Wellcome Trust Case Control Consortium. Genome–wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661–78.
59. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome–wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 2007;39(5):596–604.
60. Elson CO, Cong Y, McCracken VJ, Dimmitt RA, Lorenz RG, Weaver CT. Experimental models of inflammatory bowel disease reveal innate, adaptive, and regulatory mechanisms of host dialogue with the microbiota. Immunol Rev 2005;206:260–76.
61. Levine B, Deretic V. Unveiling the roles of autophagy in innate and adaptive immunity. Nat Rev Immunol 2007;7(10):767–77.

62. Saitoh T, Fujita N, Jang MH, Uematsu S, Yang BG, Satoh T, et al. Loss of the autophagy protein Atg16L1 enhances endotoxin–induced IL–1β production. Nature 2008;456:264–8.
63. Kuballa P, Huett A, Rioux JD, Daly MJ, Xavier RJ. Impaired autophagy of an intracellular pathogen induced by a Crohn's disease associated ATG16L1 variant. PLoS One 2008;3(10):e3391.
64. Travassos LH, Carneiro LA, Ramjeet M, Hussey S, Kim YG, Magalhaes JG, et al. Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. Nat Immunol 2010;11(1):55–62.
65. Cooney R, Baker J, Brain O, Danis B, Pichulik T, Allan P, et al. NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. Nat Med 2010;16(1):90–7.
66. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, et al. Novel Crohn disease locus identified by genome–wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PloS Genet 2007;3(4):e58.
67. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, et al. Genome–wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008;40(8):955–62.
68. Fisher SA, Tremelling M, Anderson CA, Gwilliam R, Bumpstead S, Prescott NJ, et al. Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. Nat Genet 2008;40(6):710–2.
69. Peltekova VD, Wintle RF, Rubin LA, Amos CI, Huang Q, Gu X, et al. Functional variants of OCTN cation transporter genes are associated with Crohn disease. Nat Genet 2004;36(5):471–5.
70. Buhner S, Buning C, Genschel J, Kling K, Herrmann D, Dignass A, et al. Genetic basis for increased intestinal permeability in families with Crohn's disease: role of CARD15 3020insC mutation? Gut 2006;55(3):342–7.
71. Su L, Turner JR. Got guts? Need nerve! Gastroenterology 2007;132(4):1615–8.
72. Rutgeerts P, Goboes K, Peeters M, Hiele M, Penninckx F, Aerts R, et al. Effect of faecal stream diversion on recurrence of Crohn's disease in the neoterminal ileum. Lancet 1991;338:771–4.
73. Gionchetti P, Rizzello F, Lammers KM, Morselli C, Sollazzi L, Davies S, et al. Antibiotics and probiotics in treatment of inflammatory bowel disease. World J Gastroenterol 2006;12:3306–13.
74. Macfarlane GT, Blackett KL, Nakayama T, Steed H, Macfarlane S. The Gut Microbiota in Inflammatory Bowel Disease. Current Pharmaceutical Design 2009;15:1528–36.
75. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular phylogenetic characterization of microbial community imbalances in

human inflammatory bowel diseases. Proc Natl Acad Sci U S A 2007;104(34): 13780–5.
76. Hugot JP, Cezard JP, Colombel JF, Belaiche J, Almer S, Tysk C, et al. Clustering of Crohn's disease within affected sibships. Eur J Hum Genet 2003; 11:179–84.
77. Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet 2001;2(2):91–9.
78. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008;322(5903):881–8.
79. Kere J. Genetics of complex disorders. Biochem Biophys Res Commun 2010;396:143–6.
80. Borecki IB, Province MA. Linkage and association: basic concepts. Adv Genet 2008;60:51–74.
81. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461(7265): 747–53.
82. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet 2003;33 Suppl:228–37.
83. Teare MD, Barrett JH. Genetic linkage studies. Lancet 2005;366:1036–44.
84. Kruglyak L, Lander ES. Complete multipoint sib–pair analysis of qualitative and quantitative traits. Am J Hum Genet 1995;57:439–54.
85. Cordell HJ, Clayton DG. Genetic association studies. Lancet 2005;366:1121–31.
86. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273(5281):1516–7.
87. Risch NJ. Searching for genetic determinants in the new millennium. Nature 2000;405(6788):847–56.
88. Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genome wide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 2001; 69(5):936–50.
89. Perkel J. SNP genotyping: six technologies that keyed a revolution. Nat Methods 2008;5(5):447–54.
90. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. (International Human Genome Sequencing Consortium) Initial sequencing and analysis of the human genome. Nature 2001;409: 860–921.
91. International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–96.
92. Lander ES. The new genomics: global views of biology. Science 1996; 274(5287):536–9.

93. Collins FS, Guyer MS, Chakravarti A. Variations on a theme: cataloging human DNA sequence variation. Science 1997;278:1580–1.
94. SNP database dbSNP. Available at: http://www.ncbi.nlm.nih.gov/projects/SNP
95. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007; 449(7164):851–61.
96. Kruglyak L. The road to genome–wide association studies. Nat Rev Genet 2008;9(4):314–8.
97. International HapMap Consortium. A Haplotype Map of the Human Genome. Nature 2005;437(7063):1299–1320.
98. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, et al. Haplotype tagging for the identification of common disease genes. Nat Genet 2001;29(2):233–7.
99. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science 2002; 296(5576):2225–9.
100. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single–nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004;74(1):106–20.
101. Ke X, Miretti MM, Broxholme J, Hunt S, Beck S, Bentley DR, et al. A comparison of tagging methods and their tagging space. Hum Mol Genet 2005;14(18):2757–67.
102. Barrett JC, Cardon LC. Evaluating coverage of genome–wide association studies. Nat Genet 2006;38(6):659–62.
103. Manolio TA, Collins FS. The HapMap and genome–wide association studies in diagnosis and therapy. Annu Rev Med 2009;60:443–56.
104. Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Meth 2004;1:109–11.
105. Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. Nat Rev Genet 2006;7:632–44.
106. Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, et al. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome–wide SNP platforms. Am J Hum Genet 2008;83:112–9.
107. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome–wide association studies for complex traits: consensus, uncertainty and challenges.Nat Rev Genet 2008;9(5):356–69.
108. Cardon LR. Genetics. Delivering new disease genes. Science 2006;314(5804): 1403–5.

109. Sen S, Burmeister M. Hardy–Weinberg analysis of a large set of published association studies reveals genotyping error and a deficit of heterozygotes across multiple loci. Hum Genomics 2008;3(1):36–52.
110. Maker B. Personal genomes: the case of the missing heritability. Nature 2008; 456(7218):18–21.
111. Hirschhorn JN, Daly MJ. Genome–wide association studies for common diseases and complex traits. Nat Rev Genet 2005;6(2):95–108.
112. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome–wide association signals. Nat Rev Gen 2009;10(5):318–29.
113. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome–wide association studies in diverse populations. Nat Rev Genet 2010;11(5):356–66.
114. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome–wide association studies. Nat Rev Genet 2010;11(7): 459–63.
115. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol 2009;38(1):263–73.
116. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet 2004;36(5): 512–7.
117. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet 2003;361:598–604.
118. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet 2004;36:388–93.
119. Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case–control studies. Am J Hum Genet 2007;80:921–30.
120. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, et al. Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 2008;16(12):1413–29.
121. Lao O, Lu TT, Nothnagel M, Junge O, Freitag–Wolf S, Caliebe A, et al. Correlation between genetic and geographic structure in Europe. Curr Biol 2008;18(16):1241–8.
122. Drineas P, Lewis J, Paschou P. Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers. PLoS One 2010 ;5(8):e11892.

123. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, et al. Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 2008;4(1):e4.
124. Nelis M, Esko T, Magi R, Zimprich F, Zimprich A, Toncheva D, et al. Genetic structure of Europeans: a view from the North–East. PLoS One. 2009;4(5): e5472.
125. 1000 Genomes: a deep catalog of human genetic variation. Available at: http://www.1000genomes.org/page.php
126. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med 2010;363(2):166–76.
127. Hindorff LA, Junkins HA, Manolio TA. NHGRI Catalog of published genome–wide association studies. Available at: http://www.genome.gov/gwastudies
128. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insight into the genetics of common disease. J Clin Invest 2008;118:1590–1605.
129. Moore JH. A global view of epistasis. Nat Genet 2005;37(1)**:**13–4.
130. Cordell JH. Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet 2009;10(6):392–404.
131. McGovern DP, Rotter JI, Mei L, Haritunians T, Landers C, Derkowski C, et al. Genetic epistasis of IL23/IL17 pathway genes in Crohn's disease. Inflamm Bowel Dis 2009;15(6):883–9.
132. Seiderer J, Elben I, Diegelmann J, Glas J, Stallhofer J, Tillack C, et al. Role of the novel Th17 cytokine IL–17F in inflammatory bowel disease (IBD): upregulated colonic IL–17F expression in active Crohn's disease and analysis of the IL17F p.His161Arg polymorphism in IBD. Inflamm Bowel Dis 2008;14(4):437–45.
133. Schmechel S, Konrad A, Diegelmann J, Glas J, Wetzke M, Paschos E, et al. Linking genetic susceptibility to Crohn's disease with Th17 cell function: IL–22 serum levels are increased in Crohn's disease and correlate with disease activity and IL23R genotype status. Inflamm Bowel Dis 2008;14(2):204–12.
134. Torok HP, Glas J, Endres I, Tonenchi L, Teshome MY, Wetzke M, et al. Epistasis between Toll–like receptor–9 polymorphisms and variants in NOD2 and IL23R modulates susceptibility to Crohn's disease. Am J Gastroenterol 2009;104(7):1723–33.
135. Glas J, Seiderer J, Wetzke M, Konrad A, Torok HP, Schmechel S, et al. rs1004819 is the main disease–associated IL23R variant in German Crohn's disease patients: combined analysis of IL23R, CARD15, and OCTN1/2 variants. PLoS One 2007;2(9):e819.
136. Glas J, Konrad A, Schmechel S, Dambacher J, Seiderer J, Schroff F, et al. The ATG16L1 gene variants rs2241879 and rs2241880 (T300A) are strongly

associated with susceptibility to Crohn's disease in the German population. Am J Gastroenterol 2008;103(3):682–91.
137. Emily M, Mailund T, Hein J, Schauser L, Schierup MH. Using biological networks to search for interacting loci in genome–wide association studies. Eur J Hum Genet 2009;17(10):1231–40.
138. Orholm M, Iselius L, Sorensen TI, Munkholm P, Langholz E, Binder V. Investigation of inheritance of chronic inflammatory bowel diseases by complex segregation analysis. BMJ. 1993;306:20–24.
139. Orholm M, Munkholm P, Langholz E, Nielsen OH, Sorensen IA, Binder V. Familial occurrence of inflammatory bowel disease. New Engl J Med 1991;324(2):84–8.
140. 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 1988;29(7):990–6.
141. Halfvarson J, Bodin L, Tysk C, Lindberg E, Jarnerot G. Inflammatory bowel disease in a Swedish twin cohort: a long–term follow–up of concordance and clinical characteristics. Gastroenterology 2003;124(7):1767–73.
142. Orholm M, Binder V, Sorensen TI, Rasmussen LP, Kyvik KO. Concordance of inflammatory bowel disease among Danish twins. Results of a nationwide study. Scand J Gastroenterol 2000;35(10):1075–81.
143. Thompson NP, Driscoll R, Pounder RE, Wakefield AJ. Genetics versus environment in inflammatory bowel disease: results of a British twin study. BMJ 1996;312(7023):95–6.
144. Spehlmann ME, Begun AZ, Burghardt J, Lepage P, Raedler A, Schreiber S. Epidemiology of Inflammatory Bowel Disease in a German Twin Cohort: Results of a Nationwide Study. Inflamm Bowel Dis 2008;14(7):968–76.
145. Halme L, Paavola–Sakki P, Turunen U, Lappalainen M, Farkkila M, Kontula K. Family and twin studies in inflammatory bowel disease. World J Gastroenterol 2006;12(23):3668–72.
146. Halfvarson J, Jess T, Bodin L, Jarnerot G, Munkholm P, Binder V, et al. Longitudinal Concordance for Clinical Characteristics in a Swedish–Danish Twin Population with Inflammatory Bowel Disease. Inflamm Bowel Dis 2007;13(12):1536–44.
147. Duerr RH. The genetics of inflammatory bowel disease. Gastroenterol Clin North Am 2002;31:63–76.
148. van Heel DA, Fisher SA, Kirby A, Daly MJ, Rioux JD, Lewis CM. Inflammatory bowel disease susceptibility loci defined by genome scan meta–analysis of 1952 affected relative pairs. Hum Mol Genet 2004;13(7):763–70.

149. Bonen DK, Cho JH. The genetics of inflammatory bowel disease. Gastroenterology 2003;124:521–36.
150. Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. Am J Gastroenterol 2004;99:2393–404.
151. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome–wide association study identifies IL23R as an inflammatory bowel disease gene. Science 2006;314(5804):1461–3.
152. Franke A, Hampe J, Rosenstiel P, Becker C, Wagner F, Hasler R, et al. Systematic association mapping identified NELL1 as a novel IBD disease gene. PLoS ONE 2007;2:e691.
153. McGovern DP, Jones MR, Taylor KD, Marciante K, Yan X, Dubinsky M, et al. Fucosyltransferase 2 (FUT2) non–secretor status is associated with Crohn's disease. Hum Mol Genet 2010;19(17):3468–76.
154. Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 2001;29(2):223–8.
155. Giallourakis C, Stoll M, Miller K, Hampe J, Lander ES, Daly MJ, et al. IBD5 is a general risk factor for inflammatory bowel disease: replication of association with Crohn disease and identification of novel association with ulcerative colitis. Am J Hum Genet 2003;73(1):205–11.
156. Mirza MM, Fisher SA, King K, Cuthbert AP, Hampe J, Sanderson J, et al. Genetic evidence for interaction of the 5q31 cytokine locus and the CARD15 gene in Crohn disease. Am J Hum Genet 2003;72(4):1018–22.
157. Yamazaki K, McGovern D, Ragoussis J, Paolucci M, Butler H, Jewell D, et al. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum Mol Genet 2005;14(22):3499–506.
158. Raelson JV, Little RD, Ruether A, Fournier H, Paquin B, Van Eerdewegh P, et al. Genome–wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. Proc Natl Acad Sci USA 2007;104:14747–52.
159. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 2007; 39:830–2.
160. Franke A, Fischer A, Nothnagel M, Becker C, Grabe N, Till A, Lu T, et al. Genome–wide association analysis in sarcoidosis and Crohn's disease unravels a common susceptibility locus on 10p12.2. Gastroenterology 2008;135(4):1207–15.

161. Franke A, Balschun T, Karlsen TH, Sventoraityte J, Nikolaus S, Mayr G, et al. Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. Nat Genet 2008;40(11):1319–23.
162. Kugathasan S, Baldassano RN, Bradfield JP, Sleiman PM, Imielinski M, Guthery SL, et al. Loci on 20q13 and 21q22 are associated with pediatric–onset inflammatory bowel disease. Nat Genet 2008;40(10):1211–5.
163. Silverberg MS, Cho JH, Rioux JD, McGovern DP, Wu J, Annese V, et al. Ulcerative colitis–risk loci on chromosomes 1p36 and 12q15 found by genome–wide association study. Nat Genet 2009;41:216–20.
164. Asano K, Matsushita T, Umeno J, Hosono N, Takahashi A, Kawaguchi T, et al. A genome–wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. Nat Genet 2009;41(12):1325–9.
165. UK IBD Genetics Consortium, Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, et al. Genome–wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat Genet 2009;41(12):1330–4.
166. Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M, et al. Common variants at five new loci associated with early–onset inflammatory bowel disease. Nat Genet 2009;41(12):1335–40.
167. Franke A, Balschun T, Sina C, Ellinghaus D, Hasler R, Mayr G, et al. Genome–wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). Nat Genet 2010;42(4):292–4.
168. Wang K, Baldassano R, Zhang H, Qu HQ, Imielinski M, Kugathasan S, et al. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. Hum Mol Genet 2010;19(10): 2059–67.
169. McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, Taylor KD, et al. Genome–wide association identifies multiple ulcerative colitis susceptibility loci. Nat Genet 2010;42(4):332–7.
170. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, et al. Meta–analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 2011
171. Melum E, Franke A, Karlsen TH. Genome–wide association studies—a summary for the clinical gastroenterologist.World J Gastroenterol 2009;15(43): 5377–96.
172. Nakagome S, Takeyama Y, Mano S, Sakisaka S, Matsui T, Kawamura S, et al. Population–specific susceptibility to Crohn's disease and ulcerative colitis; dominant and recessive relative risks in the Japanese population. Ann Hum Genet 2010;74(2):126–36.

173. Torkvist L, Halfvarson J, Ong RT, Lordal M, Sjoqvist U, Bresso F, et al. Analysis of 39 Crohn's Disease Risk Loci in Swedish Inflammatory Bowel Disease Patients. Inflamm Bowel Dis 2010;16(6):907–9.
174. Festen EA, Stokkers PC, van Diemen CC, van Bodegraven AA, Boezen HM, Crusius BJ, et al. Genetic Analysis in a Dutch Study Sample identifies more ulcerative colitis susceptibility loci and shows their additive role in disease risk. Am J Gastroenterol 2010;105(2):395–402.
175. Cotterill L, Payne D, Levinson S, McLaughlin J, Wesley E, Feeney M, et al. Replication and meta–analysis of 13,000 cases defines the risk for interleukin–23 receptor and autophagy–related 16–like 1 variants in Crohn's disease. Can J Gastroenterol 2010;24(5):297–302.
176. Gazouli M, Pachoula I, Panayotou I, Mantzaris G, Chrousos G, Anagnou NP, et al. NOD2/CARD15, ATG16L1 and IL23R gene polymorphisms and childhood–onset of Crohn's disease.World J Gastroenterol.2010;16(14):1753–8.
177. Cheng JF, Ning YJ, Zhang W, Lu ZH, Lin L. T300A polymorphism of ATG16L1 and susceptibility to inflammatory bowel diseases: a meta–analysis. World J Gastroenterol 2010 14;16(10):1258–66.
178. Prescott NJ, Dominy KM, Kubo M, Lewis CM, Fisher SA, Redon R, et al. Independent and population–specific association of risk variants at the IRGM locus with Crohn's disease. Hum Mol Genet 2010;19(9):1828–39.
179. Tanaka M, Arimura Y, Goto A, Hosokawa M, Nagaishi K, Yamashita K, et al. Genetic variants in surfactant, pulmonary–associated protein D (SFTPD) and Japanese susceptibility to ulcerative colitis.Inflamm Bowel Dis 2009;15(6):918–25.
180. Franke A, Balschun T, Karlsen TH, Hedderich J, May S, Lu T, et al. Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. Nat Genet 2008;40(6):713–5.
181. Sfar I, Ben Aleya W, Mouelhi L, Aouadi H, Ben Rhomdhane T, Makhlouf M, et al. Lymphoid tyrosine phosphatase R620W variant and inflammatory bowel disease in Tunisia. World J Gastroenterol 2010;16(4):479–83.
182. Cenit MC, Alcina A, Marquez A, Mendoza JL, Diaz–Rubio M, de las Heras V, et al. STAT3 locus in inflammatory bowel disease and multiple sclerosis susceptibility. Genes Immun 2010;11(3):264–8.
183. Anderson CA, Massey DC, Barrett JC, Prescott NJ, Tremelling M, Fisher SA, et al. Investigation of Crohn's disease risk loci in ulcerative colitis further defines their molecular relationship. Gastroenterology 2009;136(2):523–9.e3.
184. Andersen V, Ernst A, Christensen J, Ostergaard M, Jacobsen BA, Tjonneland A, et al. The polymorphism rs3024505 proximal to IL–10 is associated with risk of ulcerative colitis and Crohns disease in a Danish case–control study. BMC Med Genet 2010;11:82.

185. Amre DK, Mack DR, Morgan K, Fujiwara M, Israel D, Deslandres C, et al. Investigation of Reported Associations Between the 20q13 and 21q22 Loci and Pediatric–Onset Crohn's Disease in Canadian Children. Am J Gastroenterol 2009;104(11):2824–8.
186. Podolsky DK. Inflammatory bowel disease (1). N Engl J Med 1991;325:928–37.
187. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray–Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 2002;99(8):5261–6.
188. Langmore JP. Rubicon Genomics, Inc. Pharmacogenomics 2002;3:557–60.
189. Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, et al. Unbiased whole–genome amplification directly from clinical samples. Genome Res 2003;13(5): 954–64.
190. Lovmar L, Syvanen AC. Multiple displacement amplification to create a long–lasting source of DNA for genetic studies. Hum Mutat 2006;27(7):603–14.
191. Lasken RS, Egholm M. Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. Trends Biotechnol 2003;21(12):531–5.
192. De la Vega FM, Lazaruk KD, Rhodes MD, Wenz MH. Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. Mutat Res 2005;573(1–2):111–35.
193. Holland PM, Abramson RD, Watson R, Gelfand DH. 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 1991; 88:7276–80.
194. Livak KJ, Flood SJ, 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 1995;4:357–62.
195. Livak KJ. Allelic discrimination using fluorogenic probes and the 5' nuclease assay. Genet Anal 1999;14(5–6):143–9.
196. McGuigan FE, Ralston SH. Single nucleotide polymorphism detection: allelic discrimination using TaqMan. Psychiatr Genet 2002;12(3):133–6.
197. Afonina I, Zivarts M, Kutyavin I, Lukhtanov E, Gamper H, Meyer RB. Efficient priming of PCR with short oligonucleotides conjugated to a minor groove binder, Nucl Acids Res 1997;25:2657–60.
198. Dupont WD, Plummer WD. PS power and sample size program available for free on the Internet. Controlled Clin Trials 1997;18(3):274.
199. Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P. How to track and assess genotyping errors in population genetics studies. Mol Ecol 2004;13:3261–73.

200. Goring HH, Terwilliger JD. Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. Am J Hum Genet 2000; 66:1298–309.
201. Abecasis GR, Cherny SS, Cardon LR. The impact of genotyping error on family–based analysis of quantitative traits. Eur J Hum Genet 2001;9:130–4.
202. Terwilliger JD, Weeks DE, 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 1990;Suppl 47:A201.
203. Gordon D, Matise TC, Heath SC, Ott J. Power loss for multiallelic transmission/disequilibrium test when errors introduced: GAW11 simulated data. Genet Epidemiol 1999;17 Suppl 1:S587–92.
204. Pompanon F, Bonin A, Bellemain E, Taberlet P. Genotyping errors: causes, consequences and solutions. Nat Rev Genet 2005;6:847–59.
205. Taberlet P, Griffin S, Goossens B, Questiau S, Manceau V, Escaravage N, et al. Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Res 1996;24(16):3189–94.
206. Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, et al. Detection of genotyping errors by Hardy–Weinberg equilibrium testing. Eur J Hum Genet 2004;12(5):395–9.
207. Xu J, Turner A, Little J, Bleecker ER, Meyers DA. Positive results in association studies are associated with departure from Hardy–Weinberg equilibrium: hint for genotyping error? Hum Genet 2002;111(6):573–4.
208. Balding DJ, Bishop M, Canings C, editors. Handbook of statistical genetics, 3rd edition. West Sussex: John Wiley & Sons, Ltd; 2007.
209. Sasieni PD. From genotypes to genes: doubling the sample size. Biometrics 1997;53:1253–61.
210. Schaid DJ, Jacobsen SJ. Biased tests of association: comparisons of allele frequencies when departing from Hardy–Weinberg proportions. Am J Epidemiol 1999;149:706–11.
211. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin–dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52, 506–16.
212. Purcell S, Neale B, Todd–Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole–genome association and population–based linkage analyses. Am J Hum Genet 2007;81:559–75.
213. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication–based analysis for two–stage genome–wide association studies. Nat Genet 2006;38:209–31.
214. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310(6973):170.

215. De Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. Nat Genet 2005;37(11):1217–23.
216. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21(2):263–5.
217. Sachs L. Angewandte Statistik, 11th edition. Berlin: Springer Verlag; 2003.
218. Warde–Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 2010;38:W214–20.
219. Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 2010;6(2):e1000864.
220. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 2009;5: e1000337.
221. Lu Q, Elston RC. Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet 2008;82(3):641–51.
222. van der Net JB, Janssens AC, Defesche JC, Kastelein JJ, Sijbrands EJ, Steyerberg EW. Usefulness of Genetic Polymorphisms and Conventional Risk Factors to Predict Coronary Heart Disease in Patients With Familial Hypercholesterolemia. Am J Cardiol 2009; 103(3): 375–80.
223. Uitenbroek DG. Binomial – SISA. 1997. Available at: http://www.quantitativeskills.com/sisa/distributions/binomial.html
224. Gasche C, Grundtner P. Genotypes and phenotypes in Crohn's disease: do they help in clinical management? Gut 2005;54(1):162–7.
225. Evans DM, Marchini J, Morris AP, Cardon LR. Two–stage two–locus models in genome–wide association. PLoS Genet 2006;2(9):e157.
226. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304–51.
227. Medici V, Mascheretti S, Croucher PJ, Stoll M, Hampe J, Grebe J, et al. Extreme heterogeneity in CARD15 and DLG5 Crohn disease–associated polymorphisms between German and Norwegian populations. Eur J Hum Genet 2006; 14(4): 459–68.
228. Vind I, Vieira A, Hougs L, Tavares L, Riis L, Andersen PS, et al. NOD2/CARD15 gene polymorphisms in Crohn's disease: a genotype–phenotype analysis in Danish and Portuguese patients and controls. Digestion 2005; 72(2–3):156–63.
229. Rosenstiel P, Schreiber S. NOD–like receptors—pivotal guardians of the immunological integrity of barrier organs. Adv Exp Med Biol. 2009;653:35–47.

230. Carneiro LA, Magalhaes JG, Tattoli I, Philpott DJ, Travassos LH. Nod–like proteins in inflammation and disease. J Pathol 2008;214(2):136–48.
231. Hampe J, Cuthbert A, Croucher PJ, Mirza MM, Mascheretti S, Fisher S, et al. Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. Lancet 2001; 357: 1925–8.
232. 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 2002; 47: 469–72.
233. Croucher PJ, Mascheretti S, Hampe J, Huse K, Frenzel H, Stoll M, et al. Haplotype structure and association to Crohn's disease of CARD15 mutations in two ethnically divergent populations. Eur J Hum Genet 2003; 11: 6–16.
234. Ardley HC. Ring finger ubiquitin protein ligases and their implication to the pathogenesis of human diseases. Curr Pharm Des 2009;15(31):3697–715.
235. Ferguson LR, Han DY, Fraser AG, Huebner C, Lam WJ, Morgan AR, et al. Genetic factors in chronic inflammation: single nucleotide polymorphisms in the STAT–JAK pathway, susceptibility to DNA damage and Crohn's disease in a New Zealand population. Mutat Res 2010;690(1–2):108–15.
236. Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC; International Schizophrenia Consortium, Purcell SM, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet 2009;5(6):e1000534.
237. Stepkowski SM, Chen W, Ross JA, Nagy ZS, Kirken RA. STAT3: an important regulator of multiple cytokine functions. Transplantation 2008;85(10):1372–7.
238. Abraham C, Cho JH. IL–23 and autoimmunity: new insights into the pathogenesis of inflammatory bowel disease. Annu Rev Med 2009;60:97–110.
239. Lankford CS, Frucht DM. A unique role for IL–23 in promoting cellular immunity. J Leukoc Biol 2003;73:49–56.
240. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 2007; 448:470–3.
241. Verlaan DJ, Berlivet S, Hunninghake GM, Madore AM, Lariviere M, Moussette S, et al. Allele–specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. Am J Hum Genet 2009;85(3):377–93.
242. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome–wide association study meta–analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 2010;42(6):508–14.
243. Hirschfield GM, Liu X, Xu C, Lu Y, Xie G, Lu Y, et al. Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants. N Engl J Med 2009; 360(24):2544–55.

244. Laukens D, Georges M, Libioulle C, Sandor C, Mni M, Vander Cruyssen B, et al. Evidence for Significant Overlap between Common Risk Variants for Crohn's Disease and Ankylosing Spondylitis. PLoS One 2010;5(11):e13795.
245. Hjelmqvist L, Tuson M, Marfany G, Herrero E, Balcells S, Gonzalez–Duarte R. ORMDL proteins are a conserved new family of endoplasmic reticulum membrane proteins. Genome Biol. 2002;3(6):RESEARCH0027.
246. Todd DJ, Lee AH, Glimcher LH. The endoplasmic reticulum stress response in immunity and autoimmunity. Nat Rev Immunol 2008;8:663–74.
247. Kitamura M. Biphasic, bidirectional regulation of NF–κB by endoplasmic reticulum stress. Antioxid. Redox Signal 2009;11:2353–64.
248. Kallies A, Hawkins ED, Belz GT, Metcalf D, Hommel M, Corcoran LM, et al. Transcriptional repressor Blimp–1 is essential for T cell homeostasis and self–tolerance. Nat Immunol 2006;7:466–74.
249. Martins GA, Cimmino L, Shapiro–Shelef M, Szabolcs M, Herron A, Magnus-dottir E, et al. Transcriptional repressor Blimp–1 regulates T cell homeostasis and function. Nat Immunol 2006;7:457–65.
250. Ohinata Y, Payer B, O'Carroll D, Ancelin K, Ono Y, Sano M, et al. Blimp1 is a critical determinant of the germ cell lineage in mice. Nature 2005;436:207–13.
251. Turner CA Jr, Mack DH, Davis MM. Blimp–1, a novel zinc finger–containing protein that can drive the maturation of B lymphocytes into immunoglobulin–secreting cells. Cell 1994;77(2):297–306.
252. Martins GA, Cimmino L, Liao J, Magnusdottir E, Calame K. Blimp–1 directly represses Il2 and the Il2 activator Fos, attenuating T cell proliferation and survival. J Exp Med 2008;205(9):1959–65.
253. Kallies A, Xin A, Belz GT, Nutt SL. Blimp–1 transcription factor is required for the differentiation of effector CD8(+) T cells and memory responses. Immunity 2009;31(2):283–95.
254. Zhang Z, Gerstein M. The human genome has 49 cytochrome c pseudogenes, including a relic of a primordial gene that still functions in mouse. Gene 2003;312:61–72.
255. Fritah A, Christian M, Parker MG. The metabolic coregulator RIP140: an update. Am J Physiol Endocrinol Metab 2010;299(3):E335–40.
256. Zschiedrich I, Hardeland U, Krones–Herzig A, Berriel Diaz M, Vegiopoulos A, Müggenburg J, et al. Coactivator function of RIP140 for NFkappaB/RelA–dependent cytokine gene expression. Blood 2008;112(2):264–76.
257. Runge MS, Patterson C, editors. Principles of Molecular Mediciene. 2nd ed. Totowa: Humana Press Inc.; 2006. p. 19–21.
258. Kinouchi Y, Negoro K, Takagi S, Takahashi S, Shimosegawa T. Genotype and phenotype relation in inflammatory bowel disease. J Gastroenterol 2005;40 Suppl 16:21–4.

259. Meggyesi N, Kiss LS, Koszarska M, Bortlik M, Duricova D, Lakatos L, et al. NKX2–3 and IRGM variants are associated with disease susceptibility to IBD in Eastern European patients. World J Gastroenterol 2010;16(41):5233–40.
260. Ahmad T, Marshall SE, Jewell D. Genetics of inflammatory bowel disease: the role of the HLA complex. World J Gastroenterol 2006;12(23):3628–35.
261. Yun J, Xu CT, Pan BR. Epidemiology and gene markers of ulcerative colitis in the Chinese. World J Gastroenterol 2009;15(7):788–803.
262. Galandiuk S, Kimberling J, Al–Mishlab TG, Stromberg AJ. Perianal Crohn disease: predictors of need for permanent diversion. Ann Surg 2005;241(5):796–801.
263. Annese V, Piepoli A, Latiano A, Lombardi G, Napolitano G, Caruso N, et al. HLA–DRB1 alleles may influence disease phenotype in patients with inflammatory bowel disease: a critical reappraisal with review of the literature. Dis Colon Rectum 2005;48(1):57–64.
264. Gleeson MH, Walker JS, Wentzel J, Chapman JA, Harris R. Human leucocyte antigens in Crohn's disease and ulcerative colitis. Gut 1972; 13: 438–40.
265. Yap LM, Ahmad T, Jewell DP. The contribution of HLA genes to IBD susceptibility and phenotype. Best Pract Res Clin Gastroenterol 2004;18(3):577–96.
266. Kuhn R, Lohler J, Rennick D, Rajewsky K, Muller W. Interleukin–10–deficient mice develop chronic enterocolitis. Cell 1993,75:263–74.
267. Ishizuka K, Sugimura K, Homma T, Matsuzawa J, Mochizuki T, Kobayashi M, et al. Influence of interleukin–10 on the interleukin–1 receptor antagonist/ interleukin–1 beta ratio in the colonic mucosa of ulcerative colitis. Digestion 2001,63(Suppl 1):22–7.
268. Kim SC, Tonkonogy SL, Karrasch T, Jobin C, Sartor RB. Dual–association of gnotobiotic IL–10–/– mice with 2 nonpathogenic commensal bacteria induces aggressive pancolitis. Inflamm Bowel Dis 2007;13(12):1457–66.
269. Correa I, Veny M, Esteller M, Pique JM, Yague J, Panes J, et al. Defective IL–10 production in severe phenotypes of Crohn's disease. J Leukoc Biol 2009, 85:896–903.
270. Schreiber S, Heinig T, Thiele HG, Raedler A. Immunoregulatory role of interleukin 10 in patients with inflammatory bowel disease. Gastroenterology 1995,108:1434–44.
271. Makarova KS, Aravind L, Koonin EV. A novel superfamily of predicted cysteine proteases from eukaryotes, viruses and Chlamydia pneumoniae. Trends Biochem Sci 2000;25:50–2.
272. Murakami M, Yoshihara K, Shimbara S, Lambeau G, Singer A, Gelb MH, et al. Arachidonate release and eicosanoid generation by group IIE phospholipase A(2). Biochem Biophys Res Commun 2002;292(3):689–96.

273. Suzuki N, Ishizaki J, Yokota Y, Higashino K, Ono T, Ikeda M, et al. Structures, enzymatic properties, and expression of novel human and mouse secretory phospholipase A(2)s. J Biol Chem 2000;275(8):5785–93.
274. Virgin HW, Levine B. Autophagy genes in immunity. Nat Immunol 2009; 10(5):461–70.
275. Crotzer VL, Blum JS. Autophagy and adaptive immunity. Immunology 2010;131(1):9–17.
276. Latiano A, Palmieri O, Cucchiara S, Castro M, D'Inca R, Guariso G, et al. Polymorphism of the IRGM gene might predispose to fistulizing behavior in Crohn's disease. Am J Gastroenterol 2009;104(1):110–6.
277. Henckaerts L, Van Steen K, Verstreken I, Cleynen I, Franke A, Schreiber S, et al. Genetic risk profiling and prediction of disease course in Crohn's disease patients. Clin Gastroenterol Hepatol 2009;7(9):972–80.
278. Roberts RL, Hollis–Moffatt JE, Gearry RB, Kennedy MA, Barclay ML, Merriman TR. Confirmation of association of IRGM and NCF4 with ileal Crohn's disease in a population–based cohort. Genes Immun 2008;9(6):561–5.
279. Lee YH, Rho YH, Choi SJ, Ji JD, Song GG, Nath SK, et al. The PTPN22 C1858T functional polymorphism and autoimmune diseases – a meta–analysis. Rheumatology (Oxford) 2007;46(1):49–56.
280. Zhang FR, Huang W, Chen SM, Sun LD, Liu H, Li Y, et al. Genomewide association study of leprosy. N Engl J Med 2009;361(27):2609–18.
281. Cloutier JF, Veillette A. Association of inhibitory tyrosine protein kinase p50csk with protein tyrosine phosphatase PEP in T cells and other hemopoietic cells. EMBO J 1996;15: 4909–18.
282. Stanford SM, Mustelin TM, Bottini N. Lymphoid tyrosine phosphatase and autoimmunity: human genetics rediscovers tyrosine phosphatases. Semin Immunopathol 2010;32(2):127–36.
283. Rieck M, Arechiga A, Onengut–Gumuscu S, Greenbaum C, Concannon P, Buckner JH. Genetic variation in PTPN22 corresponds to altered function of T and B lymphocytes. J Immunol 2007;179(7):4704–10.
284. Vang T, Congia M, Macis MD, Musumeci L, Orru V, Zavattari P, et al. Autoimmune–associated lymphoid tyrosine phosphatase is a gain–of–function variant. Nat Genet 2005;37(12):1317–9.
285. Hasegawa K, Martin F, Huang G, Tumas D, Diehl L, Chan AC. PEST domain–enriched tyrosine phosphatase (PEP) regulation of effector/memory T cells. Science 2004;303(5658):685–9.
286. Wong HR, Shanley TP, Sakthivel B, Cvijanovich N, Lin R, Allen GL, et al. Genome–level expression profiles in pediatric septic shock indicate a role for altered zinc homeostasis in poor outcome. Physiol Genomics 2007;30(2):146–55.

287. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001;98(26):15149–54.
288. Brauer PM, Tyner AL. RAKing in AKT: a tumor suppressor function for the intracellular tyrosine kinase FRK. Cell Cycle 2009;8(17):2728–32.
289. Bantscheff M, Eberhard D, Abraham Y, Bastuck S, Boesche M, Hobson S, et al. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nat Biotechnol 2007;25(9):1035–44.
290. Mustelin T, Tasken K. Positive and negative regulation of T–cell activation through kinases and phosphatases. Biochem J 2003;371(Pt 1):15–27.
291. Weetman AP, editor. Autoimmune Diseases in Endocrinology. Totowa, NJ: Humana Press, 2008, p. 24–27.
292. Bjorgo E, Tasken K. Novel mechanism of signaling by CD28. Immunol Lett 2010;129(1):1–6.
293. Paolino M, Penninger JM. E3 ubiquitin ligases in T–cell tolerance. Eur J Immunol 2009;39(9):2337–44.
294. van Assche G, Dignass A, Panes J, Beaugerie L, Karagiannis J, Allez M, et al. The second European evidence–based Consensus on the diagnosis and management of Crohn's disease: Definitions and diagnosis. J Crohns Colitis 2010;4(1):7–27.
295. McGovern DP, Butler H, Ahmad T, Paolucci M, van Heel DA, Negoro K, et al. TUCAN (CARD8) genetic variants and inflammatory bowel disease. Gastro-enterology 2006;131(4):1190–6.
296. Weersma RK, Stokkers PC, Cleynen I, Wolfkamp SC, Henckaerts L, Schreiber S, et al. Confirmation of multiple Crohn's disease susceptibility loci in a large Dutch–Belgian cohort. Am J Gastroenterol 2009;104(3):630–8.
297. Stoll M, Corneliussen B, Costello CM, Waetzig GH, Mellgard B, Koch WA, et al. Genetic variation in DLG5 is associated with inflammatory bowel disease. Nat Genet 2004;36(5):476–80.
298. Gunnlaugsdottir B, Skaftadottir I, Ludviksson BR. Naive human T–cells become non–responsive towards anti–TNFalpha (infliximab) treatment in vitro if co–stimulated through CD28. Scand J Immunol 2008;68(6):624–34.

LIST OF PUBLICATIONS

1. **Šventoraitytė J**, Žvirblienė A, Franke A, Kwiatkowski R, Kiudelis G, Kupčinskas L, Schreiber S. *NOD2, IL23R* and *ATG16L1* polymorphisms in Lithuanian patients with inflammatory bowel disease. World J Gastroenterol. 2010;16(3):359-64. PMID: 20082483

2. Franke A, Balschun T, Karlsen TH, **Šventoraitytė J**, Nikolaus S, Mayr G, et al. Sequence variants in *IL10, ARPC2* and multiple other loci contribute to ulcerative colitis susceptibility. Nat Genet. 2008;40(11):1319-23. PMID: 18836448

3. Franke A, Balschun T, Sina C, Ellinghaus D, Häsler R, Mayr G, Albrecht M, Wittig M, Buchert E, Nikolaus S, Gieger C, Wichmann HE, **Šventoraitytė J**, et al. Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (*IL17REL*). Nat Genet. 2010; 42(4):292-4. PMID: 20228798

4. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, Lee JC, Goyette P, Imielinski M, Latiano A, Lagacé C, Scott R, Amininejad L, Bumpstead S, Baidoo L, Baldassano RN, Barclay M, Bayless TM, Brand S, Büning C, Colombel JF, Denson LA, De Vos M, Dubinsky M, Edwards C, Ellinghaus D, Fehrmann RS, Floyd JA, Florin T, Franchimont D, Franke L, Georges M, Glas J, Glazer NL, Guthery SL, Haritunians T, Hayward NK, Hugot JP, Jobin G, Laukens D, Lawrance I, Lémann M, Levine A, Libioulle C, Louis E, McGovern DP, Milla M, Montgomery GW, Morley KI, Mowat C, Ng A, Newman W, Ophoff RA, Papi L, Palmieri O, Peyrin-Biroulet L, Panés J, Phillips A, Prescott NJ, Proctor DD, Roberts R, Russell R, Rutgeerts P, Sanderson J, Sans M, Schumm P, Seibold F, Sharma Y, Simms LA, Seielstad M, Steinhart AH, Targan SR, van den Berg LH, Vatn M, Verspaget H, Walters T, Wijmenga C, Wilson DC, Westra HJ, Xavier RJ, Zhao ZZ, Ponsioen CY, Andersen V, Torkvist L, Gazouli M, Anagnou NP, Karlsen TH, Kupcinskas L, **Šventoraitytė J**, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 2011; 43(3): 246–52. PMID:21297633

Other publications:

1. **Šventoraitytė J**, Balschun T, Kiudelis G, Žvirblienė A, Leja M, Kupčinskas L, Schreiber S, Franke A. Genetic analysis of multiple inflammatory bowel disease susceptibility markers in the Lithuanian and Latvian patient populations (oral presentation). "European Bridging Meeting 2010 and EAGE Postgraduate Course", Berlin (Germany), 18–20 November 2010.
2. **Šventoraitytė J,** Balschun T, Kiudelis G, Žvirblienė A, Leja M, Kupčinskas L, Schreiber S, Franke A. Comprehensive genetic analysis of inflammatory bowel disease susceptibility markers in the Lithuanian and Latvian patient populations (oral presentation). IIIrd National conference: "Science in human health", Kaunas, 7 April 2010.
3. **Šventoraitytė J**, Franke A, Žvirblienė A, Kiudelis G, Kupčinskas L, Schreiber S. Replication of Signals from Recent inflammatory Bowel Disease Genome Wide Scans in a Lithuanian cohort (poster presentation). "ESHG: European Human Genetics Conference 2009", Vienna (Austria), 23–26 May 2009. European Journal of Human Genetics 2009, vol. 17, suppl. 2, May, p. 244-245, no. P09.056.
4. **Šventoraitytė J**, Žvirblienė A, Kupčinskas L, Kiudelis G, Schreiber, S, Franke A. Replication of markers from recent inflammatory bowel disease genome-wide association scans in a Lithuanian cohort (poster presentation). "59th Annual ASHG Meeting", Honolulu (Hawaii, USA), 20–24 October 2009.
5. **Šventoraitytė J**, Franke A, Žvirblienė A, Kiudelis G, Kupčinskas L, Schreiber S. Association analysis of markers from recent genome wide association scans (GWAS) with inflammatory bowel disease in Lithuanian population (oral presentation). "European Bridging Meeting 2008 and EAGE Postgraduate Course", Cluj Napoca (Romania), 30 October – 1 November 2008.
6. **Šventoraitytė J**, Žvirblienė A, Kwiatkowski R, Kiudelis G, Kupčinskas L, Schreiber S. CARD15, TNF-alpha IL23R and ATG16L1 genes polymorphisms in Lithuanian patients with inflammatory bowel disease (poster presentation). "ESHG: European Human Genetics Conference 2008", Barcelona (Spain), 31 May – 3 June 2008. European Journal of Human Genetics 2008, vol. 16, suppl. 2, p. 320, no. P09.139.

ACKNOWLEDGEMENT

Thank you so much for all that every single one of you did. This study would not have been possible without the contribution of many competent and kind persons. Therefore, I would like to acknowledge:

My academic supervisor Prof. N. Pauziene for accepting me as a doctoral student. Her support and supervision throughout the years.

My academic co-supervisor Prof. L. Kupcinskas for providing me an attractive research project, large resources of patient material, continuous scientific support, encouragement and faith in my skills.

My academic co-supervisor Prof. A. Franke for his supervision, introduction into the genetics of complex diseases. His scientific supervision, support, encouragement provided the basis of this thesis.

Prof. S. Schreiber and Prof. A. Franke for the possibility to use excellent working conditions provided by the state-of-art infrastructutre of the Institute for Clinical Molecular Biology and the financial study support.

All clinicians and nurses, especially the team of the Kaunas University of Medicine Hospital, Department of Gastroenterology (Prof. G. Kiudelis, Prof. L. Jonaitis, Dr. A. Zvirbliene), Vilnius University Hospital Santariskiu Clinics (Dr. G. Denapiene, Prof. J. Valantinas), M. Marcinkeviciaus Hospital (Prof. A. Irnius), Siauliai Hospital (V. Svalbonas), Panevezys Hospital (Dr. D. Krukas), Klaipeda Seamen hospital (G. Simulionis), Klaipeda University Hospital (I. Vilcinskaite), National Blood center (Dr. V. Kalibata), Blood donor center (J. Venslauskas) ”Linezers” Hospital (M. Leja) and P. Stradin Clinical University Hospital and Riga Seamen Medical Center (A. Derovs and Derova), for their expert help.

Patients and their families, who agreed on the use of blood samples for academic research, for their cooperation.

I would like to thank the members of the Genotyping Platform (Institute for Clinical Molecular Biology, Kiel, Germany) and Laboratory of Gastroenteroloy and Hepatology (Lithuanian University of Health Sciences, Kaunas, Lithuania) for creating a friendly and enjoyable research environment.

I also thank the technical staff at the Institute for Clinical Molecular Biology (Kiel, Germany) for their expert help, pleasant and warm atmosphere.

APPENDIX

Patient's questionnaire

Blood collection date:

☐☐☐☐ *Year* ☐☐ *Month* ☐☐ *Day*

Inflammatory bowel disease genetics in the Baltic countries
Patient observation form

PERSONAL INFORMATION

Patients number: ☐☐☐☐☐ (first two numbers are the numbers of the hospital)

Hospital:

- ☐ **1** Kaunas Medical University Hospital, Kaunas, Lithuania
- ☐ **2** Vilnius University Hospital Santariskiu Klinikos, Vilnius, Lithuania
- ☐ **3** M.Marcinkeviciaus Hospital, Vilnius, Lithuania
- ☐ **4** Klaipeda Seamen hospital, Klaipeda, Lithuania
- ☐ **5** Klaipeda University Hospital, Klaipeda, Lithuania
- ☐ **6** Panevezys Hospital, Panevezys, Lithuania
- ☐ **7** Siauliai Hospital, Siauliai, Lithuania
- ☐ **8** East Tallinn Central Hospital, Tallinn, Estonia
- ☐ **9** Tartu University Hospital, Tartu, Estonia
- ☐ **10** P. Stradin Clinical University Hospital, Riga, Latvia
- ☐ **11** Riga Seamen Medical Center, Riga, Latvia
- ☐ **12** "Linezers" clinics, Riga, Latvia

Date of birth: ☐☐☐☐ *Year* ☐☐ *Month* ☐☐ *Day* **Age (years):** ☐☐

Gender: ☐ Male ☐ Female

Ethnicity: ..... **Parent's ethnicity** Mother ..... Father .....

Region: .....

Environment: ☐ Rural ☐ Urban

Weight (Kg): ☐☐☐ **Height (cm):** ☐☐☐

CLINICAL CHARACTERISTIC

Diagnosis ☐ Chron's disease ☐ Ulcerative colitis

Date at onset of disease ☐☐☐☐ *Year* ☐☐ *Month* ☐☐ *Day*

Activity of disease ☐ Relapse ☐ Remission

Case ☐ New ☐ Established

If disease is established indicate the number of flares per year:

- ☐ 1 time a year
- ☐ >1 time a year
- ☐ 1 time in 2-3 years
- ☐ 1 time in 4-5 years
- ☐ 1 time in 5-10 years

Disease extension in UC

- o Proctitis
- o Proctosigmoiditis
- o Left side colitis
- o Pancolitis

Disease location in CD

- o Terminal ileum, L1
- o Colon, L2
- o Ileocolon, L3
- o Upper GI, L4

Disease location in CD: ***Upper GI modifier, L4***

- o Terminal ileum + Upper GI, L1+L4
- o Colon + Upper GI, L2+L4
- o Ileocolon + Upper GI, L3+L4

Disease behavior in CD

- o Nonstricturing, nonpenetrating, B1
- o Stricturing, B2
- o Penetrating, B3

Disease behavior in CD: ***Perianal disease modifier (p)***

- o Nonstricturing, nonpenetrating + Perianal, B1p
- o Stricturing + Perianal, B2p
- o Penetrating + Perianal, B3p

Extraintestinal manifestations ☐ YES ☐ NO

If yes, indicate: o Joints o Ocular o Cutaneous o Hepatobiliary

First degree relatives affected ☐ YES ☐ NO

If yes, indicate their relationship to the patient: o Mother o Father o Sibs

Surgery treatment ☐ YES ☐ NO

If yes, indicate the surgery type:

- o Total colectomy with ileostomy
- o Total colectomy with pouch
- o Total colectomy with residual rectum
- o Segmental colon resection
- o Stricturoplasty
- o Segmental small bowel + colon resection
- o Surgery for fistulas

Previous appendectomy ☐ YES ☐ NO

Smoking status ☐ Current smoker ☐ Non-smoker ☐ Ex-smoker

CURRENT DRUG TREATMENT

- ☐ Mesalasin oral
- ☐ Mesalasin rectal
- ☐ Sulfasalasin
- ☐ Corticosteroid
  - o Systemic > 10 mg
  - o Systemic <= 10 mg
  - o Rectal
- ☐ Budesonid oral
- ☐ Azathioprin
- ☐ Metothrexate
- ☐ Infliximab
- ☐ Other biologics
- ☐ Cyclosporine
- ☐ Metronidazole
- ☐ Other antibiotics
- ☐ Probiotics
- ☐ Loperamid
- ☐ Other……

Date, name and signature (doctor) _____

Table 1. *List of the genotyped SNPs*

Src	Marker	A1	Chr	Nearby gene (relative position)	Disease	Study
A	rs8176785	A	11p15.1	*NELL1(missence)*	CD, UC	152
A	rs1992662	T	5p13.1	*PTGER4 (+286kb)*	CD	152
A	rs1992660	A	5p13.1	*PTGER4 (+265kb)*	CD	152
A	rs10484545	G	6p22.1	*OR14J1 (+40kb)*	CD	152
A	rs1553575	G	5p13.1	*PTGER4 (+177kb)*	CD	152
A	rs830772	A	8q21.11	*HNF4G*	CD	152
A	rs4743484	C	9q31.1	*PPP3R2 (intronic)*	CD	152
A	rs7868736	T	9q32	*ZNF618 (+110kb); RGS3 (-168kb)*	CD	152
A	rs6947579	G	7q31.33	*GRM8 (+545kb)*	CD	152
A	rs2925757	G	2q24.2	*ITGB6*	CD	152
A	rs10521209	T	16q12.1	*NOD2 (intronic)*	CD	**54, 55,** 152
A	rs2076756	G	16q12.1	*NOD2(intronic)*	CD	**54, 55,** 152
A	rs272867	T	5q31.1	*SLC22A4*	CD	**69**, 152
A	rs2631372	G	5q31.1	*SLC22A5*	CD	**69,** 152
A	rs2241880	G	2q37.1	*ATG16L1(missence)*	CD	**57,** 58, 67; 152
A	rs2066845	C	16q12.1	*NOD2(missence)*	CD	**54, 55,** 67, 152;
A	rs2066844	T	16q12.1	*NOD2(missence)*	CD	**54, 55,** 58, 152
A	rs2289310	C	10q22.3	*DLG5(missence)*	CD	**297,** 152
A	rs1248696	G	10q22.3	*DLG5(missence)*	CD	**297,** 152
B	rs17234657	G	5p13.1	*PTGER4(+278kb)*	CD	58, 159
B	rs9858542	A	3p21.31	*BSN (coding-synon)*	CD, UC	58, 159
B	rs10761659	G	10q21.2	*ZNF365 (-14kb)*	CD, UC	58, 159
B	rs10883365	G	10q24.2	*NKX2-3*	CD, UC	58, 159
B	rs2836754	T	21q22.2	*ETS2 (-95kb)*	CD	58, 159
B	rs9292777	T	5p13.1	*PTGER4 (+242kb)*	CD	58, 159
B	rs10077785	C	5q31.1	*IRF1, C5orf56*	CD	58, 159
B	rs13361189	C	5q33.1	*IRGM (+3kb)*	CD	58, 159
B	rs4958847	A	5q33.1	*IRGM*	CD	58, 159
B	rs6887695	C	5q33.3	*IL12B (-65kb)*	CD, UC	58, 159

Table 1 continued

Src	Marker	A1	Chr	**Nearby gene** (relative position)	Disease	Study
B	rs12035082	C	1q24.3	*TNFSF18 (+112kb)*	CD	58, 159
B	rs2542151	G	18p11.21	*PTPN2 (+5.5kb)*	CD, UC	58, 159, 67
C	rs10758669	C	9p24.1	*JAK2 (+3.5kb)*	CD, UC	67, 169
C	rs1736135	T	21q21.1	*USP25 (+297 kb); NRIP1 (-368kb)*	CD, UC	67, 169
C	rs2872507	A	17q12	*ORMDL3 (-6.6kb)*	CD, UC	67, 169, 166
C	rs17582416	G	10p11.21	*CUL2 (+11kb)*	CD, UC	67
C	rs744166	T	17q21.2	*STAT3 (intronic)*	CD, UC	67
C	rs11175593	T	12q12	*LRRK2 (+17kb)*	CD	67
C	rs3764147	G	13q14.11	*C13orf31(missence)*	CD	67
C	rs762421	G	21q22.3	*ICOSLG (+31kb)*	CD	67
C	rs1456893	A	7p12.2	*IKZF1(+75kb)*	CD	67
C	rs7927894	T	11q13.5	*C11orf30 (-45kb)*	CD	67
C	rs7746082	C	6q21	*PRDM1 (+99kb)*	CD, UC	67, 169
C	rs2476601	G	1p13.2	*PTPN22(coding-synon)*	CD	67
C	rs1551398	A	8q24.13	*TRIB1 (-98kb)*	CD	67
C	rs10045431	C	5q33.3	*IL12B (-57kb)*	CD, UC	67
C	rs2274910	C	1q23.3	*ITLN1 (intronic)*	CD	67
C	rs2301436	T	6q27	*CCR6 (intronic)*	CD	67
C	rs11584383	T	1q32.1	*C1orf81 (-70b)*	CD, UC	67
C	rs6908425	T	6p22.3	*CDKAL1(intronic)*	CD, UC	67
C	rs9286879	G	1q24.3	*TNFSF18 (+148 kb)*	CD	67
C	rs10995271	C	10q21.2	*ZNF365 (-7kb)*	CD	67
C	rs2188962	T	5q31.1	*C5orf56(intronic)*	CD	67
C	rs3828309	G	2q37.1	*ATG16L1(intronic)*	CD	67
C	rs11747270	G	5q33.1	*IRGM*	CD	180, 67
C	rs11190140	T	10q24.2	*NKX2-3(+1kb)*	CD, UC	180, 67
C	rs4263839	G	9q32	*TNFSF15(intronic)*	CD	157, 67
C	rs4613763	C	5p13.1	*PTGER4 (+287 kb)*	CD	66; 67
C	rs11465804	A	1p31.3	*IL23R (intronic)*	CD, UC	151, 67, 163

Table 1 continued

Src	Marker	A1	Chr	**Nearby gene** (relative position)	Disease	Study
C	rs3197999	T	3p21.31	*MST1(missence)*	CD, UC	68, 67
C	rs2066847	insC	16q12.1	*NOD2(missence)*	CD	54, 67, 152
D	rs10974944	G	9p24.1	*JAK2 (intronic)*	UC	161
D	rs9268858	A	6p21.32	*HLA-DRA (-17kb), BTNL2 (-55kb), HLA-DRB1, HLA-DRB5 (55kb)*	UC	161
D	rs9268877	T	6p21.32	*HLA-DRA (-18kb), BTNL2 (-56kb), HLA-DRB1, HLA-DRB5 (54kb)*	UC	161
D	rs7712957	G	5q13.3	*S100Z (+7kb)*	UC	161
D	rs7611991	G	3p12.1	*CADM2*	UC	161
D	rs12612347	T	2q35	*ARP2C (+24kb)*	UC	161
D	rs3024505	T	1q32.1	*IL10(+1kb)*	UC	161
D	rs9268480	G	6p21.32	*BTNL2 (missence)*	CD, UC	68, 161
E	rs10753575	A	1p36.13	*RNF186 (-22kb)*	UC	163
E	rs6426833	A	1p36.13	*OTUD3 (+37kb)*	UC	163
E	rs2395185	G	6p21.32	*HLA-DRA (-20kb), BTNL2 (-58kb), HLA-DRB1, HLA-DRB5 (52kb)*	UC	163
E	rs7134599	A	12q15	*IFN-γ (+48kb)*	UC	163
E	rs1558744	A	12q15	*IFN-γ (+44kb)*	UC	163
E	rs3806308	G	1p36.13	*RNF186 (- 1kb)*	UC	163
E	rs1004819	A	1p31.3	*IL23R (intronic)*	CD, UC	151, 163
E	rs11209026	A	1p31.3	*IL23R (missence)*	CD, UC	58, 151, 152, 163
E	rs10889677	A	1p31.3	*IL23R (untranslated-3)*	CD, UC	151, 163
F	rs7809799	G	7q22.1	*SMURF1 (-19kb), KPNA7 (11kb)*	UC	167
F	rs5771069	G	22q13.33	*IL17REL*	UC	167

Src – source, A1 – risk allele, chr – chromosome, CD – Crohn's disease, UC – ulcerative colitis, kb - kilo bases. Original studies which results were replicated in the above mentioned GWA scans are presented in **bold**.

Table 2. *SNPlex™ genotyping pools*

Marker	Gene	Position (bp)	A1	A2	Probe sequence
rs8176785	*NELL1*	20761862	A	G	ACTCTTGTTC[T/C]GGAACAGCTG
rs1992662	*5p13.1*	40429609	C	T	ATTCTCCTTA[G/A]CATTCCTTTA
rs1992660	*5p13.1*	40450824	A	G	AAATTAGTTA[T/C]CATCTGCATG
rs10484545	6p22.1	29342489	C	G	GGATTATAAC[C/G]ATGAAGCGGC
rs1553575	5p13.1	40538689	A	G	AACAGCATTC[A/G]TTCTTAACAT
rs830772	*HNF4G*	76515133	G	T	GAGCTACTTT[C/A]TTTGCAGTGG
rs4743484	*PPP3R2*	103519301	C	T	TTCAGTGAGG[C/T]TAAAATTCTA
rs7868736	9q32	115568004	C	T	CATTGATAAA[C/T]TCTGAAGCCA
rs6947579	7q31.33	125320242	C	G	TCCACTGTTT[G/C]ACTTATCCAT
rs2925757	*ITGB6*	160809415	C	T	GCTATGTAAC[G/A]TGACTTCAGA
rs10521209	*NOD2*	49313210	G	T	TTGAAAAATG[C/A]GGTCAGGCTG
rs2076756	*NOD2*	49314382	A	G	TATCTTAAGG[A/G]CCAATTCCAA
rs272867	*SLC22A4*	131708956	C	T	TTGTATCTAC[C/T]GGCAAAATAT
rs2631372	*SLC22A5*	131731477	C	G	TTCTTACTTC[C/G]TGAAGATGGA
rs2241880	*ATG16L1*	233848107	C	T	CAATGTGGAT[G/A]CTCATCCTGG
rs2066845	*NOD2*	49314041	C	G	CAGATTCTGG[C/G]GCAACAGAGT
rs2289310	*DLG5*	79240879	A	C	AGCACCCCCC[A/C]AGCCAAGCAG
rs1248696	*DLG5*	79286611	C	T	CTCACTGACC[G/A]GCAAGTGAAT
rs17234657	5p13.1	40437266	G	T	CAGTCACGTT[G/T]TCAAATAGCT
rs9858542	*BSN*	49676987	A	G	GCATACCTTC[T/C]GTCAGTTTGC
rs10761659	*ZNF365*	64115570	A	G	CTCTCAAACT[A/G]TAACAGAAGG
rs10883365	*NKX2-3*	101277754	A	G	TTGGCACAAA[T/C]ACCTTCAAAC
rs2836754	21q22.2	39213610	C	T	TCAGTTCTCA[C/T]AATCTTCTCT
rs9292777	5p13.1	40473705	C	T	GGTTCCCCAA[C/T]ATATCAGTTA
rs10077785	*IBD5*	131829057	C	T	GCTTTGCCTC[C/T]GTTACCTACA
rs13361189	*IRGM*	150203580	C	T	GCTTGAAAAT[C/T]GGATGTATAT
rs4958847	*IRGM*	150219780	A	G	TGCCCAATAT[A/G]GCTAAATAAT
rs6887695	*IL12B*	158755223	C	G	CCAGACTATT[G/C]ACCACTACAC
rs12035082	1q24.3	171165000	C	T	AAGTGAGAGA[C/T]GTTCTTAGTA

Table 2 continued

Marker	Gene	Position (bp)	A1	A2	Probe sequence
rs2542151	*PTPN2*	12769947	G	T	TGGTTCGGGC[G/T]CTTCCTGAGA
rs10758669	*JAK2*	4971602	A	C	ATACCTCCTC[T/G]GTACTTCAGC
rs1736135	*21q21.1*	15727091	C	T	AGTGATATTC[C/T]CTCCAGTGTT
rs2872507	*ORMDL3*	35294289	A	G	GTATCCTGCC[A/G]TGGTTTTCTA
rs17582416	*CUL2*	35327656	G	T	TACATGTAGA[G/T]TGTGAAAGAC
rs744166	*STAT3*	37767727	C	T	ATTACTGTCA[G/A]GCTCGATTCC
rs11175593	*LRRK2*	38888207	C	T	CACTTTTCCC[G/A]TTTAGGTGAA
rs3764147	*C13orf31*	43355925	A	G	ATAATCCAGA[T/C]GTCATTGGAA
rs762421	*ICOSLG*	44439989	A	G	AATCTGCTCT[T/C]TTGATTTTTG
rs1456893	7p12.2	50240218	A	G	CGGAAGAGAA[A/G]AATTCAGGAA
rs7927894	*C11orf30*	75978964	C	T	TCAAATGCCC[G/A]ATTCAAAACT
rs7746082	*6q21*	106541962	C	G	AAGAACTTTT[C/G]ATGGCCTCAG
rs2476601	*PTPN22*	114179091	A	G	ACTTCCTGTA[T/C]GGACACCTGA
rs1551398	8q24.13	126609233	C	T	AGCCGCCTGT[G/A]TTCCAGTTCC
rs10045431	*IL12B*	158747111	A	C	CACAGCCCAG[A/C]ATTAAACTCT
rs2274910	*ITLN1*	159118670	C	T	GAGGGTTCAT[C/T]TCAGCCCCAT
rs2301436	*CCR6*	167357978	A	G	AAAGGGCTTC[T/C]GAAAAAAATC
rs11584383	1q32.1	199202489	C	T	AAGGCGGCTT[G/A]CAAGTGGCTC
rs10995271	*ZNF365*	64108492	C	G	AACTCATGCT[C/G]TCTCTCAGGT
rs2188962	*C5orf56*	131798704	C	T	TCTCTGACCC[C/T]GTGTTCTGGC
rs3828309	*ATG16L1*	233845149	C	T	GGCTCAGCTC[G/A]TATTTGCAGT
rs11747270	*IRGM*	150239060	A	G	ATTTATGTAA[T/C]ACAGACCTCA
rs11190140	*NKX2-3*	101281583	C	T	TTTCAATAGG[C/T]GGAAAAGAAG
rs4263839	*TNFSF15*	116606261	A	G	TATCATTAAA[T/C]TCATCTTCCT
rs4613763	*PTGER4*	40428485	C	T	TTTATTCCCA[C/T]CACATTTCTT
rs11465804	*IL23R*	67475114	G	T	ATGGGCAATT[C/A]CTAAAAGACT
rs3197999	*MST1*	49696536	C	T	GCTGGCCAGC[G/A]GGACCTTGCG
rs2066847	*NOD2*	49321279	-	C	CTCCTGCAGG[-/C]CCCTTGAAAG
rs10974944	*JAK2*	5060831	C	G	AAATGTGGCT[G/C]ATCATCAACC

Table 2 continued

Marker	Gene	Position (bp)	A1	A2	Probe sequence
rs9268858	6p21.32	32537736	C	T	CCTGCATTGA[C/T]TGAATGGATT
rs9268877	6p21.32	32539125	A	G	TAGCTTGCAT[A/G]GTTAGCACTG
rs7712957	*S100Z*	76174452	C	T	AATCAAAGCT[C/T]GTGCCTAGAG
rs7611991	*CADM2*	85842248	A	G	GGACAAATAG[T/C]GTAAATGATA
rs12612347	*ARP2C*	218765583	A	G	AGTGTAGGAG[A/G]TTGCGGCCAC
rs9268480	*BTNL2*	32471822	C	T	AACTGGCCTC[C/T]TGGTAGACAT
rs10753575	*RNF186*	20036455	C	T	CTAAATGAAA[C/T]GGACAATCTC
rs6426833	*OTUD3*	20044447	A	G	TCTCCGTTGC[T/C]GACTCAGCTG
rs2395185	6p21.32	32541145	G	T	CCAGGGAAGA[C/A]AAATTTTTGG
rs7134599	12q15	66786342	A	G	ATTATAAGCA[T/C]GTCTTGATCT
rs1558744	12q15	66790859	A	G	ATGTTGTCAC[A/G]TTGAAAACCA
rs1004819	*IL23R*	67442801	C	T	GATTCTTACT[G/A]TGCTATCTGC
rs11209026	*IL23R*	67478546	A	G	AGATCATTCC[A/G]AACTGGGTAG
rs10889677	*IL23R*	67497708	A	C	TCTTCTGCCT[A/C]ATTTCTTAAA

Probe sequences listed were provided by Applied Biosystems. A1 = allele1, A2 = allele2, bp – base pairs.

Table 3. *TaqMan® genotyping assays*

SNP	Gene	Position (bp)	Sequence ([VIC/FAM])	Assay ID	Type
rs3806308	*RNF186*	20015453	TTACTTTGCCCTCAAGGGCAAATGG [C/T] GGGGTGGCATGTGCACTCCCTCAAA	C_399646_10	AoD
rs6908425	*CDKAL1*	20836710	AAGAAGAGTGAATATGTATGGCTTA [C/T] TGAGATTATTTATTATGGGGCCACA	C_2504037_10	AoD
rs9286879	1q24.3	171128857	GGGAGGAATGAAAATAGAAGCATAT [A/G] TTGAGGGACTACTCCAGGGGAAGAG	C_2475289_10	AoD
rs2066844	*NOD2*	49303427	CCAGACATCTGAGAAGGCCCTGCTC [C/T] GGCGCCAGGCCTGTGCCCGCTGGTG	C_11717468_20	AoD
rs3024505	*IL10*	205006527	GGGCTGCCCAGGCAGAGCGTGAGGG [A/G] GACTAGTGTTTACTCAGCTCATTTT	C_15983681_20	AoD
rs7809799	7q22.1	98760504	AATCTGTATTCCAATCAGATTCTTT [A/G] AAAAAAAGTATATGTAAGGCTGGAC	C_30202907_10	AoD
rs5771069	*IL17REL*	50435480	CCCTGGGAAGGTCTAGGAAGGCAAA [A/G] GCAGGGGCGGCTGCCAGGTCACCCT	C_29975365_10	AoD

Primer/probe sequences listed were provided by Applied Biosystems. AoD = Assay-on-Demand. In general, allele 1 is labeled with VIC and allele 2 with FAM; bp – base pairs, SNP – single nucleotide polymorphism.

Table 4. *Tagging SNPs.* 78 SNPs were captured with 69 Tag SNPs at r^2 >= 0.8.

Test	Alleles captured
rs3197999	rs9858542, rs3197999
rs3828309	rs2241880, rs3828309
rs9268858	rs9268858, rs2395185
rs17234657	rs4613763, rs17234657
rs11190140	rs10883365, rs11190140
rs1558744	rs1558744, rs7134599
rs1004819	rs1004819, rs10889677
rs9292777	rs9292777, rs1992660
rs13361189	rs11747270, rs13361189
rs4263839	rs4263839
rs1736135	rs1736135
rs1992662	rs1992662
rs11584383	rs11584383
rs272867	rs272867
rs10045431	rs10045431
rs2925757	rs2925757
rs10077785	rs10077785
rs7746082	rs7746082
rs11465804	rs11465804
rs9268480	rs9268480
rs830772	rs830772
rs3024505	rs3024505
rs10521209	rs10521209
rs2631372	rs2631372
rs2066847	rs2066847
rs4958847	rs4958847
rs6947579	rs6947579
rs4743484	rs4743484
rs7868736	rs7868736
rs10761659	rs10761659
rs10753575	rs10753575

Test	Alleles captured
rs11209026	rs11209026
rs2301436	rs2301436
rs744166	rs744166
rs762421	rs762421
rs7712957	rs7712957
rs2066845	rs2066845
rs1248696	rs1248696
rs1456893	rs1456893
rs3806308	rs3806308
rs2274910	rs2274910
rs2542151	rs2542151
rs12035082	rs12035082
rs9268877	rs9268877
rs8176785	rs8176785
rs9286879	rs9286879
rs6887695	rs6887695
rs2836754	rs2836754
rs7611991	rs7611991
rs2066844	rs2066844
rs2872507	rs2872507
rs12612347	rs12612347
rs10995271	rs10995271
rs17582416	rs17582416
rs10484545	rs10484545
rs10758669	rs10758669
rs10974944	rs10974944
rs2076756	rs2076756
rs6908425	rs6908425
rs2476601	rs2476601
rs1551398	rs1551398
rs2188962	rs2188962

Table 4 continued

Test	Alleles captured
rs1553575	rs1553575
rs6426833	rs6426833
rs3764147	rs3764147

Test	Alleles captured
rs7927894	rs7927894
rs11175593	rs11175593
rs7809799	rs7809799
rs5771069	rs5771069

Table 5. *Summary of results for SNPs significantly associated with Crohn's disease*

Gene marker	Gene	A1	Controls (n=1097)			CD (n=128)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CCA}	P_{CCG}	OR (95% CI)	P_{CORR}
rs2066847	*NOD2*	C	3/79/1000	0.039	0.226	6/27/92	0.156	**2.46×10^{-15}**	**4.31×10^{-14}**	4.52 (3.02–6.78)	**1.62×10^{-13}**
rs2076756	*NOD2*	G	50/312/705	0.193	0.049	16/56/54	0.349	**8.43×10^{-9}**	**2.53×10^{-8}**	2.24 (1.69–2.97)	**5.56×10^{-7}**
rs10521209	*NOD2*	G	211/532/333	0.443	1.00	13/57/55	0.332	**7.71×10^{-4}**	**7.70×10^{-4}**	0.62 (0.47–0.82)	0.051
rs2066845	*NOD2*	C	0/18/1065	0.008	1.00	0/7/119	0.028	**3.83×10^{-3}**	**3.65×10^{-3}**	3.41 (1.41–8.25)	0.253
rs3024505	*IL10*	A	17/244/814	0.129	0.893	4/39/85	0.188	**0.010**	**0.016**	1.56 (1.11–2.19)	0.677
rs9268858	6p21.32 (*HLA*)	C	75/360/635	0.238	0.018	14/51/62	0.311	**0.011**	**0.014**	1.44 (1.09–1.92)	0.724
rs2395185	6p21.32 (*HLA*)	T	74/371/634	0.241	0.055	14/50/62	0.31	**0.016**	**0.020**	1.42 (1.07–1.88)	1
rs13361189	*IRGM*	C	2/86/985	0.042	0.710	2/15/110	0.075	**0.017**	**0.019**	1.85 (1.11–3.09)	1
rs9268877	6p21.32 (*HLA*)	G	265/526/284	0.491	0.502	40/61/23	0.569	**0.021**	**0.022**	1.37 (1.05–1.78)	1
rs7712957	*S100Z*	C	2/121/960	0.058	0.574	0/24/103	0.094	**0.021**	**0.019**	1.70 (1.08–2.69)	1
rs1736135	*21q21.1*	C	201/514/359	0.426	0.493	17/55/53	0.356	**0.033**	**0.035**	0.74 (0.57–0.98)	1
rs2301436	*CCR6*	A	257/523/292	0.484	0.463	19/66/41	0.413	**0.033**	**0.034**	0.75 (0.58–0.98)	1
rs11747270	*IRGM*	G	2/89/987	0.043	1.00	2/14/107	0.073	**0.034**	**0.036**	1.75 (1.04–2.95)	1

Table 5 continued

Gene marker	Gene	A1	Controls (n=1097)			CD (n=128)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CCA}	P_{CCG}	OR (95% CI)	P_{CORR}
rs2188962	*C5orf56*	T	110/469/504	0.318	0.944	18/60/49	0.378	0.054	0.156	1.30 (1.00–1.71)	1
rs3764147	*C13orf31*	G	77/441/553	0.278	0.446	13/57/57	0.327	0.101	0.247	1.26 (0.96–1.67)	1
rs11209026	*IL23R*	A	8/125/939	0.066	0.127	0/10/117	0.039	0.102	0.109	0.58 (0.30–1.12)	1
rs11584383	1q32.1	C	47/382/647	0.221	0.376	3/39/85	0.177	0.107	0.102	0.76 (0.54–1.06)	1
rs7927894	*C11orf30*	T	101/423/547	0.292	0.160	16/52/56	0.339	0.126	0.316	1.24 (0.94–1.64)	1
rs11465804	*IL23R*	G	1/102/969	0.049	0.509	0/7/120	0.028	0.133	0.127	0.56 (0.26–1.21)	1
rs2872507	*ORMDL3*	A	176/534/359	0.414	0.378	17/57/50	0.367	0.150	0.309	0.82 (0.62–1.08)	1
rs9292777	5p13.1	C	161/509/409	0.385	0.898	12/62/52	0.341	0.175	0.257	0.83 (0.63–1.09)	1
rs10484545	6p22.1	G	8/194/873	0.098	0.602	0/18/107	0.072	0.190	0.184	0.72 (0.43–1.18)	1
rs11190140	*NKX2-3*	T	207/510/279	0.464	0.373	29/65/27	0.508	0.191	0.381	1.20 (0.91–1.56)	1
rs6887695	*IL12B*	C	65/418/586	0.256	0.423	7/59/58	0.294	0.196	0.186	1.21 (0.91–1.62)	1
rs4958847	*IRGM*	A	7/186/884	0.093	0.585	3/23/97	0.118	0.205	0.203	1.31 (0.86–1.98)	1
rs2631372	*SLC22A5*	G	145/506/411	0.375	0.601	12/56/52	0.333	0.208	0.428	0.83 (0.63–1.11)	1

Table 5 continued

Gene marker	Gene	A1	Controls (n=1097)			CD (n=128)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CCA}	P_{CCG}	OR (95% CI)	P_{CORR}
rs11175593	*LRRK2*	T	0/43/1034	0.02	1.000	0/8/117	0.032	0.211	0.206	1.62 (0.75–3.49)	1
rs10995271	*ZNF365*	C	177/497/393	0.399	0.371	25/61/41	0.437	0.240	0.512	1.17 (0.90–1.52)	1
rs10758669	*JAK2*	C	123/518/436	0.355	0.110	17/65/45	0.39	0.270	0.515	1.16 (0.89–1.52)	1
rs272867	*SLC22A4*	C	257/501/300	0.48	0.096	22/66/36	0.444	0.281	0.243	0.86 (0.66–1.13)	1
rs1992660	*5p13.1*	G	162/502/409	0.385	0.699	13/63/51	0.35	0.284	0.340	0.86 (0.66–1.13)	1
rs7611991	*CADM2*	A	32/334/705	0.186	0.363	1/37/85	0.159	0.295	0.285	0.83 (0.58–1.18)	1
rs830772	*HNF4G*	T	43/325/707	0.191	0.490	5/30/87	0.164	0.303	0.430	0.83 (0.58–1.18)	1
rs4743484	*PPP3R2*	T	96/448/533	0.297	0.884	12/42/70	0.266	0.310	0.250	0.86 (0.64–1.15)	1
rs9858542	*BSN*	A	60/369/627	0.232	0.546	10/45/70	0.260	0.315	0.531	1.17 (0.86–1.57)	1
rs2542151	*PTPN2*	G	21/291/754	0.156	0.295	5/23/96	0.133	0.339	0.048	0.83 (0.56–1.22)	1
rs1551398	*8q24.13*	C	175/500/401	0.395	0.372	13/65/47	0.364	0.342	0.203	0.88 (0.67–1.15)	1
rs2925757	*ITGB6*	C	21/248/805	0.135	0.695	2/35/88	0.156	0.361	0.362	1.18 (0.82–1.70)	1
rs1992662	*5p13.1*	C	118/462/497	0.324	0.487	10/55/61	0.298	0.395	0.575	0.88 (0.66–1.18)	1

Table 5 continued

Gene marker	Gene	A1	Controls (n=1097)			CD (n=128)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CCA}	P_{CCG}	OR (95% CI)	P_{CORR}
rs17582416	*CUL2*	G	105/481/482	0.324	0.364	14/58/52	0.35	0.409	0.703	1.12 (0.85–1.48)	1
rs3197999	*MST1*	T	58/355/618	0.228	0.479	9/36/62	0.252	0.429	0.507	1.14 (0.82–1.58)	1
rs10883365	*NKX2–3*	G	241/538/297	0.474	0.951	31/64/31	0.50	0.434	0.730	1.11 (0.85–1.44)	1
rs6947579	*7q31.33*	G	74/401/596	0.256	0.575	6/46/72	0.234	0.442	0.660	0.89 (0.65–1.21)	1
rs10045431	*IL12B*	A	59/374/630	0.231	0.730	10/32/81	0.211	0.479	0.092	0.89 (0.65–1.23)	1
rs8176785	*NELL1*	G	79/440/554	0.279	0.544	10/54/60	0.298	0.513	0.788	1.10 (0.83–1.47)	1
rs6908425	*CDKAL1*	T	74/392/626	0.247	0.255	4/60/64	0.264	0.562	0.522	1.09 (0.81–1.47)	1
rs3806308	*RNF186*	T	207/543/334	0.441	0.623	31/58/39	0.46	0.576	0.490	1.08 (0.83–1.40)	1
rs2066844	*NOD2*	T	2/53/1034	0.026	0.165	0/8/120	0.032	0.589	0.639	1.23 (0.58–2.61)	1
rs7868736	*9q32*	T	70/432/575	0.266	0.390	8/47/71	0.25	0.596	0.814	0.92 (0.68–1.25)	1
rs17234657	*5p13.1*	G	27/316/724	0.173	0.336	2/43/81	0.187	0.604	0.596	1.09 (0.78–1.53)	1
rs4613763	*PTGER4*	C	29/318/731	0.174	0.461	2/43/81	0.187	0.633	0.626	1.09 (0.78–1.52)	1
rs2836754	*21q22.2*	T	212/539/325	0.448	0.712	27/55/44	0.433	0.651	0.380	0.94 (0.72–1.23)	1

Table 5 continued

Gene marker	Gene	A1	Controls (n=1097)			CD (n=128)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CCA}	P_{CCG}	OR (95% CI)	P_{CORR}
rs1004819	*IL23R*	T	79/409/584	0.265	0.530	5/53/67	0.252	0.672	0.309	0.94 (0.69–1.27)	1
rs12035082	*1q24.3*	C	151/523/402	0.383	0.402	21/58/47	0.397	0.678	0.704	1.06 (0.81–1.38)	1
rs1248696	*DLG5*	T	8/193/868	0.098	0.601	2/22/99	0.106	0.693	0.690	1.09 (0.71–1.68)	1
rs3828309	*ATG16L1*	C	230/562/283	0.475	0.126	29/66/32	0.488	0.698	0.920	1.05 (0.81–1.37)	1
rs9286879	*1q24.3*	G	43/385/660	0.217	0.180	7/41/83	0.207	0.729	0.722	0.95 (0.69–1.30)	1
rs762421	*ICOSLG*	G	132/503/438	0.357	0.550	16/54/54	0.347	0.741	0.779	0.95 (0.72–1.26)	1
rs2241880	*ATG16L1*	C	232/557/280	0.478	0.159	28/66/31	0.488	0.754	0.942	1.04 (0.80–1.36)	1
rs6426833	*OTUD3*	A	252/541/281	0.487	0.807	29/66/30	0.496	0.776	0.845	1.04 (0.80–1.35)	1
rs10077785	*IBD5*	T	63/405/598	0.249	0.682	10/44/71	0.256	0.811	0.595	1.04 (0.77–1.40)	1
rs7746082	*6q21*	C	76/417/577	0.266	0.938	7/53/63	0.272	0.828	0.625	1.03 (0.77–1.39)	1
rs1553575	*5p13.1*	A	163/542/366	0.405	0.113	22/55/47	0.399	0.855	0.412	0.98 (0.75–1.28)	1
rs10974944	*JAK2*	G	99/495/481	0.322	0.081	11/60/54	0.328	0.856	0.917	1.03 (0.78–1.36)	1
rs10889677	*IL23R*	A	84/405/585	0.267	0.242	7/52/67	0.262	0.869	0.556	0.98 (0.73–1.31)	1

Table 5 continued

Gene marker	Gene	A1	Controls (n=1097)			CD (n=128)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CCA}	P_{CCG}	OR (95% CI)	P_{CORR}
rs2476601	*PTPN22*	A	26/259/793	0.144	0.387	5/27/93	0.148	0.873	0.501	1.03 (0.71–1.49)	1
rs4263839	*TNFSF15*	A	114/474/486	0.327	0.945	14/52/58	0.323	0.893	0.893	0.98 (0.74–1.30)	1
rs10753575	*RNF186*	C	194/542/321	0.44	0.212	27/56/41	0.444	0.913	0.411	1.02 (0.78–1.32)	1
rs2274910	*ITLN1*	T	74/448/525	0.285	0.111	16/37/67	0.288	0.926	0.007	1.01 (0.75–1.36)	1
rs744166	*STAT3*	C	152/515/412	0.38	0.698	17/61/48	0.377	0.938	0.980	0.99 (0.76–1.30)	1
rs12612347	*ARP2C*	G	238/549/285	0.478	0.426	31/59/36	0.48	0.950	0.643	1.01 (0.78–1.31)	1
rs1456893	*7p12.2*	G	96/444/532	0.297	0.826	8/59/59	0.298	0.974	0.397	1.01 (0.76–1.34)	1
rs1558744	*12q15*	A	161/485/416	0.38	0.329	22/52/52	0.381	0.975	0.608	1.00 (0.77–1.31)	1
rs10761659	*ZNF365*	A	225/527/318	0.457	0.806	24/65/35	0.456	0.979	0.795	1.00 (0.77–1.30)	1
rs7134599	*12q15*	A	127/471/470	0.339	0.586	16/54/57	0.339	0.979	0.936	1.00 (0.76–1.31)	1

GT – genotype count (11 = homozygous for minor allele; 12 = heterozygous for common allele; 22 = homozygote for common allele); A1 – minor allele; MAF – minor allele1 frequency, P_{HWE} – P-values for distribution of genotypes within the control group; P_{CCA} – P-values from an allele-based case-control comparison with 1degree of freedom; P_{CCG} – P-values from an genotype-based case-control comparison with 2 degrees of freedom; P_{CORR} - P-values after correction for multiple testing (76 independent tests); OR (95% CI) – odds ratio for carriership of the rarer allele (95% confidence interval of OR); P-values <0.05 are highlighted in **bold**.

Table 6. *Summary of results for SNPs significantly associated with ulcerative colitis*

Gene marker	Gene	A1	Controls (n=1154)			UC (n=444)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CMH}	P_{CCG}	OR (95% CI)	P_{CORR}
rs1736135	*21q21.1*	C	209/544/377	0.426	0.626	46/201/187	0.338	**8.01×10^{-6}**	**3.65×10^{-5}**	0.69 (0.59–0.81)	**4.89×10^{-4}**
rs7746082	*6q21*	C	82/435/608	0.266	0.760	48/200/189	0.339	**6.41×10^{-5}**	**2.95×10^{-4}**	1.41 (1.19–1.67)	**3.91×10^{-3}**
rs10758669	*JAK2*	C	131/543/459	0.355	0.136	76/229/134	0.434	**8.08×10^{-5}**	**1.61×10^{-4}**	1.38 (1.17–1.62)	**4.93×10^{-3}**
rs2872507	*ORMDL3*	A	183/559/383	0.411	0.389	97/229/110	0.485	**1.24×10^{-4}**	**7.11×10^{-4}**	1.36 (1.16–1.59)	**7.59×10^{-3}**
rs3806308	*RNF186*	T	223/571/346	0.446	0.675	63/205/172	0.376	**2.40×10^{-4}**	**1.59×10^{-3}**	0.74 (0.63–0.87)	**0.015**
rs3024505	*IL10*	A	17/255/859	0.128	0.790	16/122/307	0.173	**1.04×10^{-3}**	**2.53×10^{-3}**	1.43 (1.16–1.77)	0.064
rs11209026	*IL23R*	A	8/133/987	0.066	0.143	0/33/408	0.037	**2.16×10^{-3}**	**8.20×10^{-3}**	0.55 (0.38–0.81)	0.132
rs3197999	*MST1*	T	62/371/652	0.228	0.343	44/143/227	0.279	**3.21×10^{-3}**	**2.94×10^{-3}**	1.32 (1.10–1.58)	0.196
rs9268877	6p21.32 (*HLA)*	G	278/557/296	0.492	0.634	93/194/149	0.436	**4.34×10^{-3}**	**6.79×10^{-3}**	0.80 (0.68–0.93)	0.265
rs6426833	*OTUD3*	A	266/566/297	0.486	0.953	124/224/88	0.541	**6.01×10^{-3}**	**0.019**	1.25 (1.07–1.46)	0.367
rs2395185	6p21.32 (*HLA)*	T	75/390/670	0.238	0.085	23/122/294	0.191	**6.44×10^{-3}**	**0.015**	0.76 (0.63–0.93)	0.393
rs11190140	*NKX2-3*	T	207/510/279	0.464	0.373	117/211/101	0.519	**7.27×10^{-3}**	**0.018**	1.25 (1.06–1.48)	0.443
rs4263839	*TNFSF15*	A	122/496/512	0.327	0.893	32/181/222	0.282	**0.010**	**0.042**	0.80 (0.67–0.95)	0.610

Table 6 continued

Gene marker	Gene	A1	Controls (n=1154)			UC (n=444)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CMH}	P_{CCG}	OR (95% CI)	P_{CORR}
rs9268858	6p21.32 (*HLA)*	C	76/380/671	0.236	0.032	23/124/294	0.193	**0.011**	**0.032**	0.78 (0.64–0.94)	0.664
rs11465804	*IL23R*	G	1/110/1017	0.050	0.518	0/26/413	0.030	**0.012**	**0.044**	0.58 (0.38–0.89)	0.754
rs10883365	*NKX2-3*	G	253/563/316	0.472	0.953	121/212/106	0.517	**0.021**	0.067	1.20 (1.03–1.41)	1
rs762421	*ICOSLG*	G	143/526/460	0.360	0.747	77/198/159	0.406	**0.023**	**0.028**	1.21 (1.03–1.42)	1
rs9268480	*BTNL2*	T	64/350/719	0.211	0.016	13/124/296	0.173	**0.025**	**0.046**	0.79 (0.65–0.97)	1
rs9858542	*BSN*	A	61/384/660	0.229	0.609	37/153/237	0.266	**0.026**	0.055	1.23 (1.03–1.48)	1
rs1992660	*5p13.1*	G	173/528/429	0.387	0.616	60/189/192	0.350	**0.046**	0.125	0.85 (0.72–0.99)	1
rs7809799	7q22.1	G	2/113/1036	0.05	0.55	2/59/380	0.07	0.024	0.029	1.62 (0.98–2.70)	1
rs2274910	*ITLN1*	T	79/477/543	0.289	0.067	51/182/202	0.326	0.051	**0.016**	1.18 (1.00–1.40)	1
rs9292777	5p13.1	C	172/534/429	0.387	0.802	59/193/189	0.353	0.058	0.174	0.85 (0.73–1.01)	1
rs17582416	*CUL2*	G	115/499/509	0.325	0.683	57/197/179	0.359	0.063	0.164	1.17 (0.99–1.38)	1
rs3828309	*ATG16L1*	C	248/590/293	0.480	0.137	118/219/105	0.515	0.100	0.128	1.14 (0.98–1.33)	1

Table 6 continued

Gene marker	Gene	A1	Controls (n=1154)			UC (n=444)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CMH}	P_{CCG}	OR (95% CI)	P_{CORR}
rs6887695	*IL12B*	C	71/440/614	0.259	0.535	34/184/218	0.289	0.102	0.221	1.16 (0.97–1.38)	1
rs2241880	*ATG16L1*	C	250/585/289	0.483	0.170	116/218/101	0.517	0.106	0.164	1.14 (0.97–1.33)	1
rs10753575	*RNF186*	C	205/567/340	0.439	0.273	67/220/147	0.408	0.111	0.262	0.88 (0.75–1.03)	1
rs1004819	*IL23R*	T	87/431/610	0.268	0.364	36/186/214	0.296	0.140	0.203	1.14 (0.96–1.35)	1
rs11584383	1q32.1	C	50/395/687	0.219	0.542	16/140/283	0.196	0.166	0.365	0.87 (0.72–1.06)	1
rs6947579	7q31.33	G	76/427/624	0.257	0.815	30/185/222	0.280	0.174	0.245	1.13 (0.95–1.35)	1
rs13361189	*IRGM*	C	2/89/1037	0.041	0.714	1/44/396	0.052	0.176	0.401	1.29 (0.89–1.85)	1
rs7611991	*CADM2*	A	33/357/735	0.188	0.205	13/121/303	0.168	0.181	0.295	0.87 (0.71–1.07)	1
rs11747270	*IRGM*	G	2/93/1039	0.043	1	1/45/390	0.054	0.182	0.403	1.28 (0.89–1.83)	1
rs2476601	*PTPN22*	A	26/273/835	0.143	0.545	11/117/309	0.159	0.221	0.506	1.15 (0.92–1.42)	1
rs5771069	*IL17REL*	G	271/569/271	0.480	0.750	113/215/109	0.50	0.278	0.821	1.11 (0.86–1.44)	1
rs10889677	*IL23R*	A	93/427/609	0.272	0.153	38/182/219	0.294	0.252	0.345	1.11 (0.93–1.31)	1
rs830772	*HNF4G*	T	45/340/746	0.190	0.440	22/135/277	0.206	0.267	0.548	1.12 (0.92–1.36)	1

Table 6 continued

Gene marker	Gene	A1	Controls (n=1154)			UC (n=444)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CMH}	P_{CCG}	OR (95% CI)	P_{CORR}
rs1553575	5p13.1	A	168/570/387	0.403	0.083	63/205/167	0.381	0.268	0.326	0.91 (0.78–1.07)	1
rs3764147	*C13orf31*	G	79/462/586	0.275	0.371	35/185/219	0.290	0.316	0.678	1.09 (0.92–1.30)	1
rs2631372	*SLC22A5*	G	153/535/430	0.376	0.525	49/212/172	0.358	0.336	0.460	0.92 (0.78–1.09)	1
rs10761659	*ZNF365*	A	233/553/340	0.453	0.764	83/208/143	0.431	0.353	0.539	0.93 (0.79–1.09)	1
rs2542151	*PTPN2*	G	23/303/793	0.156	0.426	18/109/304	0.168	0.356	0.059	1.11 (0.89–1.37)	1
rs2188962	*C5orf56*	T	118/491/531	0.319	0.785	49/199/194	0.336	0.375	0.624	1.08 (0.91–1.27)	1
rs2066845	*NOD2*	C	0/19/1121	0.008	1	0/10/428	0.011	0.389	0.414	1.40 (0.65–3.01)	1
rs1551398	8q24.13	C	184/523/426	0.393	0.290	63/205/173	0.375	0.404	0.606	0.93 (0.80–1.10)	1
rs1992662	*5p13.1*	C	124/488/521	0.325	0.543	42/191/208	0.312	0.425	0.702	0.93 (0.79–1.11)	1
rs2066844	*NOD2*	T	2/55/1089	0.026	0.170	0/18/426	0.020	0.435	0.551	0.81 (0.48–1.38)	1
rs10484545	6p22.1	G	8/205/918	0.098	0.403	5/67/364	0.088	0.446	0.314	0.90 (0.69–1.18)	1
rs2301436	*CCR6*	A	276/548/304	0.488	0.371	96/226/118	0.475	0.480	0.488	0.95 (0.81–1.11)	1
rs1558744	12q15	A	168/509/438	0.379	0.309	63/217/160	0.390	0.498	0.422	1.06 (0.90–1.24)	1

Table 6 continued

Gene marker	Gene	A1	Controls (n=1154)			UC (n=444)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CMH}	P_{CCG}	OR (95% CI)	P_{CORR}
rs7134599	12q15	A	132/494/499	0.337	0.593	55/193/189	0.347	0.500	0.868	1.06 (0.90–1.25)	1
rs10045431	*IL12B*	A	63/391/664	0.231	0.614	20/151/264	0.220	0.506	0.696	0.94 (0.78–1.13)	1
rs17234657	5p13.1	G	31/331/761	0.175	0.536	12/122/306	0.166	0.515	0.786	0.93 (0.76–1.15)	1
rs1248696	*DLG5*	T	8/199/916	0.096	0.603	6/77/351	0.103	0.517	0.453	1.09 (0.84–1.42)	1
rs4613763	*PTGER4*	C	33/333/768	0.176	0.759	13/122/307	0.167	0.533	0.785	0.94 (0.76–1.15)	1
rs11175593	*LRRK2*	T	0/44/1089	0.019	1	0/20/416	0.023	0.562	0.528	1.18 (0.68–2.02)	1
rs272867	*SLC22A4*	C	268/531/315	0.479	0.134	103/225/113	0.489	0.611	0.448	1.04 (0.89–1.22)	1
rs12035082	1q24.3	C	163/551/418	0.387	0.416	58/235/146	0.400	0.618	0.224	1.04 (0.89–1.22)	1
rs6908425	*CDKAL1*	T	79/415/655	0.249	0.237	30/169/243	0.260	0.621	0.722	1.05 (0.88–1.25)	1
rs7927894	*C11orf30*	T	104/447/575	0.291	0.219	38/182/213	0.298	0.661	0.700	1.04 (0.88–1.24)	1
rs1456893	7p12.2	G	103/459/565	0.295	0.475	43/165/233	0.285	0.666	0.483	0.96 (0.81–1.14)	1
rs8176785	*NELL1*	G	83/458/588	0.276	0.656	37/160/240	0.268	0.689	0.326	0.97 (0.81–1.15)	1

Table 6 continued

Gene marker	Gene	A1	Controls (n=1154)			UC (n=444)					
			GT (11/12/22)	MAF	P_{HWE}	GT (11/12/22)	MAF	P_{CMH}	P_{CCG}	OR (95% CI)	P_{CORR}
rs7712957	*S100Z*	C	2/128/1010	0.058	0.581	2/49/389	0.060	0.762	0.612	1.05 (0.76–1.46)	1
rs9286879	1q24.3	G	45/410/690	0.218	0.119	15/163/268	0.216	0.783	0.871	0.97 (0.81–1.18)	1
rs4958847	*IRGM*	A	8/191/934	0.091	0.722	2/73/361	0.088	0.795	0.857	0.96 (0.73–1.27)	1
rs744166	*STAT3*	C	157/543/435	0.378	0.570	59/210/171	0.373	0.815	0.968	0.98 (0.84–1.15)	1
rs10521209	*NOD2*	G	220/557/354	0.441	1	96/195/145	0.444	0.840	0.253	1.02 (0.87–1.19)	1
rs10995271	*ZNF365*	C	189/526/407	0.403	0.386	63/229/149	0.403	0.876	0.173	0.99 (0.84–1.16)	1
rs2836754	*21q22.2*	T	222/564/345	0.446	0.810	92/211/138	0.448	0.903	0.753	1.01 (0.86–1.18)	1
rs12612347	*ARP2C*	G	257/575/295	0.483	0.512	106/218/116	0.489	0.905	0.833	1.01 (0.86–1.18)	1
rs4743484	*PPP3R2*	T	101/474/558	0.298	1	36/186/213	0.297	0.924	0.898	0.99 (0.84–1.18)	1

GT – genotype count (11 = homozygous for minor allele; 12 = heterozygous for common allele; 22 = homozygote for common allele); A1 – minor allele; MAF – minor allele1 frequency, P_{HWE} – P-values for distribution of genotypes within the control group; P_{CCA} – P-values from an allele-based case-control comparison with 1degree of freedom; P_{CCG} – P-values from an genotype-based case-control comparison with 2 degrees of freedom; P_{CORR} – p-values after correction for multiple testing (72 independent tests); OR (95% CI) – odds ratio for carriership of the rarer allele (95% confidence interval of OR); P-values <0.05 are highlighted in **bold**.

***Table 7**. UC group SNP-SNP epistasis analysis*

CHR1	SNP1	CHR2	SNP2	OR	STAT	P	P_{CORR}
1	rs2476601	13	rs3764147	2.44	22.98	$\mathbf{1.64\times10^{-6}}$	$\mathbf{3.93\times10^{-3}}$
6	rs7746082	16	rs10521209	0.64	12.96	$\mathbf{3.18\times10^{-4}}$	0.76
5	rs272867	11	rs8176785	1.51	11.64	$\mathbf{6.47\times10^{-4}}$	1
16	rs10521209	18	rs2542151	0.62	10.02	$\mathbf{1.55\times10^{-3}}$	1
1	rs3806308	5	rs6887695	1.48	9.97	$\mathbf{1.60\times10^{-3}}$	1
2	rs2241880	21	rs762421	1.48	9.59	$\mathbf{1.96\times10^{-3}}$	1
3	rs7611991	5	rs6887695	0.57	9.45	$\mathbf{2.11\times10^{-3}}$	1
9	rs4263839	16	rs10521209	0.68	9.44	$\mathbf{2.13\times10^{-3}}$	1
2	rs3828309	21	rs762421	1.46	9.01	$\mathbf{2.69\times10^{-3}}$	1
6	rs9268877	8	rs1551398	1.37	8.77	$\mathbf{3.06\times10^{-3}}$	1
5	rs272867	5	rs13361189	0.49	8.59	$\mathbf{3.38\times10^{-3}}$	1
5	rs2631372	5	rs13361189	0.42	8.58	$\mathbf{3.40\times10^{-3}}$	1
5	rs2631372	11	rs8176785	1.49	8.51	$\mathbf{3.53\times10^{-3}}$	1
5	rs272867	5	rs4958847	0.56	8.38	$\mathbf{3.80\times10^{-3}}$	1
1	rs6426833	7	rs6947579	1.41	8.37	$\mathbf{3.81\times10^{-3}}$	1
5	rs1992662	8	rs830772	0.66	7.94	$\mathbf{4.83\times10^{-3}}$	1
2	rs12612347	10	rs17582416	0.71	7.65	$\mathbf{5.67\times10^{-3}}$	1
1	rs12035082	6	rs6908425	1.44	7.60	$\mathbf{5.83\times10^{-3}}$	1
1	rs11209026	18	rs2542151	2.60	7.49	$\mathbf{6.22\times10^{-3}}$	1
6	rs7746082	10	rs10883365	0.70	7.25	$\mathbf{7.08\times10^{-3}}$	1
5	rs7712957	10	rs10761659	1.79	7.11	$\mathbf{7.68\times10^{-3}}$	1
5	rs272867	5	rs11747270	3.41	7.04	$\mathbf{7.99\times10^{-3}}$	1
6	rs9268858	21	rs762421	1.48	7.00	$\mathbf{8.14\times10^{-3}}$	1
3	rs7611991	16	rs2066844	3.86	6.90	$\mathbf{8.61\times10^{-3}}$	1
5	rs2631372	5	rs4958847	0.55	6.90	$\mathbf{8.64\times10^{-3}}$	1
1	rs10889677	21	rs1736135	1.38	6.88	$\mathbf{8.70\times10^{-3}}$	1
5	rs2631372	5	rs11747270	0.47	6.84	$\mathbf{8.89\times10^{-3}}$	1
5	rs10045431	21	rs1736135	1.47	6.83	$\mathbf{8.97\times10^{-3}}$	1
6	rs9268480	11	rs7927894	0.67	6.70	$\mathbf{9.64\times10^{-3}}$	1

Table 7 continued

CHR1	SNP1	CHR2	SNP2	OR	STAT	P	P_{CORR}
7	rs6947579	8	rs1551398	1.36	6.67	**9.79×10^{-3}**	1
6	rs2395185	21	rs762421	1.47	6.67	**9.81×10^{-3}**	1

CHR – chromosome; STAT – χ^2 value; P-values corrected for 2404 independent SNP-SNP tests performed; OR – presented for SNP-SNP interaction. P-values <0.05 are highlighted in **bold**.

Table 8. *CD group SNP-SNP epistasis analysis*

CHR1	SNP1	CHR2	SNP2	OR	STAT	P	P_{CORR}
6	rs2301436	10	rs10761659	0.51	12.17	**4.85×10^{-4}**	1
1	rs2476601	9	rs10758669	0.36	10.22	**1.39×10^{-3}**	1
1	rs1004819	2	rs12612347	1.88	9.08	**2.59×10^{-3}**	1
1	rs3024505	10	rs10995271	2.49	8.87	**2.90×10^{-3}**	1
16	rs10521209	21	rs762421	0.53	8.83	**2.96×10^{-3}**	1
3	rs3197999	12	rs2836754	0.51	8.64	**3.28×10^{-3}**	1
10	rs17582416	12	rs1558744	1.82	8.58	**3.40×10^{-3}**	1
9	rs7868736	21	rs1736135	1.86	8.18	**4.24×10^{-3}**	1
10	rs1248696	17	rs744166	2.46	7.98	**4.73×10^{-3}**	1
1	rs12035082	3	rs9858542	0.56	7.74	**5.41×10^{-3}**	1
1	rs10889677	2	rs12612347	1.76	7.58	**5.89×10^{-3}**	1
7	rs6947579	9	rs4743484	0.46	7.49	**6.22×10^{-3}**	1
1	rs2476601	9	rs10974944	0.41	7.45	**6.34×10^{-3}**	1
5	rs10045431	13	rs3764147	1.91	7.21	**7.24×10^{-3}**	1
1	rs11209026	8	rs830772	3.85	6.90	**8.64×10^{-3}**	1
1	rs2476601	9	rs4263839	2.13	6.88	**8.72×10^{-3}**	1
1	rs3806308	5	rs1992662	0.59	6.85	**8.85×10^{-3}**	1

CHR – chromosome; STAT – χ^2 value; P-values corrected for 2763 independent SNP-SNP tests performed; OR – presented for SNP-SNP interaction. P-values <0.05 are highlighted in **bold.**

Table 9. *Statistics for pair-wise linkage disequilibrium estimation between markers*

L1	L2	D'	LOD	r^2	CIlow	CIhi	Dist (bp)	T-int
rs10758669	rs10974944	0.835	293.61	0.574	0.8	0.86	89229	293.61
rs3806308	rs10753575	0.011	0.03	0	-0.01	0.08	21002	0.04
rs3806308	rs6426833	0.005	0.01	0	-0.01	0.07	28994	-
rs10753575	rs6426833	0.994	497.63	0.771	0.98	1	7992	497.64

Table 9 continued

L1	L2	D'	LOD	r^2	CIlow	CIhi	Dist (bp)	T-int
rs8176785	rs6908425	0.135	0.95	0.002	0.03	0.25	74848	0.95
rs9268480	rs9268858	0.766	235.25	0.509	0.73	0.8	65914	535.73
rs9268480	rs9268877	0.702	65.81	0.14	0.64	0.76	67303	-
rs9268480	rs2395185	0.766	234.67	0.508	0.73	0.8	69323	-
rs9268858	rs9268877	1	201.92	0.326	0.99	1	1389	1148.22
rs9268858	rs2395185	1	645.82	0.998	0.99	1	3409	-
rs9268877	rs2395185	0.997	198.09	0.325	0.97	1	2020	1078.58
rs2872507	rs17582416	0.01	0.01	0	-0.01	0.09	33367	0.01
rs11175593	rs2836754	0.024	0.01	0	0	0.31	325403	0.01
rs4613763	rs1992662	1	56.24	0.098	0.96	1	1124	790.28
rs4613763	rs17234657	0.998	534.74	0.994	0.98	1	8781	-
rs4613763	rs1992660	1	72.81	0.125	0.97	1	22339	-
rs4613763	rs9292777	1	72.51	0.126	0.97	1	45220	-
rs4613763	rs1553575	0.921	53.98	0.116	0.86	0.96	110204	-
rs1992662	rs17234657	1	56.15	0.098	0.96	1	7657	1861.09
rs1992662	rs1992660	0.997	500.73	0.782	0.98	1	21215	-
rs1992662	rs9292777	0.997	497.54	0.778	0.98	1	44096	-
rs1992662	rs1553575	0.502	72.63	0.178	0.45	0.55	109080	-
rs17234657	rs1992660	1	72.32	0.125	0.97	1	13558	1469.43
rs17234657	rs9292777	1	71.88	0.125	0.97	1	36439	-
rs17234657	rs1553575	0.935	55.03	0.119	0.87	0.97	101423	-
rs1992660	rs9292777	0.993	716.81	0.98	0.98	1	22881	1588.24
rs1992660	rs1553575	0.367	47.86	0.121	0.32	0.41	87865	-
rs9292777	rs1553575	0.356	45.1	0.115	0.31	0.4	64984	274.6
rs2066844	rs10521209	1	12.22	0.021	0.86	1	9783	60.3
rs2066844	rs2066845	1	0.47	0	0.07	0.98	10614	-
rs2066844	rs2076756	1	47.54	0.104	0.95	1	10955	-
rs2066844	rs2066847	0.263	0.04	0	0.03	0.92	17852	-
rs2066844	rs9858542	0.083	0.03	0	0	0.46	373560	-
rs2066844	rs3197999	0.01	0	0	-0.01	0.17	393109	-
rs10521209	rs2066845	1	4.86	0.008	0.67	1	831	194.32

Table 9 continued

L1	L2	D'	LOD	r^2	CIlow	CIhi	Dist (bp)	T-int
rs10521209	rs2076756	0.99	116.99	0.2	0.96	1	1172	-
rs10521209	rs2066847	1	24.18	0.037	0.92	1	8069	-
rs10521209	rs9858542	0.046	0.2	0.001	-0.01	0.14	363777	-
rs10521209	rs3197999	0.011	0.01	0	-0.01	0.11	383326	-
rs2066845	rs2076756	1	17.49	0.04	0.88	1	341	207.37
rs2066845	rs2066847	0.047	0.15	0	-0.01	0.19	7238	-
rs2066845	rs9858542	0.198	0.46	0.001	0.02	0.43	362946	-
rs2066845	rs3197999	0.16	0.28	0.001	0.01	0.41	382495	-
rs2076756	rs2066847	0.991	79.04	0.177	0.95	1	6897	104.4
rs2076756	rs9858542	0.005	0	0	-0.01	0.16	362605	-
rs2076756	rs3197999	0.016	0.01	0	-0.01	0.17	382154	-
rs2066847	rs9858542	0.059	0.02	0	0	0.39	355708	1.06
rs2066847	rs3197999	0.092	0.05	0	0	0.41	375257	-
rs9858542	rs3197999	0.974	531.18	0.946	0.96	0.99	19549	531.53
rs10995271	rs10761659	0.899	235.55	0.448	0.87	0.93	7078	235.55
rs7134599	rs1558744	1	546.03	0.831	0.99	1	4517	546.03
rs1004819	rs11465804	0.878	7.11	0.013	0.65	0.96	32313	473.61
rs1004819	rs11209026	0.96	13.12	0.021	0.81	1	35745	-
rs1004819	rs10889677	0.898	453.38	0.803	0.87	0.92	54907	-
rs11465804	rs11209026	0.972	169.15	0.719	0.93	1	3432	645.97
rs11465804	rs10889677	1	10.32	0.018	0.83	1	22594	-
rs11209026	rs10889677	1	15.14	0.023	0.88	1	19162	478.84
rs7927894	rs7712957	0.115	0.13	0	0	0.36	195488	0.13
rs7712957	rs830772	0.096	0.89	0.003	0.01	0.19	340681	0.89
rs10883365	rs11190140	0.987	651.22	0.968	0.97	1	3829	651.22
rs272867	rs2631372	0.998	378.79	0.625	0.98	1	22521	636.3
rs272867	rs2188962	0.98	256.35	0.437	0.96	1	89748	-
rs272867	rs10077785	0.092	1.16	0.003	0.02	0.17	120101	-
rs2631372	rs2188962	0.976	148.88	0.275	0.95	1	67227	414.76
rs2631372	rs10077785	0.198	8.37	0.023	0.14	0.26	97580	-
rs2188962	rs10077785	0.994	87.82	0.163	0.96	1	30353	97.35

Table 9 continued

L1	L2	D'	LOD	r^2	CIlow	CIhi	Dist (bp)	T-int
rs13361189	rs4958847	0.978	132.06	0.465	0.94	1	16200	350.11
rs13361189	rs11747270	0.993	218.05	0.967	0.96	1	35480	-
rs4958847	rs11747270	0.986	139.1	0.475	0.95	1	19280	357.15
rs10045431	rs6887695	1	58.11	0.109	0.97	1	8112	58.16
rs10045431	rs2274910	0.033	0.05	0	-0.01	0.17	371559	-
rs6887695	rs2274910	0.001	0	0	-0.01	0.06	363447	0.05
rs9286879	rs12035082	0.802	124.58	0.279	0.75	0.84	36143	124.58
rs3828309	rs2241880	1	764.4	0.999	0.99	1	2958	764.4

L1, L2 – the two loci in question; D' – the value of D prime between the two loci; LOD – the log of the likelihood odds ratio, a measure of confidence in the value of D'; r^2 – the correlation coefficient between the two loci; CI low – 95% confidence lower bound on D'; CI hi – the 95% confidence upper bound on D'; Dist – is the distance (in bases) between the loci; T-int – a statistic used by the HapMap Project to measure the completeness of information represented by a set of markers in a region

Table 10. *Statistics for the estimation of the area under the receiver operator characteristic (ROC) curve in CD group*

SNP	Area under the ROC curve	P-value
rs2066847	0.591	0.001
rs2076756	0.616	< 0.0001
rs10521209	0.543	0.112
rs2066845	0.519	0.478
rs13361189	0.526	0.330
rs1736135	0.524	0.381

SNP – single nucleotide polymorphism; ROC – receiver operating characteristic

Table 11. *Statistics for the estimation of the area under the receiver operator characteristic (ROC) curve in UC group*

SNP	Area under the ROC curve	P-value
rs1736135	0.535	0.030
rs7746082	0.553	0.001
rs10758669	0.549	0.002
rs2872507	0.543	0.007
rs3806308	0.524	0.130
rs3024505	0.537	0.023
rs11209026	0.508	0.618
rs3197999	0.522	0.180
rs9268877	0.514	0.381

Table 11 continued

SNP	Area under the ROC curve	P-value
rs6426833	0.529	0.074
rs11190140	0.556	< 0.0001
rs9268858	0.512	0.449
rs9858542	0.520	0.220

SNP – single nucleotide polymorphism; ROC – receiver operating characteristic

Table 12. *Statistics of the β-coefficients for each CD-associated SNP estimated by binary logistic regression analysis*

SNP	β-coefficient	Standart error	Wald	df	P-value
rs2066847	0.910	0.206	19.495	1	<0.0001
rs2076756	1.615	0.129	156.633	1	<0.0001
rs10521209	2.044	0.100	414.381	1	<0.0001
rs2066845	0.944	0.445	4.496	1	0.034
rs13361189	1.644	0.265	38.513	1	<0.0001
rs1736135	2.090	0.102	419.740	1	<0.0001

SNP – single nucleotide polymorphism, df – degrees of freedom.

Table 13. *Statistics of the β-coefficients for each UC-associated SNP estimated by binary logistic regression analysis*

SNP	β-coefficient	Standart error	Wald	df	P-value
rs1736135	0.864	0.061	204.003	1	<0.0001
rs7746082	0.735	0.077	90.448	1	<0.0001
rs10758669	0.793	0.069	132.018	1	<0.0001
rs2872507	0.822	0.066	153.204	1	<0.0001
rs3806308	0.889	0.061	211.079	1	<0.0001
rs3024505	0.679	0.105	42.153	1	<0.0001
rs11209026	0.932	0.056	274.867	1	<0.0001
rs3197999	0.840	0.088	92.069	1	<0.0001
rs9268877	1.063	0.114	87.294	1	<0.0001
rs6426833	0.872	0.064	186.417	1	<0.0001
rs11190140	0.782	0.067	137.645	1	<0.0001
rs9268858	0.922	0.058	254.234	1	<0.0001
rs9858542	0.851	0.087	96.438	1	<0.0001

SNP – single nucleotide polymorphism, df – degrees of freedom.

Table 14. *List of network types by which members of interactive network are associated*

Network	Description	Weight
Physical interactions		***90.22%***
Bantscheff-Drewes-2007	Direct interaction; Pubmed 17721511; Tags: Epithelial Cells; Cultured Cells; Cell Line; Signal Transduction; Cancer	7.74%
Rachez-Freedman-1999	Direct interaction; Pubmed 10235266; Tags: Epithelial Cells; Transcription Factors; Cell Line; Cultured Cells; Cancer	6.2%
Wood	Direct interaction;Pubmed 10214908; Tags: Transcription Factors	5.85%
Green-Lorsch-2002	Direct interaction;Pubmed 12297040	4.98%
Gordon-Regnier-2000	Direct interaction;Pubmed 10747208; Tags: Muscle	4.86%
Lukas-Lukas-1999	Direct interaction; Pubmed 10548110; Tags: Cell Proliferation; Transcription Factors; Cell Line; Cultured Cells; Cancer	4.59%
Boyer-Momsen-1973	Direct interaction; Pubmed 4517936	3.57%
Chen-Karin-2001	Direct interaction; Pubmed 11719186; Tags: Cultured Cells; Cancer; Cell Line; Cell Signalling; Immune System; Transcription Factors	3.41%
Blagoev-Mann-2003	Direct interaction; Pubmed 12577067; Tags: Epithelial Cells; Cell Line; Cultured Cells; Signal Transduction; Cancer	2.54%
Vermeulen-Timmers-2007	Direct interaction; Pubmed 17884155; Tags: Epithelial Cells; Transcription Factors; Cell Line; Cultured Cells; Cancer	2.4%
Borman-Kean-2000	Direct interaction; Pubmed 11058101	2.13%
Koch-Hermeking-2007	Direct interaction; Pubmed 17314511; Tags: Epithelial Cells; Transcription Factors; Cell Line; Cultured Cells; Cancer	2.1%
Tomomori-Sato-Conaway-2004	Direct interaction; Pubmed 14638676; Tags: Cultured Cells; Cancer; Epithelial Cells; Cell Line; Liver; Transcription Factors	2.06%
Merrick-Anderson-1975	Direct interaction; Pubmed 1095581	1.91%
Miles-Elenitoba-Johnson-2005	Direct interaction; Pubmed 16147992; Tags: Transcription Factors; Cell Line; Cultured Cells; Cancer	1.89%
Cuello-Wieland-2003	Direct interaction; Pubmed 12486123; Tags: Brain; Nervous System	1.74%
Trachsel-Staehelin-1977	Direct interaction; Pubmed 592399	1.68%
Jacob-Luse-1991	Direct interaction; Pubmed 1939271; Tags: Epithelial Cells; Cell Line; Cultured Cells; Cancer	1.62%

Table 14 continued

Network	Description	Weight
Bernhard-Sheil-2004	Direct interaction; Pubmed 15047060; Tags: Cell Line; Cultured Cells; Cancer	1.57%
Le Hir-Moore-2001	Direct interaction; Pubmed 11532962; Tags: Transcription Factors	1.41%
Fierro-Monti-Roepstorff-2006	Direct interaction; Pubmed 16739988; Tags: Cultured Cells; Cell Line	1.4%
Czubaty-Staro#-2005	Direct interaction; Pubmed 15848144; Tags: Epithelial Cells; Cell Line; Cultured Cells; Cancer	1.38%
Hoshino-Katada-1999	Direct interaction; Pubmed 10358005	1.37%
Liu-Yen-2006	Direct interaction; Pubmed 17030981; Tags: Epithelial Cells; Transcription Factors; Cell Line; Cultured Cells; Cancer	1.15%
Foster-Klip-2006	Direct interaction; Pubmed 16396496; Tags: Cell Line; Cultured Cells; Stem Cells	1.05%
Sato-Conaway-2004	Direct interaction; Pubmed 15175163; Tags: Epithelial Cells; Transcription Factors; Cell Line; Cultured Cells; Cancer	1.02%
Squatrito-Draetta-2004	Direct interaction; Pubmed 15064750; Tags: Cultured Cells; Cancer; Cell Proliferation; Cell Line; Fibroblasts; Epithelial Cells; Stem Cells	0.95%
Haren-Merdes-2006	Direct interaction; Pubmed 16461362; Tags: Epithelial Cells; Cultured Cells; Cell Line; Nervous System; Cancer	0.9%
Dodson-Steiner-1998	Direct interaction; Pubmed 9631292	0.89%
Jin-Conaway-2005	Direct interaction; Pubmed 16230350; Tags: Epithelial Cells; Cultured Cells; Cell Line; Cancer	0.86%
Hakimi-Shiekhattar-2003	Direct interaction; Pubmed 12493763; Tags: Cultured Cells; Cancer; Epithelial Cells; Cell Line; Time Series; Transcription Factors	0.86%
Mayor-Nigg-2000	Direct interaction; Pubmed 11076968; Tags: Cultured Cells; Cancer; Cell Proliferation; Cell Line; Fibroblasts; Epithelial Cells; Stem Cells	0.84%
Schröder-Hasilik-2007	Direct interaction; Pubmed 17174955; Tags: Pregnancy	0.78%
Brajenovic-Drewes-2004	Direct interaction; Pubmed 1467619; Tags: Cultured Cells; Cell Line	0.64%
Frolova-Philippe-1994	Direct interaction; Pubmed 7990965	0.58%
Cai-Conaway-2005	Direct interaction; Pubmed 15647280; Tags: Epithelial Cells; Transcription Factors; Cell Line; Cultured Cells; Cancer	0.58%

Table 14 continued

Network	Description	Weight
Crockett-Lim-2004	Direct interaction; Pubmed 14968112; Tags: Cultured Cells; Cell Line; Signal Transduction; Cancer	0.57%
Barrios-Rodiles-Wrana-2005	Direct interaction; Pubmed 15761153; Tags: Cell Signalling; Cell Line; Cancer; Epithelial Cells; Signal Transduction; Cultured Cells; Transcription Factors	0.56%
Cottrell-Bredesen-2005	Direct interaction; Pubmed 16049941; Tags: Brain; Nervous System; Immune System	0.55%
Daulat-Jockers-2007	Direct interaction; Pubmed 17215244; Tags: Cultured Cells; Cell Line	0.52%
Sablina-Hahn-2007	Direct interaction; Pubmed 17540176; Tags: Cell Line; Cultured Cells; Cancer	0.48%
Conaway-Conaway-1988	Direct interaction; Pubmed 2449431; Tags: Liver	0.42%
Zhao-Moore-1999	Direct interaction; Pubmed 10357856	0.42%
Zhou-Conrads-2004	Direct interaction; Pubmed 15174051	0.42%
Wysocka-Herr-2003	Direct interaction; Pubmed 12670868; Tags: Cultured Cells; Cancer; Cell Proliferation; Cell Line; Epithelial Cells; Transcription Factors	0.41%
Higa-Zhang-2006	Direct interaction; Pubmed 17041588; Tags: Epithelial Cells; Transcription Factors; Cell Line; Cancer; Cultured Cells	0.4%
Cai-Conaway-2003	Direct interaction; Pubmed 12963728; Tags: Epithelial Cells; Transcription Factors; Cell Line; Cultured Cells; Cancer	0.4%
Thelemann-Haley-2005	Direct interaction; Pubmed 15657067; Tags: Cultured Cells; Cell Line; Signal Transduction; Cancer	0.39%
Cramer-Kornberg-2001	Direct interaction; Pubmed 11313498; Tags: Transcription Factors	0.34%
Goudreault-Gingras-2009	Direct interaction; Pubmed 18782753; Tags: Epithelial Cells; Cell Line; Cultured Cells; Nervous System; Cancer	0.34%
Budd-Campbell-2000	Direct interaction; Pubmed 10748138	0.31%
Price-2000	Direct interaction; Pubmed 10733565; Tags: Transcription Factors	0.28%
Jones-MacBeath-2006	Direct interaction; Pubmed 16273093; Tags: Cultured Cells; Cell Line	0.27%
Villacé-Ortín-2004	Direct interaction; Pubmed 15121898; Tags: Cultured Cells; Nervous System; Cell Line	0.25%
Ravasi-Hayashizaki-2010_human	Pubmed 20211142; Tags: Cell Proliferation; Transcription Factors	0.22%

Table 14 continued

Network	Description	Weight
Hinkley-Henry-2003	Direct interaction; Pubmed 12621023; Tags: Transcription Factors	0.22%
PATHWAYCOMMONS	Direct interaction	0.21%
Catimel-Nice	Direct interaction; Pubmed 16212417; Tags: Transcription Factors; Cell Line; Time Series; Cancer; Cultured Cells	0.2%
Litovchick-DeCaprio-2007	Direct interaction; Pubmed 17531812; Tags: Cell Proliferation; Transcription Factors; Nervous System; Cancer	0.19%
Mikula-Ostrowski-2006	Direct interaction; Pubmed 16518874; Tags: Cultured Cells; Signal Transduction; Cancer; Cell Line; Immune System; Transcription Factors	0.19%
Sowa-Harper-2009	Direct interaction; Pubmed 19615732	0.18%
Jin-Pawson-2004	Direct interaction; Pubmed 15324660; Tags: Cell Proliferation; Cultured Cells; Immune System	0.16%
Berggard-James-2006	Direct interaction; Pubmed 16512683; Tags: Brain; Nervous System	0.16%
BIOGRID	Direct interaction	0.14%
Wang-Balch-2006	Direct interaction; Pubmed 17110338	0.12%
Goumans-Benne-1980	Direct interaction; Pubmed 6901506	0.12%
Vertegaal-Lamond-2006	Direct interaction; Pubmed 17000644; Tags: Epithelial Cells; Cell Line; Cultured Cells; Cancer	0.11%
Falsone-Kungl-2005	Direct interaction; Pubmed 16263121; Tags: Epithelial Cells; Cell Line; Cultured Cells; Cancer	0.11%
Will-Lührmann-2004	Direct interaction; Pubmed 15146077; Tags: Epithelial Cells; Cell Line; Cultured Cells; Cancer	0.11%
Yamaguchi-Handa-1999	Direct interaction; Pubmed 10199401; Tags: Epithelial Cells; Transcription Factors; Cell Line; Cultured Cells; Cancer	0.1%
Melikyan-Cohen-2000	Direct interaction; Pubmed 11038187 Tags: Immune System	0.09%
Nakayama-Ohara-2002	Direct interaction; Pubmed 12421765 Tags: Cultured Cells; Cancer	0.08%
Sleeman-Lamond-1999	Direct interaction; Pubmed 10531003; Tags: Cultured Cells; Cancer; Localization; Epithelial Cells; Cell Line; Time Series	0.08%
Benzinger-Hermeking-200	Direct interaction; Pubmed 15778465; Tags: Cell Proliferation; Transcription Factors; Cultured Cells; Kidney; Cancer	0.08%

Table 14 continued

Network	Description	Weight
Goehler-Wanker-2004	Direct interaction; Pubmed 15383276 Tags: Fibroblasts; Cultured Cells; Nervous System; Cancer; Cell Line	0.06%
Camargo-Brandon-2007	Direct interaction; Pubmed 17043677 Tags: Cell Proliferation; Nervous System	0.05%
Lim-Zoghbi-2006	Direct interaction; Pubmed 16713569 Tags: Brain; Nervous System	0.05%
Rappsilber-Mann-2002	Direct interaction; Pubmed 12176931 Tags: Transcription Factors; Signal Transduction	0.05%
Ewing-Figeys-2007 A	Direct interaction; Pubmed 12176931 Tags: Transcription Factors; Signal Transduction	0.04%
Colland-Gauthier-2004	Direct interaction; Pubmed 15231748 Tags: Cell Signalling; Signal Transduction; Cancer; Cell Line; Cultured Cells; Transcription Factors; Kidney	0.04%
Babusiak-Vyoral-2005	Direct interaction; Pubmed 15627969 Tags: Cell Line; Cultured Cells; Cancer	0.04%
Mazumder-Fox-2003	Direct interaction; Pubmed 14567916; Tags: Cultured Cells; Cancer; Cell Line; Time Series; Cell Signalling; Immune System	0.04%
Bouwmeester-Superti-Fu	Direct interaction ;Pubmed 14743216; Tags: Cell Line; Cancer; Signal Transduction; Cultured Cells; Cell Signalling; Transcription Factors	0.02%
Shiio-Eisenman-2006	Direct interaction; Pubmed 16449650; Tags: Cultured Cells; Cancer; Epithelial Cells; Cell Line; Fibroblasts; Transcription Factors	0.02%
Vera-Jaumot-2007	Direct interaction; Pubmed 17309103 Tags: Transcription Factors	0.02%
Ewing-Figeys-2007 B	Direct interaction; One of 2 datasets produced from this publication; Pubmed 17353931	0.02%
Melki-Cowan-1993	Direct interaction; Pubmed 8104191 Tags: Nervous System	0.01%
Rual-Vidal-2005 A	Direct interaction; One of 2 datasets produced from this publication; Pubmed 16189514	0.01%
Zawel-Reinberg-1995	Direct interaction; Pubmed 7601352 Tags: Transcription Factors	0.01%
Co-expression		***3.5%***
Bild-Nevins-2006 B	Pearson correlation; One of 3 datasets produced from this publication; Pubmed 16273092 GEO GSE3143; Tags: Cultured Cells; Signal Transduction; Cancer; Epithelial Cells; Cell Line; Disease; Breast; Transcription Factors; Breast Cancer	0.27%

Table 14 continued

Network	Description	Weight
Wang-Maris-2006	Pearson correlation; Pubmed 16778177 GEO GSE3960; Tags: Transcription Factors; Cancer	0.26%
Hummel-Siebert-2006	Pearson correlation; Pubmed 16760442 GEO GSE4475; Tags: Cancer	0.25%
Rieger-Chu-2004	Pearson correlation; Pubmed 15096622 GEO GSE1725; Tags: Cultured Cells; Cell Line	0.25%
Ramaswamy-Golub-2001	Pearson correlation; Pubmed 11742071 Tags: Cancer	0.21%
Nakayama-Hasegawa-2007	Pearson correlation; Pubmed 17464315 GEO GSE6481; Tags: Transcription Factors; Cancer	0.21%
Jones-Libermann-2005	Pearson correlation Pubmed 16115910 GEO GSE15641 Tags: Transcription Factors; Disease; Cancer	0.19%
Perou-Botstein-2000	Pearson correlation; Pubmed 10963602 Tags: Cultured Cells; Breast Cancer; Cancer	0.18%
Hannenhalli-Cappola-2006	Pearson correlation; Pubmed 16952980 GEO GSE5406; Tags: Transcription Factors	0.17%
Agnelli-Neri-2007	Pearson correlation Pubmed 17367409 GEO GSE6401 Tags: Transcription Factors; Cancer; Localization	0.16%
Burczynski-Dorner-2006	Pearson correlation Pubmed 16436634 GEO GSE3365	0.16%
Tian-Shaughnessy-2003 B	Pearson correlation; One of 2 datasets produced from this publication; Pubmed 14695408 GEO GSE755; Tags: Transcription Factors; Signal Transduction; Cancer	0.15%
Chng-Fonseca-2007	Pearson correlation; Pubmed 17409404 GEO GSE6477; Tags: Cancer; Localization	0.15%
Ross-Brown-2000	Pearson correlation; Pubmed 10700174 Tags: Transcription Factors; Cultured Cells; Breast Cancer; Breast; Cancer	0.15%
Noble-Diehl-2008	Pearson correlation Pubmed 18523026 GEO GSE11223	0.14%
Chowdary-Mazumder-2006	Pearson correlation; Pubmed 16436632 GEO GSE3726; Tags: Breast Cancer; Cancer	0.14%
Ross-Perou-2001	Pearson correlation; Pubmed 11673656 Tags: Cell Line; Cultured Cells; Breast Cancer; Breast; Cancer	0.13%
Zangrando-Basso-2009	Pearson correlation Pubmed 19549311 GEO GSE14062	0.11%

Table 14 continued

Network	Description	Weight
Tian-Shaughnessy-2003 A	Pearson correlation; One of 2 datasets produced from this publication; Pubmed 14695408 GEO GSE754; Tags: Transcription Factors; Signal Transduction; Cancer	0.1%
Wong-Aronow-2007	Pearson correlation; Pubmed 17374846 GEO GSE4607; Tags: Immune System	0.1%
Pathway		***2.13%***
PATHWAYCOMMONS-CELL_MAP	Direct interaction PATHWAY COMMONS CELL_MAP	0.88%
PATHWAYCOMMONS-HUMANCYC	Direct interaction PATHWAY COMMONS HUMANCYC	0.5%
PATHWAYCOMMONS-REACTOME	Direct interaction PATHWAY COMMONS REACTOME	0.34%
PATHWAYCOMMONS-NCI_NATURE	Direct interaction PATHWAY COMMONS NCI_NATURE	0.33%
PATHWAYCOMMONS-IMID	Direct interaction PATHWAY COMMONS IMID	0.08%
Predicted		***2.08%***
I2D_vonMering-Bork-2002_High_Yeast2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using vonMering-Bork-2002 Saccharomyces cerevisiae data; Pubmed 12000970 I2D YeastHigh	0.27%
I2D_Li-Vidal-2004_interolog_Worm2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using Li-Vidal-2004 Caenorhabditis elegans data Pubmed 14704431 I2D INTEROLOG	0.22%
I2D_vonMering-Bork-2002_Medium_Yeast2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using vonMering-Bork-2002 Saccharomyces cerevisiae data; Pubmed 12000970 I2D YeastMedium	0.19%
I2D_BioGRID_Yeast2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using BioGRID; Saccharomyces cerevisiae data; I2D BioGRID_Yeast	0.19%
I2D_Krogan-Greenblatt-2006_Core_Yeast2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using Krogan-Greenblatt-2006 Saccharomyces cerevisiae data; Pubmed 16554755; I2D Krogan_Core	0.14%
I2D_small_scale	Direct interaction; I2D predictions combined small-scale datasets; I2D under_threshold	0.13%
I2D_IntAct_Worm2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using IntAct Caenorhabditis elegans data; I2D IntAct_Worm	0.1%

Table 14 continued

Network	Description	Weight
I2D_BIND_Mouse2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using BIND Mus musculus data; I2D BIND_Mouse	0.09%
I2D_MGI_Mouse2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using MGI Mus musculus data; I2D MGI	0.09%
I2D_IntAct_Yeast2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using IntAct Saccharomyces cerevisiae data; I2D IntAct_Yeast	0.08%
I2D_IntAct_Fly2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using IntAct Drosophila melanogaster data; I2D IntAct_Fly	0.08%
I2D_Ptacek-Snyder-2005_Yeast2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using Ptacek-Snyder-2005 Saccharomyces cerevisiae data; Pubmed 16319894 I2D Yeast_Kinome	0.06%
I2D_IntAct_Mouse2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using IntAct Mus musculus data; I2D IntAct_Mouse	0.06%
I2D_vonMering-Bork-2002_Low_Yeast2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using vonMering-Bork-2002 Saccharomyces cerevisiae data; Pubmed 12000970 I2D YeastLow	0.06%
Stuart-Kim-2003	Pubmed 12934013; Tags: Cell Proliferation; Cultured Cells; Signal Transduction; Cancer	0.06%
I2D_Krogan-Greenblatt-2006_NonCore_Yeast2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using Krogan-Greenblatt-2006 Saccharomyces cerevisiae data; Pubmed 16554755; I2D Krogan_NonCore	0.06%
I2D_Formstecher-Daviet-2005-Embryo_Fly2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using Formstecher-Daviet-2005 Drosophila melanogaster data; Pubmed 15710747; I2D FlyEmbryo; Tags: Cancer	0.05%
I2D_BIND_Rat2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using BIND Rattus norvegicus data; I2D BIND_Rat	

Table 14 continued

Network	Description	Weight
I2D_IntAct_Rat2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using IntAct Rattus norvegicus data; I2D IntAct_Rat	0.04%
I2D_BIND_Yeast2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using BIND Saccharomyces cerevisiae data; I2D BIND_Yeast	0.04%
I2D_Stanyon-Finley-2004-CellCycle_Fly2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using Stanyon-Finley-2004 Drosophila melanogaster data; Pubmed 15575970; I2D FlyCellCycle Tags: Transcription Factors	0.02%
I2D_MINT_Worm2Human	Direct interaction; I2D predictions of protein protein interactions for Homo sapiens using MINT Caenorhabditis elegans data; I2D MINT_Worm	0.01%
Co-localization		**1.86%**
Schadt-Shoemaker-2004	Pearson correlation; Predicted transcript array Pubmed 15461792	1.29%
Johnson-Shoemaker-2003	Pearson correlation; Pubmed 14684825 Tags: Cultured Cells; Cell Line	0.56%
Genetic interactions		0.21%
BIOGRID	Direct interaction	0.21%

Table 15. *List of the interacting genes and the type of their interactions*

Gene 1	Gene 2	Weight	Network group	Networks
VAV1	CD247	0.0264	Pathway	PATHWAYCOMMONS-NCI_NATURE
VAV1	ZAP70	0.0117	Co-expression	Nakayama-Hasegawa-2007 Bild-Nevins-2006 B
CBL	CSK	0.0238	Physical interactions	PATHWAYCOMMONS BIOGRID
FRK	CSK	0.0069	Co-expression	Hummel-Siebert-2006
LCK	CD3E	0.0191	Pathway	PATHWAYCOMMONS-NCI_NATURE
CD3E	VCP	0.0158	Co-expression	Zangrando-Basso-2009
LCK	VAV1	0.0225	Pathway	PATHWAYCOMMONS-NCI_NATURE
CBL	ZAP70	0.0309	Pathway	PATHWAYCOMMONS-NCI_NATURE
CSK	VCP	0.0194	Co-expression	Wong-Aronow-2007 Hummel-Siebert-2006
CSK	GRB2	0.0159	Co-expression	Ramaswamy-Golub-2001
ZAP70	CD3E	0.0359	Pathway	PATHWAYCOMMONS-NCI_NATURE
VAV1	PTPN22	0.0440	Physical interactions	PATHWAYCOMMONS
VAV1	CD3E	0.0281	Pathway	PATHWAYCOMMONS-NCI_NATURE

Table 15 continued

Gene 1	Gene 2	Weight	Network group	Networks
CBL	GRB2	0.1902	Predicted	I2D_IntAct_Mouse2Human I2D_BIND_Mouse2Human
CSK	CD247	0.0054	Co-expression	Bild-Nevins-2006 B
CD247	PTPN22	0.0592	Physical interactions	PATHWAYCOMMONS
CBL	ZAP70	0.0179	Physical interactions	PATHWAYCOMMONS BIOGRID
CBL	GRB2	0.0050	Physical interactions	PATHWAYCOMMONS BIOGRID
GRB2	PTPN22	0.1587	Physical interactions	PATHWAYCOMMONS BIOGRID
ZAP70	CSK	0.0050	Co-expression	Bild-Nevins-2006 B
CD3E	VCP	0.0144	Physical interactions	PATHWAYCOMMONS
CBL	CD247	0.0054	Physical interactions	PATHWAYCOMMONS
LCK	CSK	0.0373	Co-expression	Nakayama-Hasegawa-2007 Bild-Nevins-2006 B Noble-Diehl-2008
CBL	CSK	0.0312	Co-expression	Noble-Diehl-2008
CSK	PTPN22	0.0568	Physical interactions	PATHWAYCOMMONS
CSK	CD3E	0.0134	Co-expression	Nakayama-Hasegawa-2007 Bild-Nevins-2006 B
CSK	GRB2	0.0030	Physical interactions	PATHWAYCOMMONS
CSK	CD247	0.0099	Physical interactions	PATHWAYCOMMONS
ZAP70	PTPN22	0.0550	Physical interactions	PATHWAYCOMMONS
LCK	ZAP70	0.0567	Co-expression	Zangrando-Basso-2009 Wong-Aronow-2007 Bild-Nevins-2006 B Ramaswamy-Golub-2001 Noble-Diehl-2008
FRK	VCP	0.0045	Co-expression	Ramaswamy-Golub-2001
LCK	CBL	0.0115	Physical interactions	PATHWAYCOMMONS BIOGRID
CD247	CD3E	0.0136	Co-localization	Johnson-Shoemaker-2003
CBL	GRB2	0.0479	Pathway	PATHWAYCOMMONS-REACTOME PATHWAYCOMMONS-IMID
CBL	VAV1	0.0242	Pathway	PATHWAYCOMMONS-NCI_NATURE
FRK	GRB2	0.0138	Co-expression	Hummel-Siebert-2006
CBL	CD247	0.0193	Pathway	PATHWAYCOMMONS-NCI_NATURE

Table 15 continued

Gene 1	Gene 2	Weight	Network group	Networks
LCK	CSK	0.0049	Physical interactions	PATHWAYCOMMONS
CD247	GRB2	0.0031	Physical interactions	PATHWAYCOMMONS
VAV1	VCP	0.0130	Co-expression	Jones-Libermann-2005
CBL	VAV1	0.0129	Physical interactions	PATHWAYCOMMONS BIOGRID
CD3E	PTPN22	0.0678	Physical interactions	PATHWAYCOMMONS
ZAP70	CD247	0.0449	Co-expression	Zangrando-Basso-2009 Wong-Aronow-2007 Nakayama-Hasegawa-2007 Burczynski-Dorner-2006 Bild-Nevins-2006 B Wang-Maris-2006
VAV1	CD247	0.0149	Co-expression	Bild-Nevins-2006 B Jones-Libermann-2005 Noble-Diehl-2008
CSK	CD3E	0.0113	Physical interactions	PATHWAYCOMMONS
CD247	CD3E	0.0604	Co-expression	Wong-Aronow-2007 Nakayama-Hasegawa-2007 Hummel-Siebert-2006 Bild-Nevins-2006 B Jones-Libermann-2005 Ramaswamy-Golub-2001 Wang-Maris-2006
CBL	PTPN22	0.0309	Physical interactions	PATHWAYCOMMONS
CBL	CD3E	0.0206	Pathway	PATHWAYCOMMONS-NCI_NATURE
LCK	ZAP70	0.0208	Physical interactions	PATHWAYCOMMONS BIOGRID
FRK	CD3E	0.0471	Co-expression	Tian-Shaughnessy-2003 A Tian-Shaughnessy-2003 B
VCP	PTPN22	0.0723	Physical interactions	PATHWAYCOMMONS
CD247	CD3E	0.0118	Physical interactions	PATHWAYCOMMONS
ZAP70	CD3E	0.0379	Co-expression	Wong-Aronow-2007 Nakayama-Hasegawa-2007 Hummel-Siebert-2006 Burczynski-Dorner-2006 Bild-Nevins-2006 B Ramaswamy-Golub-2001
LCK	CBL	0.0164	Pathway	PATHWAYCOMMONS-NCI_NATURE
VAV1	CD3E	0.0066	Co-expression	Bild-Nevins-2006 B
FRK	C13orf 31	0.0340	Co-expression	Wong-Aronow-2007 Ramaswamy-Golub-2001

Table 15 continued

Gene 1	Gene 2	Weight	Network group	Networks
CD247	CD3E	0.1344	Pathway	PATHWAYCOMMONS-REACTOME PATHWAYCOMMONS-NCI_NATURE
LCK	CD247	0.1381	Co-expression	Jones-Libermann-2005 Wang-Maris-2006 Zangrando-Basso-2009 Tian-Shaughnessy-2003 B Chowdary-Mazumder-2006 Tian-Shaughnessy-2003 A Nakayama-Hasegawa-2007 Bild-Nevins-2006 B Ramaswamy-Golub-2001 Wong-Aronow-2007
VAV1	ZAP70	0.0423	Pathway	PATHWAYCOMMONS-NCI_NATURE
ZAP70	VCP	0.0117	Physical interactions	PATHWAYCOMMONS
VAV1	GRB2	0.0742	Predicted	I2D_IntAct_Mouse2Human
VAV1	VCP	0.0093	Physical interactions	PATHWAYCOMMONS
ZAP70	GRB2	0.0090	Physical interactions	PATHWAYCOMMONS BIOGRID
VAV1	GRB2	0.0065	Physical interactions	PATHWAYCOMMONS BIOGRID
ZAP70	CD247	0.0453	Physical interactions	PATHWAYCOMMONS BIOGRID
VAV1	ZAP70	0.0233	Physical interactions	PATHWAYCOMMONS BIOGRID
VAV1	CD247	0.0076	Physical interactions	PATHWAYCOMMONS
LCK	VCP	0.0063	Physical interactions	PATHWAYCOMMONS
LCK	GRB2	0.0015	Physical interactions	PATHWAYCOMMONS
FRK	CSK	0.0162	Physical interactions	Bantscheff-Drewes-2007
LCK	ZAP70	0.0287	Pathway	PATHWAYCOMMONS-NCI_NATURE
LCK	CD247	0.0051	Physical interactions	PATHWAYCOMMONS
ZAP70	CSK	0.0092	Physical interactions	PATHWAYCOMMONS
VAV1	CSK	0.0685	Co-expression	Tian-Shaughnessy-2003 A Chng-Fonseca-2007 Agnelli-Neri-2007 Ramaswamy-Golub-2001 Tian-Shaughnessy-2003 B

Table 15 continued

Gene 1	Gene 2	Weight	Network group	Networks
LCK	CD3E	0.0707	Co-expression	Wong-Aronow-2007 Nakayama-Hasegawa-2007 Burczynski-Dorner-2006 Bild-Nevins-2006 B Jones-Libermann-2005 Ramaswamy-Golub-2001 Wang-Maris-2006
LCK	VAV1	0.0038	Physical interactions	PATHWAYCOMMONS
LCK	GRB2	0.0098	Pathway	PATHWAYCOMMONS-NCI_NATURE
LCK	CD247	0.1669	Pathway	PATHWAYCOMMONS-REACTOME PATHWAYCOMMONS-NCI_NATURE
ZAP70	CD3E	0.0563	Physical interactions	PATHWAYCOMMONS BIOGRID
VAV1	CD3E	0.0088	Physical interactions	PATHWAYCOMMONS
LCK	CD3E	0.0367	Physical interactions	PATHWAYCOMMONS BIOGRID
LCK	PTPN22	0.0296	Physical interactions	PATHWAYCOMMONS
VAV1	GRB2	0.0599	Pathway	PATHWAYCOMMONS-NCI_NATURE PATHWAYCOMMONS-IMID
ZAP70	CD247	0.2252	Pathway	PATHWAYCOMMONS-REACTOME PATHWAYCOMMONS-NCI_NATURE

Table 16. *Ontology categories*

Category	Q-value
T cell receptor complex	6.4E-4
positive regulation of immune system process	1.1E-2
positive regulation of T cell activation	1.1E-2
positive regulation of lymphocyte activation	1.35E-2
protein domain specific binding	1.35E-2
positive regulation of cell activation	1.42E-2
positive regulation of leukocyte activation	1.42E-2
regulation of T cell activation	1.42E-2
receptor complex	1.77E-2
regulation of immune system process	1.77E-2
protein tyrosine kinase activity	1.88E-2
regulation of lymphocyte activation	1.88E-2
regulation of leukocyte activation	2.4E-2
regulation of cell activation	2.98E-2
T cell activation	3.05E-2
T cell receptor signaling pathway	4.63E-2