POPULATION GENOMICS OF CROHN DISEASE SUSCEPTIBILITY

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Biotechnology

by
Bengisu NARLİ

June 2024
İZMİR

We approve the thesis of **Bengisu NARLİ**

Examining Committee Members:

Prof. Dr. Efe SEZGİN
Department of Food Engineering, İzmir Institute of Technology

Dr. Öğr. Üyesi Mehmet GÖRGÜLÜ
Department of Dentistry, Altınbaş University

Doç Dr. Şükrü GÜLEÇ
Department of Food Engineering, İzmir Institute of Technology

27 June 2024

Prof. Dr. Efe SEZGİN
Supervisor, Department of Food Engineering İzmir Institute of Technology

Assoc. Prof. Dr. Ali Oğuz BÜYÜKKİLECİ
Co-Supervisor, Department of Food Engineering İzmir Institute of Technology

Assoc. Prof. Dr. Ali Oğuz BÜYÜKKİLECİ
Head of the Department of Biotechnology

Prof. Dr. Mehtap EANES
Dean of the Graduate School Engineering and Science

ACKNOWLEDGMENTS

I appreciate to my valuable advisor and teacher, Efe SEZGİN, who was always by my side with his knowledge, experience and expertise throughout my master's education and thesis. During this process, my valuable teacher's positive energy, always helping me with my problems and motivation helped me complete my study. Thank you very much for being my thesis advisor.

I would like to thank my beloved family for being with me throughout my life and for their endless support and efforts under all circumstances. And also, I thankful to my boyfriend very much, he always motivated me throughout my thesis process and supported me with his tolerance and understanding.

ABSTRACT

POPULATION GENOMICS OF CROHN DISEASE SUSCEPTIBILITY

Crohn's Disease (CD) is an inflammatory bowel disease that causes chronic inflammation of the gastrointestinal tract. It is argued that the genetic makeup that increases susceptibility to CD in modern populations is a derived trait that confers selective advantage against certain environmental stressors, such as resistance to infections and pathogens. Therefore, there should be signs of selection on CD-associated genes. To test whether CD risk is a derived selected trait, CD-associated genes and variants were identified through literature searches. Ancient and modern population data were collected through the Allen database and 1000 Genomes Project. Data were analyzed using Plink and R. 352 CD risk alleles were identified and the disease risk status (protective and susceptible) of CD-associated alleles was compared to their ancestral and derived status. The differentiation of four metapopulations worldwide according to these alleles was examined. Furthermore, differences in allele frequency between ancient and modern populations were compared. It was observed that the genetic differentiation and segregation between modern metapopulations is due to the influence of CD-related genes *PUS10* and *PPBP_CXCL5*, while the segregation between ancient and modern metapopulations is due to the influence of *PPP5C*, *PPBP_CXCL5* and *AIMP1P2* genes. Populations with higher prevalence of CD were found to have higher risk allele frequencies. Variants in CD-related *IRGM*, *OR2B11* and *IL10* genes showed high allele frequency changes over time when comparing ancestral and modern European populations. Recent selection analyses indicated possible positive selection acting on the CD-related genes *HERC2, MACROD2, RBFOX1*, *ITLN1* and *RNFT1P2*.

ÖZET

CROHN HASTALIĞI DUYARLILIĞININ POPÜLASYON GENOMİĞİ

Crohn Hastalığı (CD), gastrointestinal sistemin kronik inflamasyonuna neden olan inflamatuar bir bağırsak hastalığıdır. Modern popülasyonlarda CD'ye yatkınlığı artıran genetik yapının, enfeksiyonlara ve patojenlere direnç gibi belirli çevresel stres faktörlerine karşı seçici avantaj sağlayan türetilmiş bir özellik olduğu iddia edilmektedir. Bu nedenle, CD ile ilişkili genlerde seçilim belirtileri olmalıdır. CD riskinin türetilmiş seçilmiş bir özellik olup olmadığını test etmek için, literatür taramaları yoluyla CD ile ilişkili genler ve varyantlar belirlenmiştir. Allen veriseti ve 1000 Genom Projesi aracılığıyla eski ve modern nüfus verileri toplanmıştır. Veriler Plink ve R kullanılarak analiz edilmiştir. 352 CD risk aleli tanımlanmıştır ve CD ile ilişkili alellerin hastalık risk durumu (koruyucu ve duyarlı) atasal ve türetilmiş durumlarıyla karşılaştırılmıştır. Dünya çapındaki dört metapopülasyonun bu alellere göre farklılaşması incelenmiştir. Ayrıca, eski ve modern popülasyonlar arasındaki alel frekansı farklılıkları karşılaştırılmıştır. Günümüz metapopülasyonları arasındaki genetik farklılaşma ve ayrımın, CD ile ilişkili *PUS10* ve *PPBP_CXCL5* genlerinin etkisine bağlı olduğu, antik ve modern metapopülasyonlardaki ayrımın ise *PPP5C, PPBP_CXCL5* ve *AIMP1P2* genlerinin etkisi olduğu gözlemlenmiştir. CD prevalansının daha yüksek olduğu popülasyonlarda daha yüksek risk alel frekanslarına sahip olduğu belirlenmiştir. CD ile ilişkili *IRGM, OR2B11* ve *IL10* genlerindeki varyantlar, atasal ve modern Avrupa popülasyonları karşılaştırıldığında zaman içinde yüksek alel frekansı değişiklikleri göstermiştir. Gen bazlı son seleksiyon analizleri, CD ile ilişkili *HERC2, MACROD2, RBFOX1, ITLN1* ve *RNFT1P2* genlerinde etkili olan olası pozitif seçilime işaret etmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

Figure **Page**

Figure **Page**

Figure **Page**

LIST OF TABLES

Table **Page**

LIST OF SYMBOLS AND ABBREVIATIONS

ACB	African Caribbeans in Barbados
aDNA	Ancient DNA
AFR	African
AIEC	Adherent-Invasive *Escherichia coli*
AMR	Ad Mixed American
ASW	Americans of African Ancestry in SW USA
BEB	Bengali from Bangladesh
CD	Crohn's Disease (CD)
CDX	Chinese Dai in Xishuangbanna, China
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry
CHB	Han Chinese in Bejing, China
CHS	Southern Han Chinese
CLM	Colombians from Medellin, Colombia
EAS	East Asian
EHH	Extended Haplotype Homozygosity
EIM	Extraintestinal Manifestations
ESN	Esan in Nigeria
EUR	European
FIN	Finnish in Finland
Fst	Fixation index
GBR	British in England and Scotland
GI	Gastrointestinal
GIH	Gujarati Indian from Houston, Texas
GO	Gene Ontology
GWAS	Genome Wide Association Studies
GWD	Gambian in Western Divisions in the Gambia
IBD	Inflammatory Bowel Disease
IBS	Iberian Population in Spain
IHS	integrated haplotype score
ITU	Indian Telugu from the UK

JPT	Japanese in Tokyo, Japan
KHV	Kinh in Ho Chi Minh City, Vietnam
LD	Linkage Disequilibrium
LWK	Luhya in Webuye, Kenya
MSL	Mende in Sierra Leone
mtDNA	mitochondrial DNA
MXL	Mexican Ancestry from Los Angeles USA
NSAIDs	Non-Steroidal Anti-Inflammatory Drugs
PCR	polymerase chain reaction
PEL	Peruvians from Lima, Peru
PJL	Punjabi from Lahore, Pakistan
PUR	Puerto Ricans from Puerto Rico
SAS	South Asian
SNPs	Single Nucleotide Polymorphisms
STU	Sri Lankan Tamil from the UK
TB	Tuberculosis
TSI	Toscani in Italia
UC	Ulcerative Colitis
YRI	Yoruba in Ibadan, Nigeria

CHAPTER 1

INTRODUCTION

1.1 What is Crohn’s Disease?

Inflammatory bowel disease (IBD) is a life-long disease that involves chronic inflammation of the gastrointestinal (GI) tract. It results from the interaction of an individual's immune responses, genetic predisposition, and environmental factors. Inflammatory bowel disease includes types of Crohn's Disease (CD) and Ulcerative Colitis (UC).[1] Crohn's Disease was first described as a Regional Ileitis in 1932 by Burrill Crohn, Leon Ginzberg and Gordon D. Oppenheimer. By collecting data on 14 patients with symptoms of diarrhea, fever, abdominal cramps and weight loss, these doctors studied the pathological and clinical details of the disease and determined the existence of Crohn's Disease as a new disease unlike any previous disease.[2] Crohn's disease is a destructive and progressive IBD characterized by chronic inflammation of any part of the GI tract from the mouth to the anus, with a rapidly increasing worldwide incidence.[3,4] Although the cause of CD is still unknown, weak immune system, altered microbiota, genetic predisposition and environmental factors play an effective role in the course of the disease.[4] Crohn's Disease is a complex disease, so the genetic background is very important because multi-allelic, multi-gene disease, and also epidemiology and clinical manifestations of CD vary significantly between ethnic groups and geographical regions.

1.2 Epidemiology: Incidence and Prevalence

The epidemiology of Crohn's Disease varies greatly depending on geographic region, environmental conditions, ethnic groups, and immigrant populations.[3] Although there are regional differences, the prevalence and incidence rates of CD are generally

increasing steadily worldwide.[3,5] In particular, the highest rates are in Europe and North America, while it remains rare in Africa and South America.[3,6] The incidence of Crohn's Disease is highest in North America (Canada), Northern Europe, New Zealand and Australia. The highest prevalence of Crohn's Disease is seen in Europe (Germany), Canada and America.[5,6] It is observed that the incidence and prevalence of Crohn's Disease is increasing in today's populations, especially in Europe and America, as well as in Asia.[7] In addition to genetic predisposition being an important factor in this increase, changing environmental conditions and habits also act an important role.

1.3 Clinical Features and Natural History of Crohn's Disease

Crohn's Disease is a chronic intestinal inflammatory disease that can be seen in any area throughout the digestive system and most commonly affects the terminal ileum, cecum, perianal region and colon.[1] Clinical diagnosis of Crohn's Disease requires a variety of data such as the patient's history, physical examination, laboratory tests, endoscopy results, pathology findings and radiographic analyses.[8] What makes diagnosis easier is the presence of chronic intestinal inflammation, and the presence of Crohn's Disease can be checked and confirmed with other test results.

Crohn's Disease varies depending on which area or areas of the gastrointestinal tract the inflammation is located in. When the distribution of the disease according to the relevant regions is examined, only ileitis is seen in 25% of the patients, only colitis is seen in 25%, and ileocolitis is seen in 50%. Additionally, approximately one-third of the patients have perianal involvement, and 5-15% have oral or gastroduodenal region involvement.[3] Depending on this condition, there are different varieties of Crohn's Disease and these types may show different symptoms and complications.

The most common type of Crohn's Disease is Ileocolitis. Ileocolitis is inflammation of the last part of the small intestine (ileum) and large intestine (colon). The symptoms that occur in this type of inflammation are diarrhea, weight loss, cramping or pain in the middle or lower abdomen. Ileitis is inflammation of only the last part of the small intestine (ileum). Ileitis shows similar symptoms to ileocolitis. Jejunoileitis, inflammation of the middle part of the small intestine (jejunum), can cause vomiting, abdominal pain, cramps and diarrhea in the individual. Long-term inflammation may

cause fistula formation in the jejunum.[9] Gastroduodenal Crohn's is inflammation that affects the stomach, esophagus, and the first part of the small intestine, the duodenum. Symptoms of this type may include nausea, vomiting, loss of appetite, and weight loss.[10] Also, Gastroduodenal CD may cause obstruction, fistula formation or biliary obstruction.[11,12] Crohn's Colitis, or Granulomatous Colitis, causes inflammation only in the large intestine. Symptoms of Crohn's colitis include diarrhea, rectal bleeding, joint pain and skin lesions.[13] Additionally, fistula, ulcer and abscess may develop as a result of this inflammation.[14]

Although symptoms vary in different types of CD, in general, symptoms such as abdominal pain, weight loss, diarrhea and hematochezia, fever are common among types of CD.[15] In addition to these symptoms, patients also present with fistulas or perianal findings such as ulcers, abscesses, and fissures at the time of diagnosis.[4,16]

Extraintestinal Manifestations (EIM) of Crohn’s Disease are likely to be seen in nearly half of the patients; Formations such as both axial and peripheral, pyoderma gangrenosum and erythema nodosum, uveitis, scleritis, iritis and episcleritis, primary sclerosing cholangitis can be observed outside the gastrointestinal system, and these symptoms can affect many body systems, including the mouth, eye, musculoskeletal system, and hepatobiliary systems.[3,4,8,17]

The different behavior or phenotypes of CD that exist; stricturing disease due to fibrosis, stricturing and/or penetrating disease, penetrating disease due to fistulas between the gut and other structures, and lastly inflammatory or non-stricturing, non-penetrating disease.[4] Changes that may occur in clinical symptoms during the course of Crohn’s Disease, for example, recurrence of inflammation and/or variation in its severity, may cause replace in the current disease phenotype.

1.4 Pathogenesis and Risk Factors of Crohn’s Disease

Crohn’s Disease, that is, inflammation occurring in the gastrointestinal system, is a disease that results from interactions between genetic predisposition, environmental factors and intestinal microflora. CD can lead to the formation of an irregular and unhealthy intestinal microbiota by deteriorating the function of the intestinal barrier and the emergence of various immune responses.[4,18]

1.4.1 Non-Genetic Factors

Non-genetic factors (environmental factors and intestinal microflora) that affect the pathophysiology of Crohn's Disease also play a major role in the development of the disease. Environmental factors in Crohn's Disease include smoking, vitamin D deficiency, oral contraceptive use, antibiotic use, regular use of non-steroidal anti-inflammatory drugs, poor hygiene and urban environment increase the risk of the disease; Having more than two siblings, exposure to pets and farm animals, physical activity, breastfeeding, sharing a bedroom, fruit consumption and high fiber intake are among the factors associated with a reduced risk of disease.[1,7,17,19,20]

1.4.1.1 Environmental Factors

Environmental factors have a significant impact on the pathogenesis of Crohn's Disease. Numerous studies have examined the link between various environmental factors and the development of CD.[1] Smoking is a known environmental risk factor for CD and is associated with a twofold increased risk of CD, including passive smoking and early life exposure (OR 1.76, 95% CI 1.40–2.22).[21] Smoking has also been associated with earlier disease recurrence, greater need for surgery, need for immunosuppression (more frequent immunosuppression), and higher postoperative disease recurrence.[18,21,22] In addition, it has been observed that smoking changes smooth muscle tone, affects intestinal mucosal integrity, causes oxidative stress and affects the intestinal microbiota.[23,24] Several meta-analyses have identified a difference in the effect of smoking on CD risk among different ethnicities.[25] In an Israeli study on the relationship between smoking and Ulcerative Colitis, smoking cessation was associated with an increased risk of Ulcerative Colitis but not with an increased risk of Crohn's Disease.[26] Also, the incidence of CD is quite low in Asia and Africa despite high smoking rates, while the incidence of Crohn's Disease is quite high in Northern European countries despite low smoking rates.[27,28] Considering the results of some analyses, the causal relationship remains to be proven.

Medications such as antibiotics, NSAIDs (non-steroidal anti-inflammatory drugs), oral contraceptives, and aspirin have been suggested as potential risk factors in Crohn's Disease.[29] The gut microbiota is unbalanced and diverse during childhood, the first years of life, which can affect the gut immune response and cause a variety of abnormal inflammatory responses. It is thought that more than one factor is responsible for triggering intestinal dysbiosis.[30,31] Exposure to antibiotics, especially in childhood, is associated with this condition and increases the risk of CD (OR 1.74; 95% CI 1.35–2.23).[25,32] Other drugs potentially associated with increased risk include oral contraceptives, aspirin, and NSAIDs, while statins have been associated with reduced risk, especially in older people.[33] In addition to these valid data, antibiotic exposure has been shown to have a protective association with Crohn's Disease in a large population-based study from Asia.[34] There are studies on the effect of infections on CD, and although supporting data are less, systemic infections may trigger relapse in people with established disease.[35]

Appendectomy, that is, surgical removal of the appendix, is associated with the risk of CD, although its relationship with the development of CD is an area of research that has been little researched.[36] A study on Crohn's Disease suggested that Crohn's Disease may be diagnosed later in those who have previously had an appendectomy.[36,37] One study, in a large cohort of 212,963 patients who had an appendectomy before age 50, found an increased risk of Crohn's Disease for up to 20 years after appendectomy; However, it was observed that the risk of Crohn's Disease decreased when patients were operated on under the age of 10.[38] Appendectomy for perforating appendicitis has been associated with an increased risk of bowel resection, whereas appendectomy for other causes has been associated with a decreased risk of Crohn's Disease.[38]

Vitamin D and Micronutrients (Zinc and Iron) have been suggested to have an effect on Crohn's Disease.[7] In a study conducted to investigate the effects of vitamin D on CD, it was observed that women with sufficient and high levels of vitamin D had a significantly lower risk of CD than those with low and insufficient vitamin D levels (OR 0.54, 95% CI 0.30–0.99).[39] In animal studies, it has been observed that vitamin D deficiency or knockout of the vitamin D receptor may be associated with an increased risk of disease by causing inflammation.[40–42] Low vitamin D levels (≤20 ng/ml) increase the risk of progression to CD, and normalization of levels is associated with reduced risk.[43] Micronutrients such as zinc and iron have many effects on the immune system. In Crohn's Disease, zinc is important for autophagy and bacterial clearance, reduces

intestinal permeability, and there are studies suggesting that it reduces the likelihood of relapse.[44,45] Studies have suggested that dietary iron may cause colonic inflammation and also, there may be a relationship between high iron content in drinking water and increased disease risk.[46]

The three most important lifestyles that have an impact on Crohn's Disease include sleep, stress and exercise. Sleep deprivation and dysregulation are more common in CD and have been associated with active disease.[47] This relationship may cause active disease to cause sleep disturbances and exacerbate inflammation in poor sleep.[3,48,49] A study involving 136 Japanese patients found that sleep disturbance was a potential risk factor for exacerbations of CD within one year, while another study of a similar type observed that 3173 patients with sleep disturbances had an increased risk of disease exacerbations in CD within 6 months.[47,49] Studies with large samples have shown an association between major life stressors, anxiety, and depression, and increased risk of CD.[50] Stress can affect intestinal inflammation through various mechanisms, especially the autonomic nervous system; By affecting the production of pro-inflammatory cytokines and the activation of macrophages, it can cause changes in intestinal permeability and intestinal microbiota.[51] In CD patients, depression or anxiety is associated with disease relapse, surgery, decreased responsiveness to immunosuppressive therapy, and impaired quality of life.[50,52] Engaging in physical activity has been linked to a reduced risk of Crohn's Disease. A study conducted in Germany observed that people who do heavy work, such as construction, cleaning, etc., have a lower risk of developing CD than those who work in sedentary professions, such as desk jobs, mechanics, etc.[53] This study was supported by a prospective sample study showing a 44% reduction in CD risk in the cohort performing intense physical activity.[53,54]

1.4.1.2 Intestinal Microflora and Diet

In Crohn's Disease, diversity in the intestinal microbiota and diet are important factors associated with the disease. Any change in the structure of the healthy microbiota in the intestine can cause unusual inflammatory responses.[1] The occurrence of intestinal dysbiosis as a result of various factors is a feature of CD, and diet is the most likely environmental factor affecting the intestinal microbiota. Dysbiosis in Crohn's patients is

seen as a decrease in *Firmicutes* and *Bacteroides* bacteria and an increase in *Gammaproteobacteria* and *Actinobacteria.*[55] In a study by Darfeuille-Michaud et al., mucosa-associated adherent-invasive *Escherichia coli* (AIEC) was observed to be prevalent in approximately one-third of patients with Crohn's Disease.[56,57] It has also been suggested that *Faecalibacterium prausnitzii*, which has anti-inflammatory properties, provides protection against such diseases.[58–60]

Diet affects the structure of the intestinal microbiota and the metabolic activity of the intestine. Changes in microbial composition may alter the impact of diet on disease risk.[3,61] In particular, the host-gut microbiota relationship has changed with changes in food composition. Depending on the eating habits of individuals, the effect of different nutritional models is associated with the intestinal microbiota. For example, long-term consumption of a Western-style diet rich in animal origin and saturated fat has been associated with enterotypes such as *Bacteroides*, and long-term consumption of a diet rich in carbohydrates and fiber has been associated with enterotypes such as *Prevotella.*[1,62] Studies have shown that there is an inverse relationship between dietary fiber intake and the risk of CD. In particular, fiber from vegetables and fruits has been associated with a lower risk of CD (OR 0.59, 95% CI 0.39–0.90).[63,64]

In a study conducted in India, it was observed that *Proteobacteria* and *Bacteriodetes* phyla were dominant in the intestinal microbiome of subjects living in rural areas, and *Firmicutes* and *Lactobacillus* phyla were dominant in subjects living in urban areas.[65] Additionally, in a study conducted in South Africa, different gut microbiome composition was observed in genetically similar populations. In this observation, rural subjects contained significantly less *Bacteriodetes* than semi-urban and urban subjects.[66] These findings suggest that exposure to different environmental conditions and lifestyles can affect microbiome composition in a similar population.[1]

1.4.2 Genetic Factors

Looking at genetic, immunological and epidemiological data, Crohn's Disease is a heterogeneous disease that interacts with genetic and environmental factors.[7] Genetic predisposition plays an important role in determining disease risk, and much research is still ongoing to identify the genomic effect. There are many genes and loci associated

with the diagnosis of Crohn's Disease, and these genes are associated with innate pattern genes; Innate pattern recognition receptors are involved in epithelial barrier homeostasis and maintenance of epithelial barrier integrity, autophagy and lymphocyte differentiation. The most strongly duplicated genes associated with CD in studies conducted so far are; *NOD2* (*CARD15*), *IL23R* and *ATG16L1*.[3,4,67]

1.4.2.1 Familial Inheritance

In studies aimed at understanding the role of genetics in the pathogenesis of Crohn's Disease, it has been obtained from familial aggregation and twin studies that the hereditary component is an important factor.[68–70] While approximately 15% of individuals diagnosed with Crohn's Disease mention a familial connection to CD, only a minority of these patients have relatives with Ulcerative Colitis.[68,69,71] The relative risk of first-degree relatives of CD patients developing Inflammatory Bowel Disease is estimated to be around 5% in non-Jewish populations and 8% in Jewish populations.[69] When both parents are affected, the likelihood of the child developing the disease is estimated to be one in three.[68]

Twin studies indicate a substantial hereditary factor in Crohn's Disease. Using the twin design, researchers assume that the environmental influence on phenotypic variation is consistent between dizygotic (DZ) and monozygotic (MZ) twins. The disparity in disease concordance rates between MZ and DZ twin pairs can help estimate the additive genetic, unique environmental, and shared environmental elements of disease risk.[72] While the concordance rate for Crohn's Disease is 20-50% in monozygotic twins, this rate is 10% in dizygotic twins. However, although the effect of genetic predisposition between twins is known, some differences may develop in the natural course of the disease as a result of being affected by environmental factors. For example, a twin who smokes may develop Crohn's Disease, while a twin who does not smoke may develop Ulcerative Colitis.[68,73–75] Meta-analysis of six twin studies with a combined set of 196 DZ and 112 MZ twin pairs yielded concordance rates of 3.6% and 30.3%, respectively, suggesting a large role for genetics in Crohn's Disease risk.[72] Family and twin studies have motivated genome-wide analyses that will contribute to a better understanding of Crohn's Disease.

1.4.2.2 Disease Related Genes and Variances in Loci

There are many genes and loci associated with Crohn's Disease, but studies so far have identified strong associations with *NOD2*, the autophagy gene *ATG16L1*, and the IL-23 receptor gene *IL23R*.[67] Many of the CD risk loci may individually increase disease risk, and these variants are often found in regulatory regions of the genome.[76] Studies examining the relationship between the disease and genetic loci first showed that there was a relationship with a locus on chromosome 16.[77] In ongoing studies, this locus has been characterized as the *NOD2* locus, with three common variants affecting susceptibility to CD.[78,79] Following the discovery of coding variation in the intracellular pattern recognition receptor gene *NOD2* in 2001, more than 150 risk loci associated with CD risk have been identified through international collaborative GWAS.[76,77,80] The *NOD2* gene has been associated with fibrostenotic disease, ileal involvement, early age of onset, and family history of CD. While the increased risk of CD in patients carrying the heterogeneous *NOD2* locus is 2-4 times, it has been observed that the risk of developing CD is 20-40 times in those with homogeneous alleles.[78,79,81] The *IL23R* gene is involved in the development of Th17 lymphocytes and leads to alteration of cytokine production, which is involved in the development of Crohn's Disease.[82,83] *ILC3* and *ILC1* are known to play a role in the development of Crohn's Disease. Innate lymphoid cells (ILCs), a heterogeneous cell population, are critical in maintaining barrier integrity. They respond to microbial cues and dietary input by producing cytokines such as TNFα, interleukin 17, interleukin 22, and interferon γ.[84,85] As a result of analysis, ILCs isolated from the inflamed colon of patients with Crohn's Disease show increased gene expression of key *ILC3* cytokines, transcription factors (*RORC*) and cytokine receptors (*IL23R*).[86]

Despite the diversity in the effects of CD-related genes on the immune system in general, most genes can be broadly divided into those that affect innate immune responses, autophagy, maintenance of the integrity of the epithelial barrier, adaptive immune responses, repair and injury repair, response to oxidative stress, and microbial defense and antimicrobial activity.[67,87] Immune pathways are dysregulated in Crohn's Disease, and disruption of a site on the intestinal epithelium by emulsifiers commonly found in the western diet or by mutations in the *MUC2* gene may promote bacterial translocation.[88,89] Various risk loci may affect immunological function in a single pathway; for example, *ATG16L1*, *NOD2*, *IRGM*, *LRRK2* all have an effect on autophagy.

1.5 Candidate Gene Studies and Genome-Wide Association Study (GWAS)

In candidate gene studies, genes are selected because they contain basic information about biological function or are located within a region revealed through linkage analysis. Candidate gene studies involve genotyping markers within a gene of interest in a sample of controls and disease cases and detecting statistically significant differences in allele frequencies between the two groups.[90] The majority of the results obtained from candidate gene studies related to Crohn's Disease did not produce the expected results and these studies were not repeated in subsequent studies.[91,92] Factors such as false positive associations, small sample sizes, publication bias, and lack of reliability or omission of multiple comparisons mean that findings from candidate gene studies can be inaccurate and unreliable.[93,94] Consequently, a combination of technological advances and much larger samples will be required to identify accurate risk loci.[90]

As the Human Genome Project began to be completed, efforts to measure the extent of human genetic diversity at the population level were ongoing.[90] Thanks to projects such as the SNP Consortium and dbSNP, more than 1.4 million Single Nucleotide Polymorphisms (SNPs) have been documented.[95,96] In ongoing research, an additional 3.1 million SNPs were identified in 270 individuals from three different ancestral groups through the International Hapmap Project.[97] Simultaneously, advances in microarray technologies have facilitated cost-effective genotyping of hundreds of thousands of SNPs across the genome.[98] Linkage Disequilibrium (LD), specifically nonrandom associations of alleles situated at distinct loci, enables direct genotyping of pertinent sequences representing only a fraction of the total variants in the genome, thus allowing for effective exploration of the majority of genetic variations prevalent in a population.[90] In both East Asians and Europeans, around five million common SNPs with a minor allele frequency exceeding 5% can be identified through the tagging of approximately 500,000 SNPs.[97,99] These advances enabled Genome-Wide Association Studies (GWAS) to understand complex traits or identify loci associated with disease risk. GWAS are the search for statistically significant differences in allele or genotype frequencies between large numbers of diseased individuals and population controls at hundreds of thousands of SNPs generally spread across the genome. In these association

studies, SNPs that show a significant association with the disease point to regions of the genome that are likely to contain disease-related genes.[90] The first GWAS study of Crohn's Disease was conducted in a Japanese population in 2005 and identified *TNFSF15* as a disease-associated gene.[100] Ongoing studies each include approximately 500-2000 Crohn's Disease cases and a similar number of controls genotyped at 100,000-600,000 SNPs.[101,102] Genome-wide association studies have shown that Ashkenazi Jews have a three to four times higher risk of disease than the non-Jewish population; African-American and Asian groups have been observed to be associated with the lowest risk of disease.[7,103] Multiple CD risk loci with small individual contributions appear to correlate better across ethnicities.[101] This transethnic association study showed that most risk loci are shared between different ancestry groups, with a small number affecting population specificity regarding risk allele frequency (*NOD2*) or effect size (*TNFSF15* and *TNFSF8*).[104]

Genes and pathways identified through GWAS have contributed to the understanding of the biological processes underlying CD. Associations in *IRGM* and *ATG16L1* first suggested a role for autophagy in disease pathogenesis.[87,101,105] Genes involved in both innate immune system pathways, such as *IL23R, NOD2, STAT3, TLR4,* and acquired immune system pathways, such as *HLA, PTPN22, TNFSF15, IRF5* genes, have also been better understood thanks to these studies.[106] GWAS have also contributed to understanding the genetic overlap of Crohn's Disease with other immune-related diseases. Approximately 30% of the associated variants in these studies are shared with Ulcerative Colitis, while close to 50% of the loci are shared with at least one other immune-mediated disease such as Type 1 Diabetes, Celiac Disease, or Rheumatoid Arthritis.[107] Unlike most of these diseases, genes in the Human Leukocyte Antigen (*HLA*) region appear to provide a small effect on the risk of Crohn's Disease (OR 1.1-1.2).[90] In the results obtained from a study conducted by Jostins et al., the role of non-coding variations in disease risk was addressed. Many of these variants have been associated with affecting gene regulation. Only nine of the disease-associated loci (*NOD2, IL23R, FUT2, ADAM30, MUC19, GPR35, CD6, GPR65, ZNF831*) contain variants in the coding regions of the genes, and 13 have variants spanning both coding and non-coding regions (*ATG16L1, CARD9, UBQLN4, ITLN1, FCGRA2A, SLC22A4, REV3L, LACC1, ZPBP2/GSDMB, TUBD1, CD226, MST1/BSN, YDJC*).[80]

1.6 Evolution of Crohn's Disease

The global increase in Crohn's Disease in recent decades suggests that epigenetic epidemiology and environmental health have an impact on this health problem. This indicates that environmental factors have the potential to increase disease incidence. Crohn's Disease, which begins in childhood and is increasingly seen in the elderly, should be taken into consideration in terms of public health. Therefore, it is important to consider Crohn's Disease from a broad perspective in terms of the effect of age, diagnosis and environmental factors.[108,109] The interrelatedness and frequency of genes associated with the response to mycobacterial diseases and infection suggest that Crohn's Disease may represent an evolutionary adaptation against environmental microbes.[103]

Tuberculosis (TB) has a significant epidemiological impact on society and has shaped population genetics. Examination of TB-related susceptibility genes revealed overlap in genes associated with Crohn's Disease and an associated risk of autoimmunity.[110,111] The observed similarities between TB and CD play an important role in defining. CD as a multigenetic environmental enteropathy affected by overlapping susceptibility genes and enteric microbiome dysbiosis. On the other hand, although *Mycobacterium paratuberculosis* was isolated from patients with CD, clinical studies failed to demonstrate a causal role, highlighting the complexity of the relationship between the diseases and the need for more details to be understood.[112,113] Although the absence of Tuberculosis symptoms in the pathology of the reported cases associated with CD has led researchers to investigate the etiology of CD other than Koch infection, its role in shaping multigenetic environmental sensitivity continues. Ongoing studies have helped to better understand the nature of the disease. In particular, nucleotide-binding oligomerization domain-containing protein 2 (*NOD2*) plays an important role in mycobacterial immunity.[114] Frameshift mutation in the *NOD2* gene has been determined to be an important susceptibility factor for Celiac Disease. Frameshift mutation in the gene encoding *NOD2* has been identified as a strong susceptibility factor for CD.[79]

1.7 Ancient DNA Studies

Ancient DNA (aDNA) studies began in 1984 with the isolation and sequencing of DNA samples from the South African Horseman (Equus quagga quagga), a zebra species that became extinct in the early 20th century, and an Egyptian child mummy.[115,116] Bacterial cloning technique was used to amplify aDNA sequences and the resulting material was found to be of fungal or microbial origin. Endogenous DNA was generally limited to very low concentrations of damaged and short fragments of multicopy loci, such as mitochondrial DNA (mtDNA).[117,118] In the following years, thanks to the invention of the Polymerase Chain Reaction (PCR), this study enabled the rapid progress and development of the field by allowing the routine amplification and examination of extant aDNA molecules.[117,119–122]

Along with these developments, PCR product amplifications have also revealed the problem of contamination that may arise from modern DNA, which can be a potential problem in terms of the reliability and accuracy of the studies carried out. Apart from the contamination factor, post-mortem DNA degradation (mutations) and insufficient amount of aDNA sequence are the difficulties that can be encountered in aDNA studies.[117,118]

Despite these difficulties, ancient DNA studies are of great importance in elucidating the history of human evolution. In particular, GWAS are studied in ancient and modern populations; It can enable the identification of thousands of loci associated with complex traits and diseases and the detection of frequencies and changes in allele frequencies occurring at genetic loci after selection.[123]

1.8 Hypothesis and Aims of the Study

The genetic structure that increases susceptibility to Crohn's Disease in modern populations has evolved as a derived selected trait, conferring selective advantages against certain environmental stresses such as infectious diseases and pathogens. As a result, it is expected that signs of selection will be observed in genes associated with Crohn's Disease.

To test whether Crohn’s Disease susceptibility is a selected derived trait, to compare the distribution of disease risk status (susceptible vs. protective) of alleles between populations with high Crohn’s Disease prevalence from their respective ancestors and derived states. To compare population differentiation among Europeans, East Asians, South East Asians, and Africans by CD-associated alleles and genome-wide. In order to understand the change in the evolutionary history of Crohn’s Disease susceptibility, it was aimed to compare allele frequency differences between today's populations and ancient populations. Performing population genetic testing to identify possible selection in genes associated with Crohn’s Disease.

CHAPTER 2

MATERIALS AND METHODS

2.1 Data Collection

In the thesis titled "Population Genomics of Susceptibility to Crohn Disease", firstly, a literature study was conducted. For the literature study, articles were scanned using various keywords such as "Crohn's Disease", "Genetics", "GWAS", "Selection", "Susceptibility", "Population Genomics" between 2000 and 2024. More than 500 articles related to candidate gene studies were scanned in the literature review. Using Ensembl and NCBI databases, data on SNPs associated with disease were collected by Type, Chromosome, Ancestral Allele, Derived Allele, Risk Allele, Effect, P Value, OR Value and disease-associated population for use in the analyses (Appendix A).[124,125]

The frequencies of alleles associated with risk alleles of disease-associated SNPs in African, American, European, East Asian and South Asian populations including 2504 modern-day individuals were collected from the Ensembl 1000 Genomes Project Phase 3 data.[124] Data on ancient populations were collected from the Allen Ancient DNA Resource on the David Reich Laboratory website.[126] By grouping populations within the data set, periods from BCE and before BCE were selected for analysis. The data required for allele frequency analyzes of the modern Turkish population were taken from the Turkish Variome data, which includes 3362 Turkish individuals.[127]

2.2 Population Genetics Test Based on Single Nucleotide Polymorphism Data

Analyzes were performed to focus on Europeans (high CD occurrence), Asia (low CD occurrence), and Africans (very low CD occurrence). PLINK (whole genome

association analysis toolset) version 2.0 and 1.9 were used to observe distribution and differentiation between populations by Principal Component Analysis (PCA) and to compare SNP-based fixation indices (Fst) between 1000 Genome populations.[128]

The 1000 Genome phase 3 data set (containing 2504 present-day individuals) from the PLINK website was downloaded for PCA and Fst analyzes to be performed using the PLINK tool set in the R environment.[129] Allele frequencies of disease-associated SNPs in modern African, American, European, East Asian and South Asian metapopulations and subpopulations of these populations were calculated with PLINK. Principal Component Analysis (PCA) was used to examine the distribution of allele frequencies of relevant SNPs among modern populations and data visualization was performed in R. Fst analysis of SNPs associated with Crohn's Disease in modern African, American, European, East Asian and South Asian meta-populations and subpopulations of these populations was performed via PLINK and data visualization was performed in R.

From the Allen ancient data set, subpopulations containing a sufficient number of samples for the accuracy and significance of the analyzes from the BCE and pre-BCE periods were selected. In addition, these ancient populations were grouped to correspond to subpopulations in the modern population so that allele frequency differences could be calculated. For the Ancient European subpopulation, data were collected from Italy (for TSI), Spain (for IBS), England and Scotland (for GBR) and Austria, Germany and France (for CEU). For the ancient East Asian subpopulation Japan (for JPT), Vietnam (for KHV), Han Chinese (for CHB), and Southern Han Chinese (for CHS) were collected from the Allen dataset. For the ancient South Asian subpopulation Pakistan (for PJL) collected from the Allen dataset. For the ancient African subpopulation, Kenya, Malawi and Tanzania (for LWK), and Morrocco and Cameron (for GWD, YRI and ESN) were collected from the Allen dataset. The ancient Turkish population data was included in the Allen data.

Allele frequencies of CD-related SNPs in the modern-ancient Turkish population and Ancient populations were calculated in PLINK. Principal component analyzes of the allele frequencies obtained as a result of the analysis were performed in R and the data were visualized. The difference in allele frequencies obtained from modern and ancient populations was calculated. PCA analysis was performed using allele frequencies obtained from the differences between these populations, and the data was visualized.

2.3 Gene Sequence Based Selection Analyzes

To examine selection in genes associated with Crohn's Disease in different populations, positive selection within a single population was examined by integrated haplotype score (IHS) analysis using the rehh (Extended Haplotype Homozygosity (EHH)) package in R.[130,131] With the analyses, genomic regions that have recently been under natural or artificial selection can be identified in the genetic data of different populations and various sign about the molecular mechanisms of adaptation can be obtained.[131]

For IHS analysis, VCF files and DNA sequences containing 14 subpopulations belonging to European, East Asian and African metapopulations of 196 genes common to both modern and ancient populations associated with CD were downloaded using the Ensembl DataSlicer tool based on 1000 Genomes data. When creating VCF files, 50K bases were added to the beginning and end of the position of the relevant gene to prevent the selection signals from other SNPs in the gene from being compromised.[124] Manhattan and other related plots of EHH and IHS values obtained as a result of gene-based selection analyzes were constructed.

2.4 Biological Pathway and Molecular Function Analyses

Gene Function (Protein Class, Biological Process, Molecular Function, Cellular Component and Pathway) information of disease-related genes was collected using the PANTHER (Protein ANalysis THrough Evolutionary Relationship) online tools (Appendix B).[132] In this way, the functions and interactions of genes and variants associated with Crohn's Disease were determined. Additionally, STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) biological database was used to examine predicted and known protein-protein interactions of Crohn's Disease-associated genes.[133]

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Data Collection

As a result of the literature review based on GWAS and candidate gene studies, 240 genes and 352 variants associated with Crohn's Disease were identified. These results are given in detail in Appendix A. Among the 240 genes associated with CD, the genes with the most variants are *IRGM*, *NOD2*, *NKX2-3* genes, and after these genes, *ATG16L1*, *IL23R*, *LOC124900967_LOC105374737*, *TNFSF15*, *OR2B11* and *PTPN2* genes contain more than one variant associated with the disease. Of the variants linked to CD, about half of these SNPs were located in the intron region, while the other SNPs were frequently located in the missense region and less frequently in upstream and intergenic regions. In the data obtained, it was observed that the majority of CD-associated genes and variants were observed in European and subpopulations, followed by Asia, with a predominance of Japanese and very rare in African Americans. When the protective and susceptible effects of disease-associated SNPs were analyzed in these three populations, 87% of the 325 variants reported for Europeans were susceptible, 91% of the 79 variants for Asians were susceptible, and 75% of 8 variants for African Americans were protective (Chi-square P value < 0.001). Indicating excess of susceptible variants in Europeans and Asians.

To determine whether susceptibility to CD is an ancestral or derived trait, the distribution of Crohn's Disease-associated genes and variants, derived and ancestral allele status, and susceptible and protective effects on CD were analyzed. Susceptible effects were slightly higher in the derived alleles (57% vs. 43%), whereas protective effects had slightly more derived alleles than ancestral alleles (67% vs. 33%). However, the distribution of susceptible and protective effects among ancestral and derived allele status was not statistically significant (Chi-square P value = 0.18).

3.2 Population Genetics Test Based on Single Nucleotide Polymorphism Data

3.2.1 Principal Component Analyses with Crohn's Disease Associated SNPs

Principal Component Analysis (PCA) of 348 CD-associated variants in PLINK was performed using allele frequencies of Africa (AFR), America (AMR), Europe (EUR), East Asia (EAS), South Asia (SAS) and 26 subpopulations of these meta-populations of 2504 modern individuals from the 1000 Genomes phase 3 dataset. Out of a total of 352 disease-associated SNPs, 348 SNPs were studied because the relevant allele frequencies of 4 variants were not available through Ensembl. The first two principal components, which provide important information about population differentiation, were observed as 35.61% (PC1) and 15.61% (PC2) (Figure 3.1). America overlaps with 4 metapopulations, overlaps with Europe and South Asia. East Asia and Europe are separated, and Africa is clearly separated from all continents (Figure 3.1).

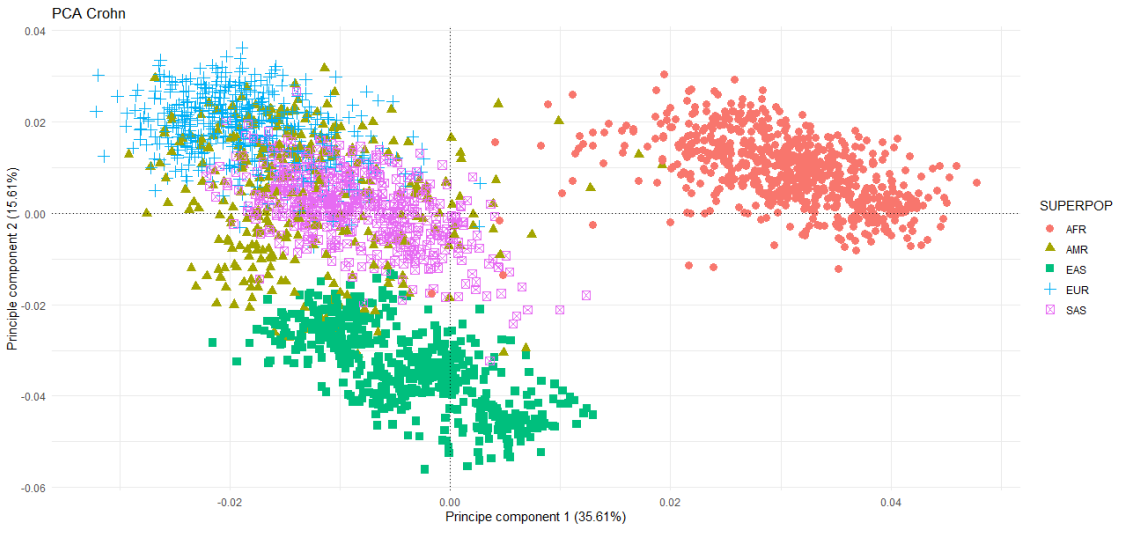

Figure 3.1 PCA analysis plot of 5 metapopulations comprising 2504 modern individuals with 348 SNPs associated with CD in PLINK.

In subsequent PCA studies, America was removed from these analyses. Thus, it was possible to clearly observe the distribution and separation of other populations among each other. In particular, it was clearly observed that Europe and parts of South Asia overlap and that East Asia is separated from these two populations (Figure 3.2).

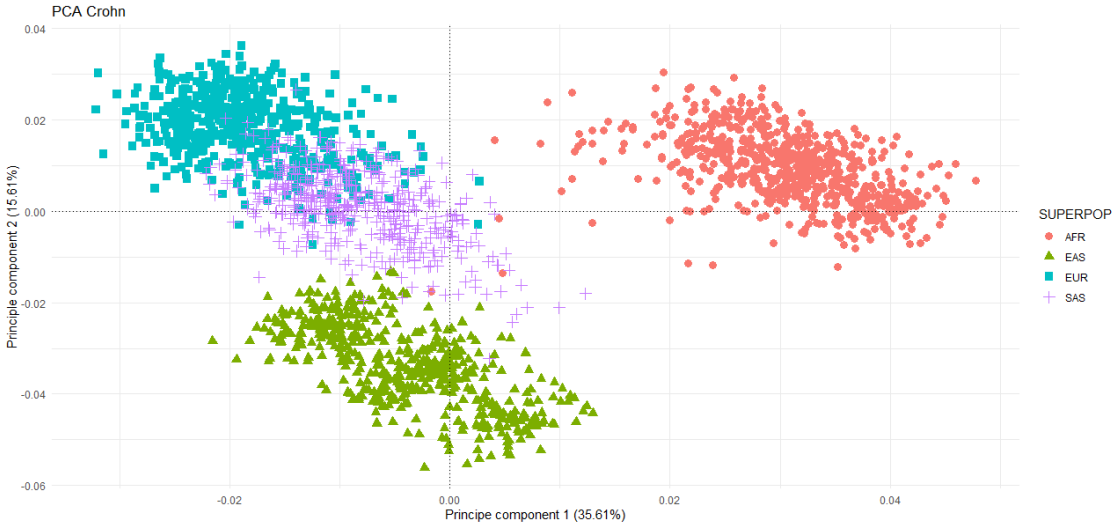

Figure 3.2 PCA analysis plot of 4 metapopulations comprising 2504 modern individuals with 348 SNPs associated with CD in PLINK.

In a PCA analysis of 4 modern worldwide metapopulations based on the allele frequencies of 348 CD-associated SNPs, two principal components were observed at 51.2% (Dim1) and 39.6% (Dim2), providing important information about population differentiation (Figure 3.3). The first PC clearly separated Africa from European, East and South Asian populations. In this distinction *PUS10* (rs13003464) is one of the variants with a significantly higher loading coefficient. *PUS10* also has a loading value that distinguishes Europeans from other populations. Variants of *STAT3* (rs4796791), *HERC2* (rs916977) and *RIT1* (rs670523) are also included, which contribute to the high effect in this distinction.

The second PC separates the East Asian population from the South Asian and European populations. *AIMP1P2* (rs9286879, rs2157453, rs7517810) and *CENPW* (rs9388489) variants are located in the second PC, which distinguishes the East Asian population from other populations. *PPBP_CXCL5* (rs2472649) has high loading in the

separation of East and South Asia. *IRGM,ZNF300* (rs4958427) is one of the variants with a loading value that contributes to the separation of Europe and South Asia.

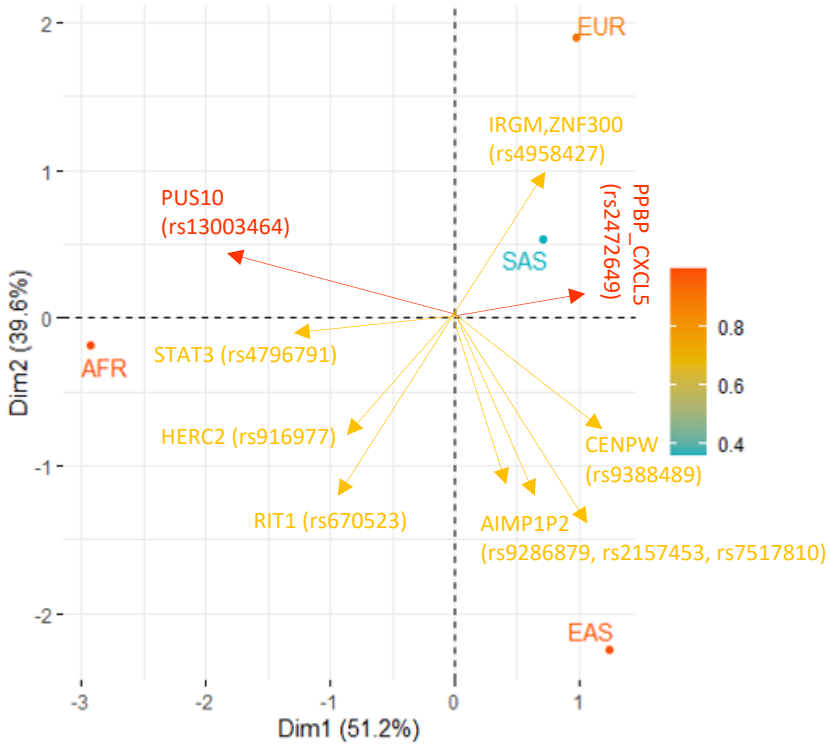

Figure 3.3 PCA analysis of 4 worldwide modern metapopulations based on allel frequences of 348 SNPs associated with CD. Colors indicate the degree of separation between metapopulations and loading of variants on the axes.

Two principal components, which provide important information about the differentiation of subpopulations of the European population, were observed at 54.9% (Dim1) and 24% (Dim2) (Figure 3.4A). The first PC clearly distinguished Finland (FIN) from other remaining populations. One of the variants that has a loading effect in this distinction is *HORMAD2* (rs713875). The second PC separated Italy (TSI) and Spain (IBS) in Southwest Europe from England and Scotland (GBR), and Caucasian (CEU) in Northwest Europe. In this distinction, *HERC2* (rs916977) is one of the variants with a significantly higher loading coefficient. In this distinction, *FUT2_MAMSTR* (rs281379) and *NKX2-3* (rs4919345) are some of the contributing variants. Additionally, *HERC2* (rs916977) effectively separated Finland from other Northwestern European populations.

Two principal components that provide important information about the differentiation of subpopulations of the East Asian population were observed at 48.2% (Dim1) and 27.8% (Dim2) (Figure 3.4B). The first PC separated Han Chinese (CHB) and Japan (JPT) from Xishuangbanna (CDX) and Vietnam (KHV). One of the variants with a loading effect in this distinction is *TMEM258* (rs102275). The second PC clearly separated Han Chinese and Japanese from each other and from the remaining populations. *ANKRD55* (rs10065637) is one of the variants that has a significantly high loading coefficient, particularly in distinguishing Han Chinese from other populations. *IFNGR2* (rs2284553) is also partially effective in this distinction. Southern Han Chinese (CHS) is less segregated than other populations.

Two principal components were observed at 33% (Dim1) and 26.9% (Dim2), providing important information about the differentiation of subpopulations of the South Asian population (Figure 3.4C). The first PC separated Sri Lankan Tamil (STU) and Indian Telugu (ITU) from Gujarati Indian (GIH), Pakistan (PJL) and Bangladesh (BEB). One of the variants that distinguishes Sri Lankan Tamil and Indian Telugu from other populations is *IFNGR2* (rs2284553). The second PC, in particular, has largely isolated Bangladesh from other populations. One of the variants with a high loading effect in this distinction is *IL23R* (rs76418789). *IRGM,ZNF300* (rs4958427) is one of the variants with loading value that contributes to the distinction between Gujarati Indian and Pakistan.

Two principal components were observed at 31.1% (Dim1) and 25.7% (Dim2), which provide important information about the differentiation of subpopulations of the African population (Figure 3.4D). The first PC separated African Caribbeans (ACB) and especially American Africans (ASW) from other remaining populations. One of the variants that has an impact on this distinction is *IKZF3* (rs12946510). Additionally, *HLA* (rs3129871) is one of the variants with a significantly high loading coefficient. It was expected that especially American Africans would separate from this group. Because this population is based on African American origin living in the Southwest America. Similarly, this differentiation is meaningful since African Caribbeans are associated with a population living in Barbados. The second PC was observed to provide a partial distinction between the East African population of Kenya (LWK) and the West African populations of Esan (ESN), Yoruba in Ibadan (YRI), Mende in Sierra Leone (MSL) and Gambian (GWD). It was observed that the *HLA* (rs3129871) variant had a significant loading value, especially in distinguishing Kenya from other populations. The second PC also provided a distinction between the Esan and Yoruba in Ibadan from the Mende in

Sierra Leone and Gambian population. In this distinction, *IRGM* (rs11957134), *IL12B* (rs6887695), and *STAT3* (rs4796791) are some of the contributing variants.

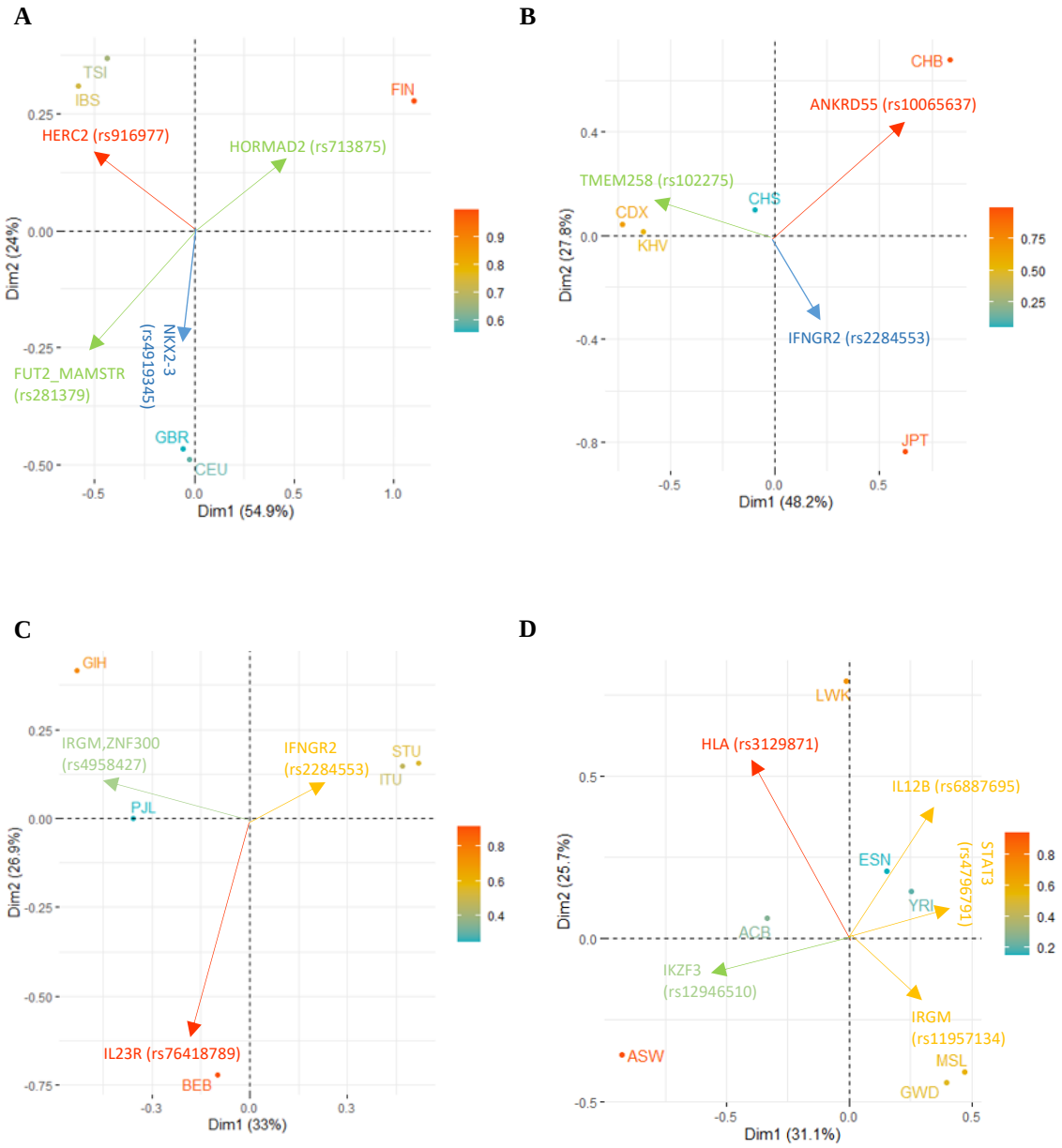

Figure 3.4 PCA analysis of 22 worldwide modern subpopulations, five EUR subpopulation (A), five EAS subpopulation (B), five SAS subpopulation (C), and seven AFR subpopulation (D) based on allel frequences of 348 SNPs associated with CD. Colors indicate the degree.

In general, when the subpopulations of all metapopulations were examined, it was observed that all of them were separated/differentiated within themselves. In this case, it made sense to examine all populations individually.

From a total of 348 SNPs associated with CD, 262 variants were found to be common in the Modern-Ancient Turkish population and 4 ancient metapopulations, and the allele frequencies of these common variants were calculated in PLINK. Principal component analysis of the allele frequencies obtained as a result of the analysis was performed. In the analysis for four ancient metapopulations and ancient Türkiye population, the first two principal components, which provide important information about population differentiation, were observed at 40.5% (Dim1) and 24.3% (Dim2) (Figure 3.5). The first PC clearly separated the East Asia population (Han Chinese and Southern Han Chinese) from the Europe (Italy, Spain, Austria-Germany-France and England-Scotland), Türkiye, South Asia (Pakistan) and Africa (Kenya-Malawi-Tanzania and Morocco-Cameroon) populations. In this distinction, *HLA-DQB1_MTCO3P1* (rs9469220) and *LOC105378327_ALDH7A1P4* (rs10761659) are some of the variants with significantly high loading coefficients. The second PC specifically separated the African population from other populations. In this differentiation, especially the *LOC107986537* (rs68191) variant is one of the variants with a significant loading value. *PPP5C* (rs4802307) had a significant impact on the separation of European and Turkish populations from South Asia. Overall, all three metapopulations were observed to be separated/grouped from each other (partially South Asia).

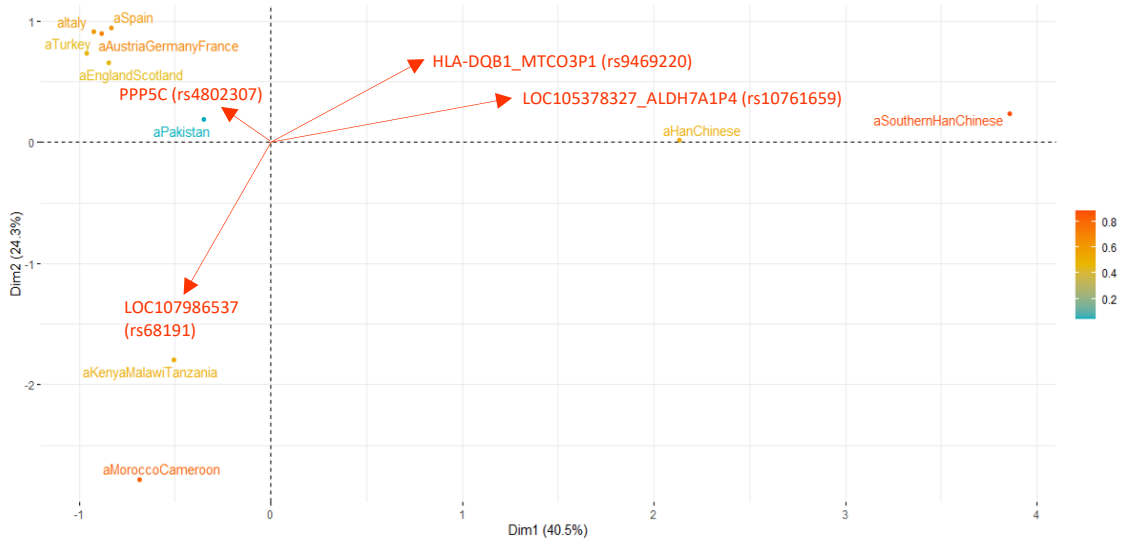

Figure 3.5 PCA analysis of subpopulations of 4 ancient metapopulations and ancient Türkiye based on allele frequencies of 262 SNPs associated with CD. Colors indicate the degree of separation between metapopulations and the loading of variables on the axes.

In the analysis of four ancient and modern metapopulations and the population of Türkiye, the first two principal components, which provide important information about population differentiation, were observed as 34.8% (Dim1) and 29.4% (Dim2) (Figure 3.6). The first PC separated the ancient-modern East population Asian (Han Chinese and Southern Han Chinese) from the entire remaining ancient-modern world population. In this distinction, *AIMP1P2* (rs7517810) in particular is one of the variants with significantly higher loading coefficients. The second PC specifically distinguished the ancient-modern African population (Kenya-Malawi-Tanzania and Morocco-Cameroon) from other ancient-modern populations. *KIF15* (rs3804583), *PUS10* (rs13003464) and *FAF1* (rs11205760) variants were especially effective in this differentiation. Especially *PPP5C* (rs4802307) and *PPBP_CXCL5* (rs2472649) variants have an important loading value in differentiating the ancient-modern European and Turkish populations, and partly from South Asia (Pakistan), from other populations. *PPBP_CXCL5* (rs2472649) also provided a significant effect on the differentiation of the African population. In general, all four metapopulations differed from each other, and ancient and modern subpopulations showed a certain separation/grouping within themselves.

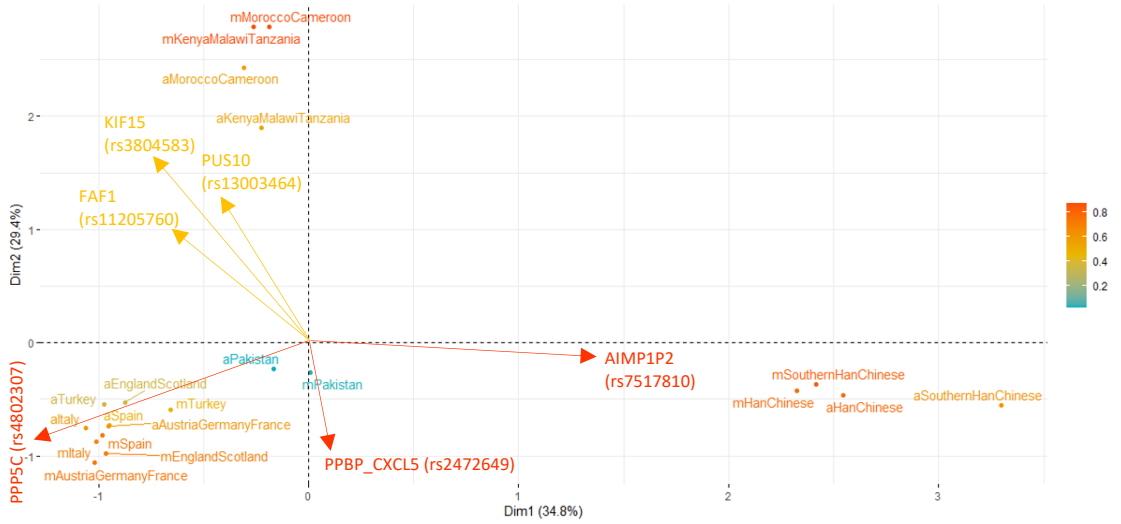

Figure 3.6 PCA analysis of subpopulations of 4 ancient and modern metapopulations and Türkiye population based on allele frequencies of 262 SNPs associated with CD. Colors indicate the degree of separation between metapopulations and the loading of variables on the axes.

In the analysis conducted for the ancient and modern European subpopulation and the Turkish population, the first two principal components, which provide information about population differentiation, were observed as 34.1% (Dim1) and 21.2% (Dim2) (Figure 3.7). The firstly PC distinguished ancient and modern European subpopulations. In this distinction, *IRGM* (rs7714584) and *OR2B11* (rs10733113) variants are some of the variants with significantly high loading coefficients. *TMEM258* (rs102275) and *CARD9* (rs4077515) are some of the variants that are effective in distinguishing ancient Europe, and *SLC22A4*, *MIR3936HG* (rs1050152) and *IRF-AS1* (rs12521868, rs2188962) are effective in distinguishing modern Europe. The second PC separated the ancient-modern Turkish population from the European subpopulation. *IL10* (rs3024505) was effective in this distinction. When the allele frequencies of ancient and modern Europe were examined, it was seen that ancient and modern populations separated over time. This difference is clearly seen in the population of ancient Türkiye and modern Türkiye. With the transition from hunter-gatherer to settled life and the advent of agriculture, our way of eating is also changing. Anatolia is one of the best examples of this. The transition from animal-based nutrition to grain-based nutrition had an impact on this differentiation.

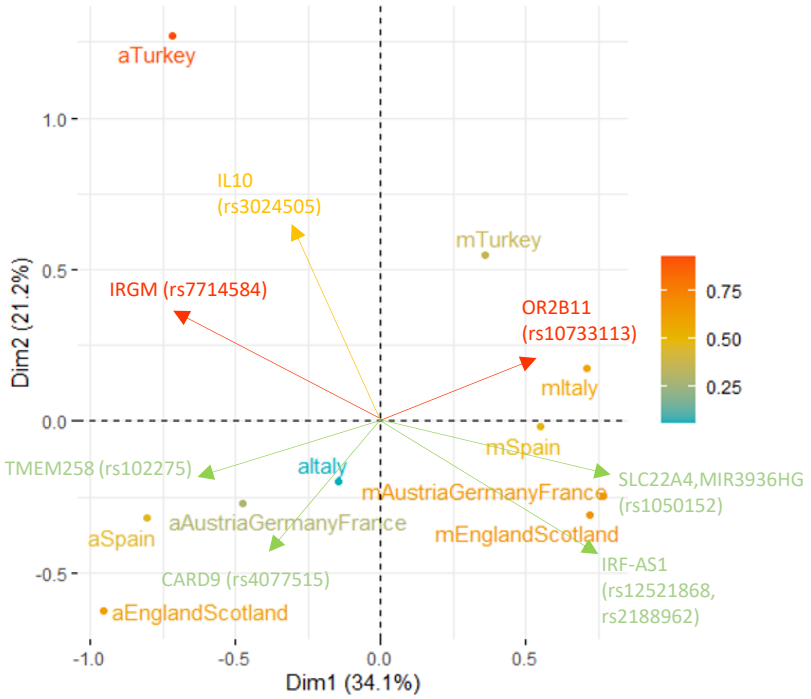

Figure 3.7 PCA analysis of ancient and modern European subpopulations and Türkiye population based on allele frequencies of 262 SNPs associated with CD. Colors indicate the degree of separation between metapopulations and the loading of variables on the axes.

SNP-based analyzes proved that both ancestral and derived alleles play a role in increasing or decreasing the risk of CD. Furthermore, while the distinction in the European population is clear, CD-associated alleles appear to be influential in differentiating subpopulations of other world populations. It shows that CD-associated alleles are not unique to Europeans but are also found at similar frequencies in other world populations. This situation reveals the effects of various genetic changes and environmental factors in today's and ancient populations over time.

3.2.2 Populations Differentiation (Fst) Analyses

The fixation index (Fst) was used to calculate the genetic distance between present-day African, European, East Asian and South Asian populations. Population

differentiation (Fst) analysis of SNPs associated with Crohn's Disease was performed using PLINK and the data obtained were visualized in R. A high value obtained as a result of this analysis can be considered as positive selection. If the result of the analysis is low, these loci may be candidates for balancing selection. The higher the Fst value, the more divergence is expected. Fst values between populations are generally below 0.2.

When examining genetic differentiation between African and European populations, the average Fst estimates of CD-associated SNPs (0.150) were higher than the genome-wide Fst estimates (0.122) (Figure 3.8).

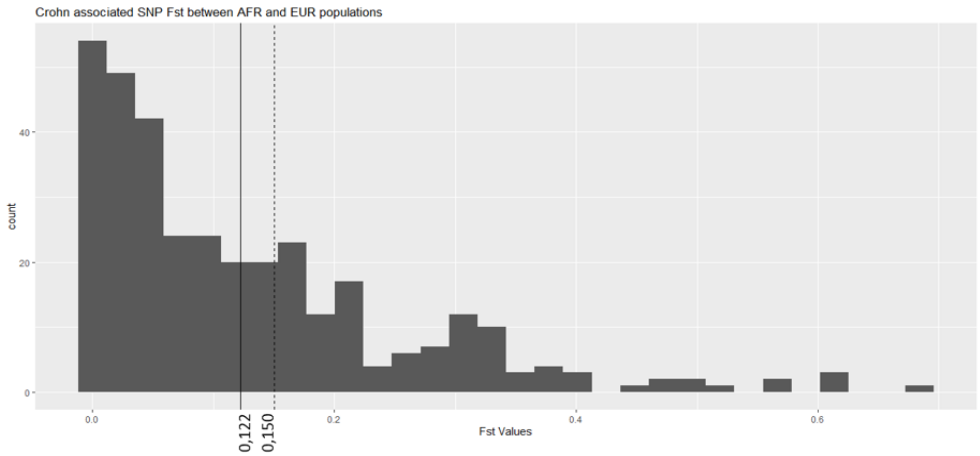

Figure 3.8 Calculation of population differentiation (Fst) for CD-linked SNPs among African (AFR) and European (EUR) populations. Solid vertical lines represent genome-wide SNP Fst, while dashed vertical lines indicate average Fst of CD-associated SNPs.

When examining genetic differentiation between Africa and East Asia populations, the average Fst estimates of CD-associated SNPs (0.173) were higher than the genome-wide Fst estimates (0.149) (Figure 3.9).

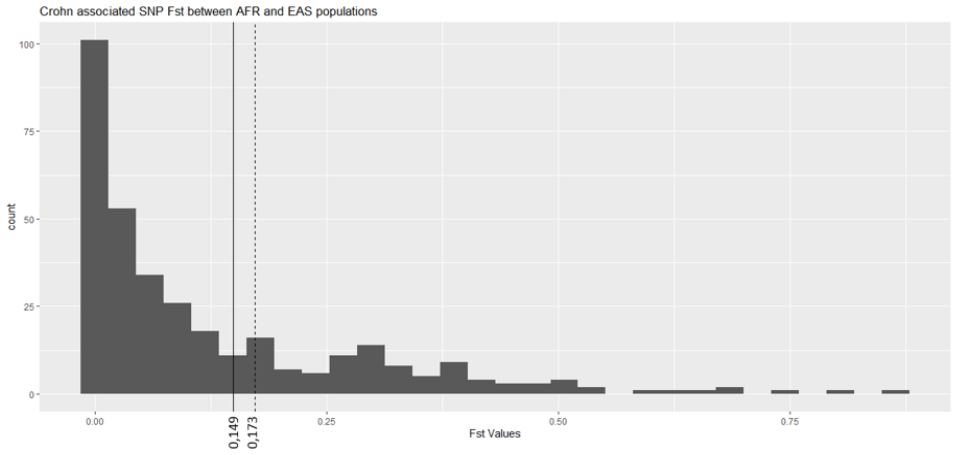

Figure 3.9 Calculation of population differentiation (Fst) for CD-linked SNPs among Africa (AFR) and East Asia (EAS) populations. Solid vertical lines represent genome-wide SNP Fst, while dashed vertical lines indicate average Fst of CD-associated SNPs.

When examining genetic differentiation between Africa and South Asia populations, the average Fst estimates of CD-associated SNPs (0.123) were higher than the genome-wide Fst estimates (0.112) (Figure 3.10).

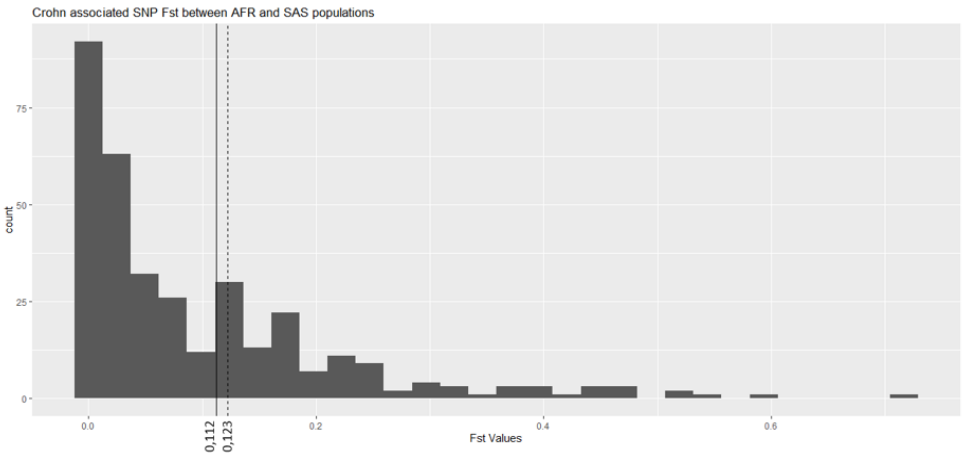

Figure 3.10 Calculation of population differentiation (Fst) for CD-linked SNPs among Africa (AFR) and South Asia (SAS) populations. Solid vertical lines represent genome-wide SNP Fst, while dashed vertical lines indicate average Fst of CD-associated SNPs.

When examining genetic differentiation between East Asia and Europe populations, the average Fst estimates of CD-associated SNPs (0.131) were higher than the genome-wide Fst estimates (0.099) (Figure 3.11).

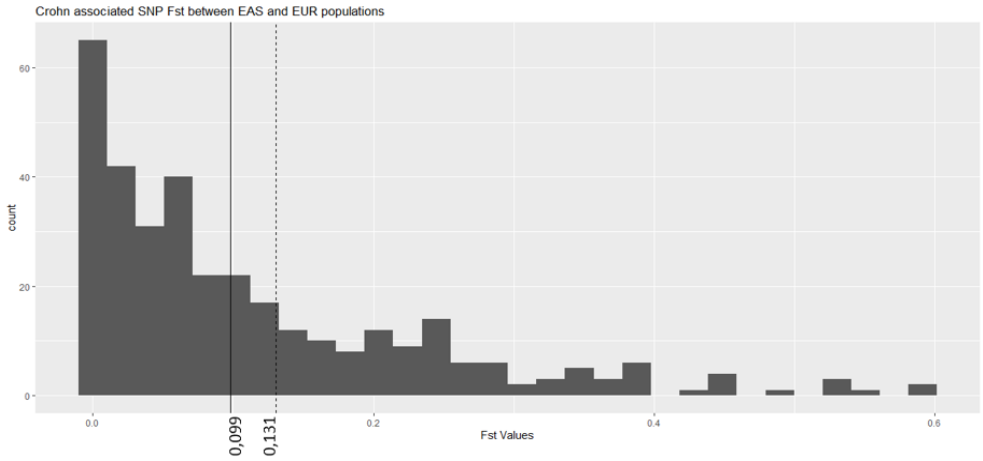

Figure 3.11 Calculation of population differentiation (Fst) for CD-linked SNPs among East Asia (EAS) and Europe (EUR) populations. Solid vertical lines represent genome-wide SNP Fst, while dashed vertical lines indicate average Fst of CD-associated SNPs.

When examining genetic differentiation between East Asia and South Asia populations, the average Fst estimates of CD-associated SNPs (0.084) were higher than the genome-wide Fst estimates (0.063) (Figure 3.12).

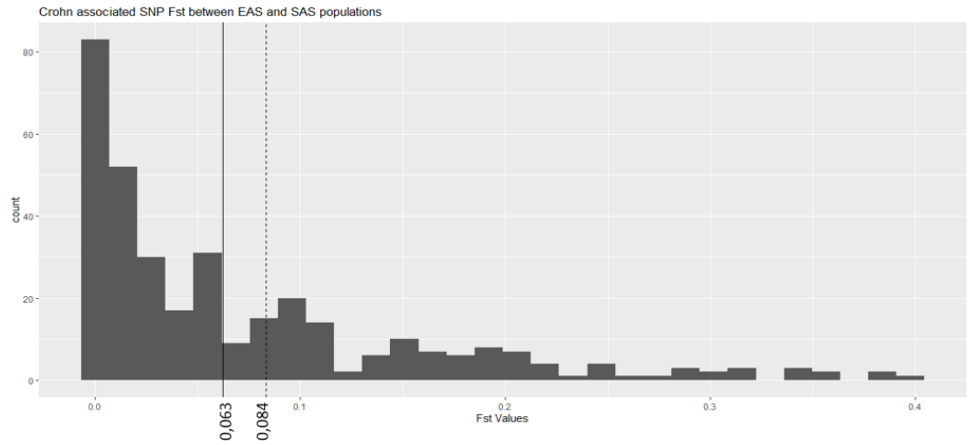

Figure 3.12 Calculation of population differentiation (Fst) for CD-linked SNPs among East Asia (EAS) and South Asia (SAS) populations. Solid vertical lines represent genome-wide SNP Fst, while dashed vertical lines indicate average Fst of CD-associated SNPs.

When examining genetic differentiation between European and South Asia populations, the average Fst estimates of CD-associated SNPs (0.044) were higher than the genome-wide Fst estimates (0.032) (Figure 3.13).

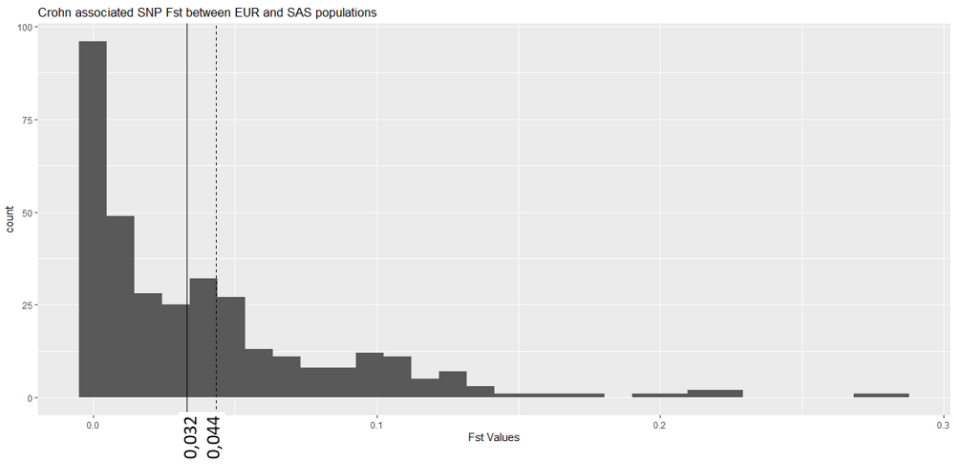

Figure 3.13 Calculation of population differentiation (Fst) for CD-linked SNPs among Europe (EUR) and South Asia (SAS) populations. Solid vertical lines represent genome-wide SNP Fst, while dashed vertical lines indicate average Fst of CD-associated SNPs.

The most variation in Fst estimates of CD-associated SNPs between populations was observed between Africa - East Asian population (0.173), Africa - Europe (0.150) population, East Asia - Europe population (0.131) and Africa - South Asia populations (0.123), respectively. The least variation between populations was observed between European - South Asian (0.044) and East - South Asian populations (0.084). When the data obtained as a result of the Fst analysis were evaluated, the genetic distance between the populations was similar to the PCA result. In these results, Africa is generally the population with the greatest genetic distance from other populations. Europe and South Asia are one of the populations with close genetic distance to each other, as shown in PCA analyses.

3.3 Gene Sequence Based Selection Analyses

To support the CD associated variant analysis, possible selection was investigated by gene-based analysis. Since it is not possible to distinguish various alternative selection scenarios by focusing solely on allele type and frequency-based analyses, the results obtained by population genetic analyzes on the genes of CD-associated alleles need to be supplemented. Population genetic analyses were performed on CD-associated genes to determine whether the increased risk of CD in specific populations was due to selection. The Integrated Haplotype Score (IHS) was analyzed in R using Extended Haplotype Homozygosity (EHH) to examine selection in 196 genes associated with Crohn's Disease in European, East Asian and African populations. The VCF files used for IHS analysis consist of polarized data. Therefore, positive IHS values indicate ancestral and negative IHS values indicate derived alleles.

IHS values and associated P values from IHS analysis for 196 genes in European populations were analyzed. In the CEU subpopulation, the *ANKRD55* (rs556200781) gene located on chromosome 5 has the highest positive IHS value of 7.89. *ITLN1* (rs549214602) gene located on chromosome 1 has the highest negative IHS value with -4.19 (Figure 3.14). In the FIN subpopulation, *TRIB1AL,LINC02964* (rs113848747) located on chromosome 8 and *ELF1* (rs9562236) genes located on chromosome 13 had the highest positive IHS value with 5.79 and 5.78, respectively. *LOC101927745* (rs57201378) gene located on chromosome 21 has the highest negative IHS value with

-4.79 (Figure 3.15). In the GBR subpopulation, the *SPATA48_IKZF1* (rs551382470) gene located on chromosome 7 has the highest positive IHS value of 5.94. *TAGAP-AS1* (rs548894351) gene located on chromosome 6 has the highest negative IHS value with -3.83 (Figure 3.16). In the IBS subpopulation, *C17orf67* (rs111407155) located on chromosome 17 and *LINC01014* (rs9822628) genes located on chromosome 13 had the highest positive IHS value with 6.78 and 6.78, respectively. *FOXP2* (rs2694928) gene located on chromosome 7 has the highest negative IHS value with -4.16 (Figure 3.17). In the TSI subpopulation, the *RBX1_RPS9P2* (rs576516444) gene located on chromosome 22 has the highest positive IHS value of 6.89. *SLC22A23* (rs143637581) located on chromosome 6 and *MROH3P* (rs7523503) genes located on chromosome 1 had highest negative IHS value with -4.62 and -4.62 respectively (Figure 3.18). The P values associated with IHS for these genes were significant ($p<0.001$) (Table 3.1).

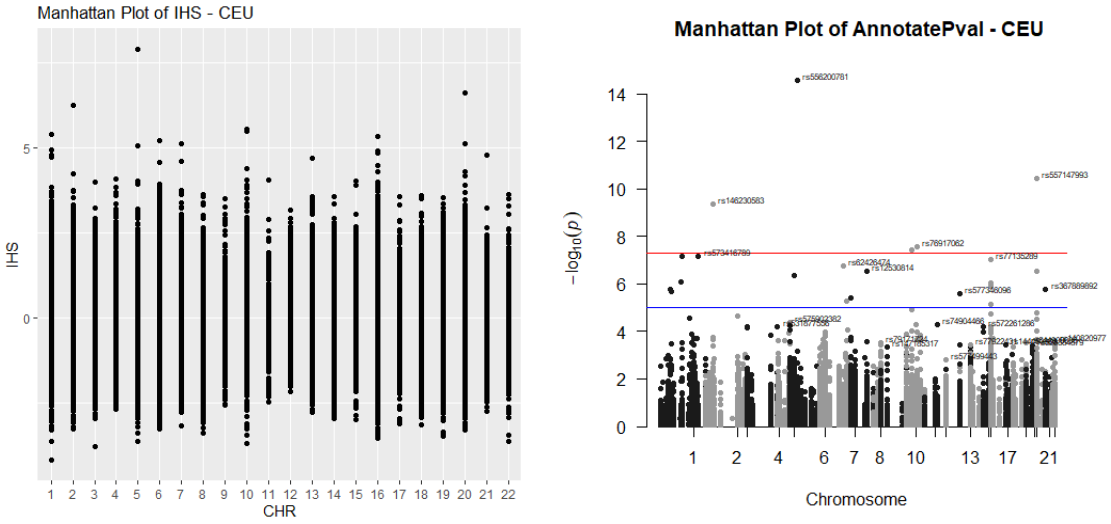

Figure 3.14 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the CEU subpopulation of the European population.

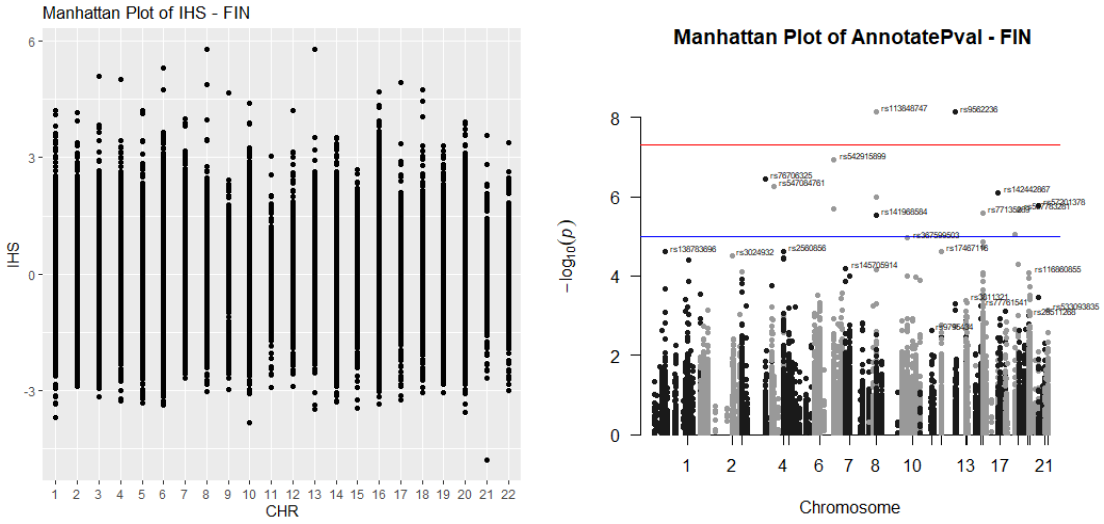

Figure 3.15 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the FIN subpopulation of the European population.

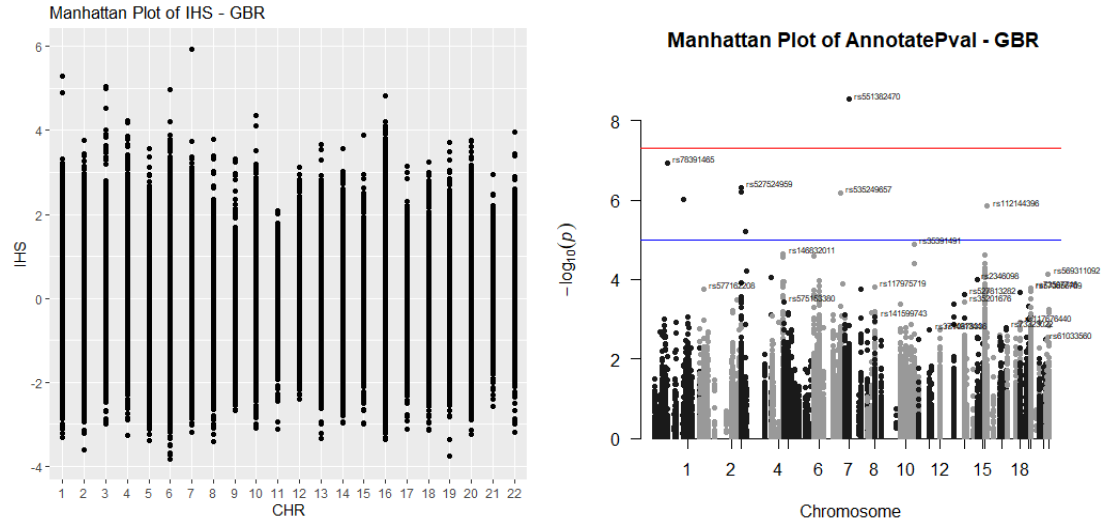

Figure 3.16 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the GBR subpopulation of the European population.

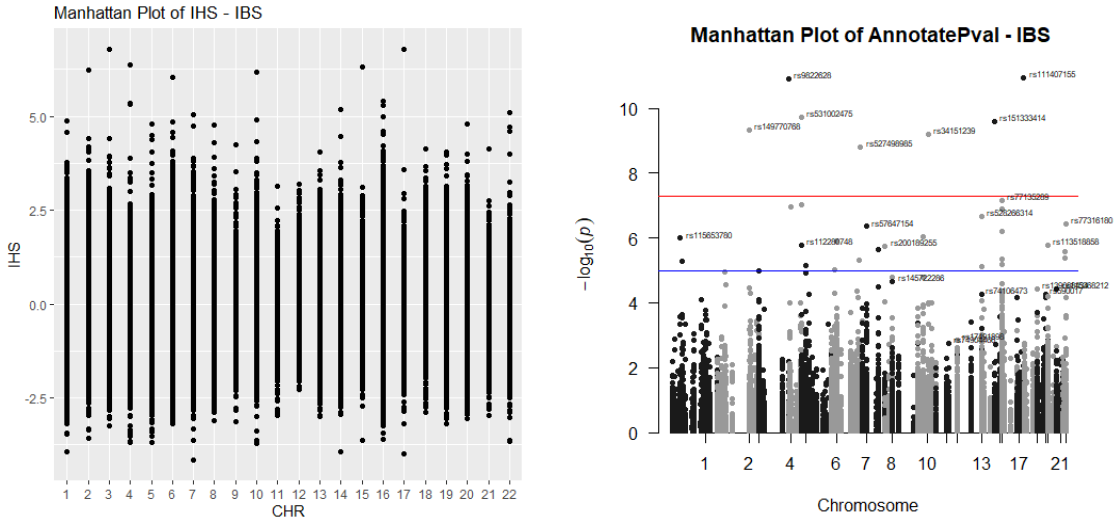

Figure 3.17 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the IBS subpopulation of the European population.

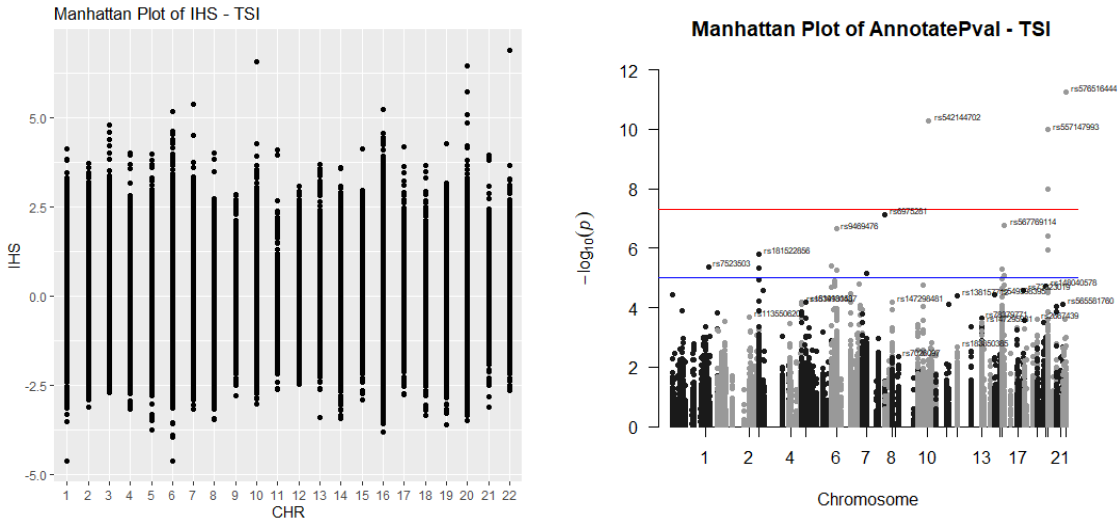

Figure 3.18 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the TSI subpopulation of the European population.

IHS values and associated P values from IHS analysis for 196 genes in East Asia populations were analyzed. In the CDX subpopulation, *LYRM4* (rs534404683) located on chromosome 6 and *LOC101928354* (rs563956224) genes located on chromosome 6 had the highest positive IHS value with 7.58 and 7.54, respectively. *PLSCR4* (rs79577099)

gene located on chromosome 3 has the highest negative IHS value with -5.37 (Figure 3.19). In the CHB subpopulation, the *LYRM4* (rs78351220) gene located on chromosome 6 has the highest positive IHS value of 8.53. *LOC107984997_RNFT1P2* (rs191740154) gene located on chromosome 1 has the highest negative IHS value with -5.21 (Figure 3.20). In the CHS subpopulation, *PHACTR2* (rs11155322) gene located on chromosome 6 and *TPPP2_NDRG2* (rs149304930) genes located on chromosome 14 had the highest positive IHS value with 7.71 and 7.67. *SKAP2* (rs189716020) gene located on chromosome 7 has the highest negative IHS value with -6.05 (Figure 3.21). In the JPT subpopulation, *TPPP2_NDRG2* (rs145618071) located on chromosome 14 and *PHACTR2* (rs11155322) genes located on chromosome 6 had the highest positive IHS value with 7.85 and 7.7, respectively. *IL23R* (rs1358748) gene located on chromosome 1 has the highest negative IHS value with -4.69 (Figure 3.22). In the KHV subpopulation, the *MAMSTR* (rs28746174) gene located on chromosome 19 has the highest positive IHS value of 6.92. *HERC2* (rs2346097) gene located on chromosome 15 has the highest negative IHS value with -4.89 (Figure 3.23).

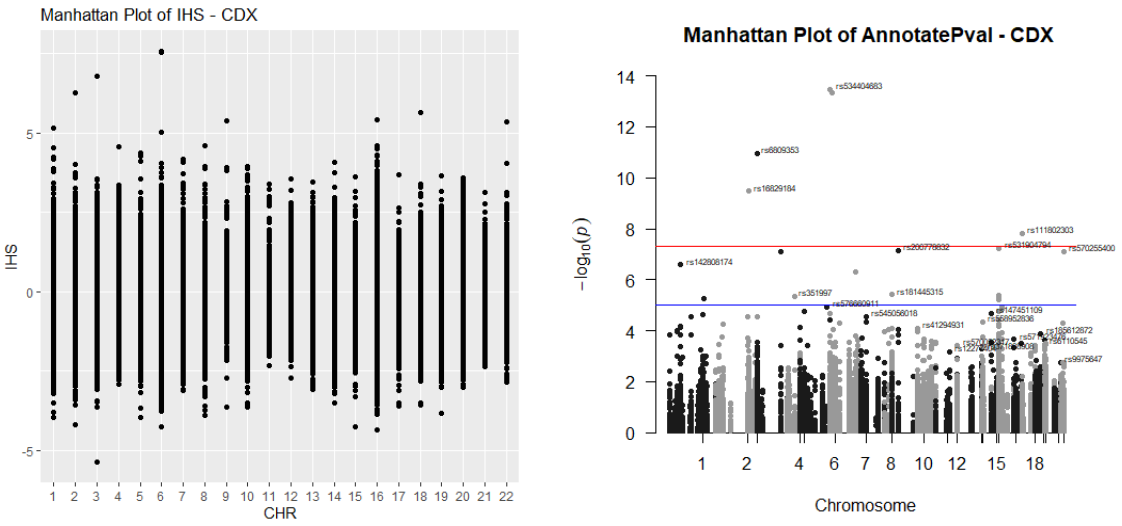

Figure 3.19 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the CDX subpopulation of the East Asia population.

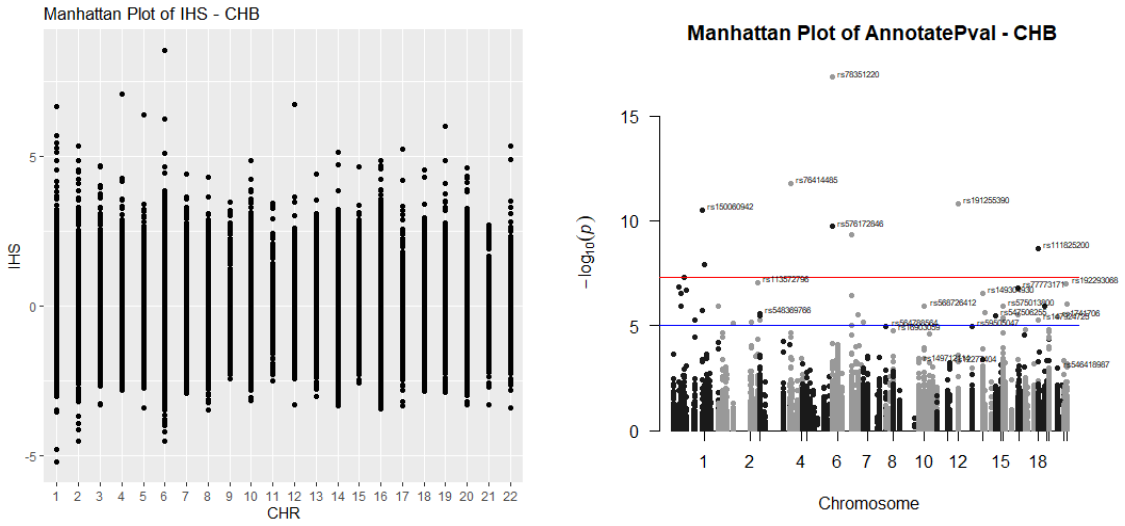

Figure 3.20 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the CHB subpopulation of the East Asia population.

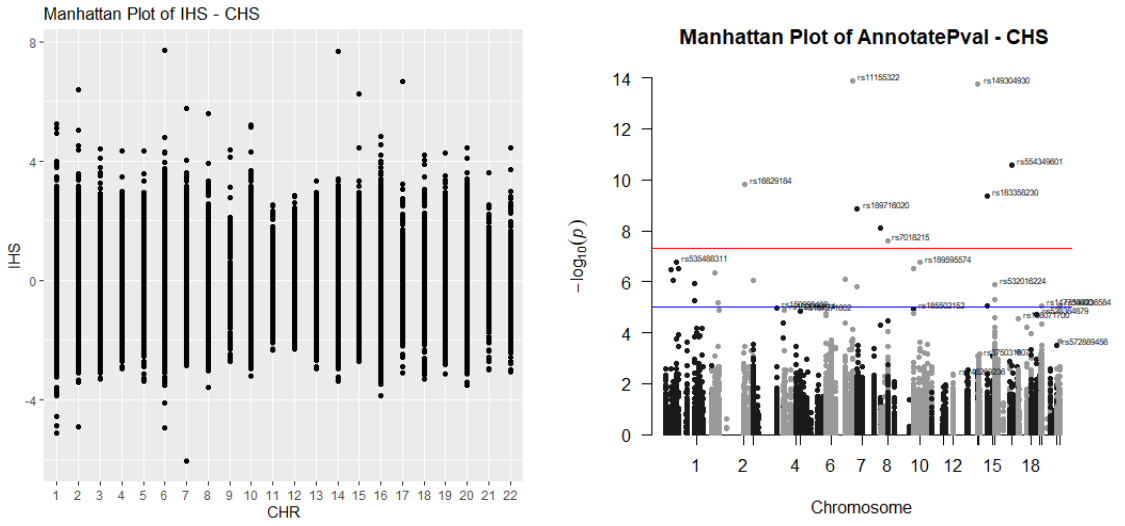

Figure 3.21 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the CHS subpopulation of the East Asia population.

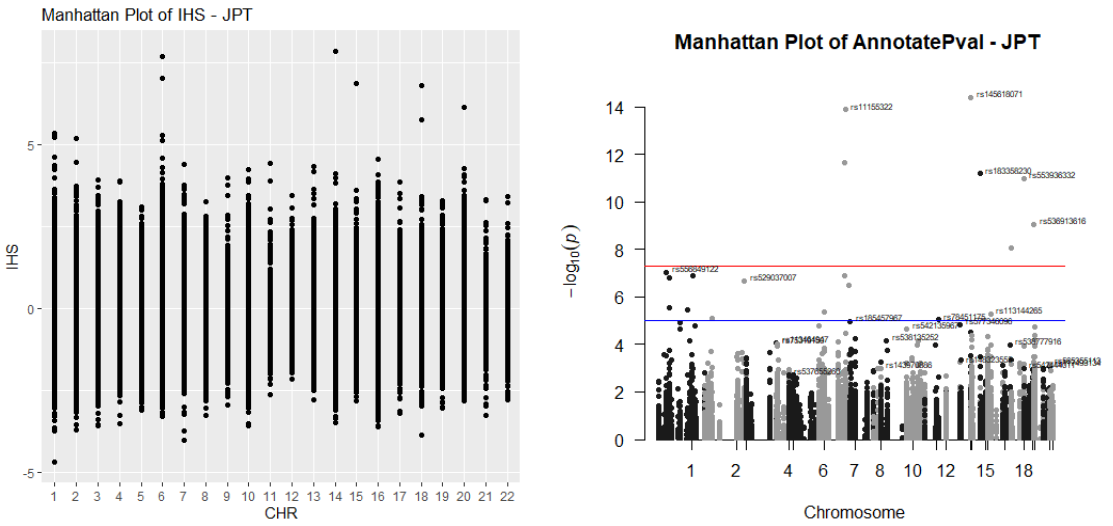

Figure 3.22 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the JPT subpopulation of the East Asia population.

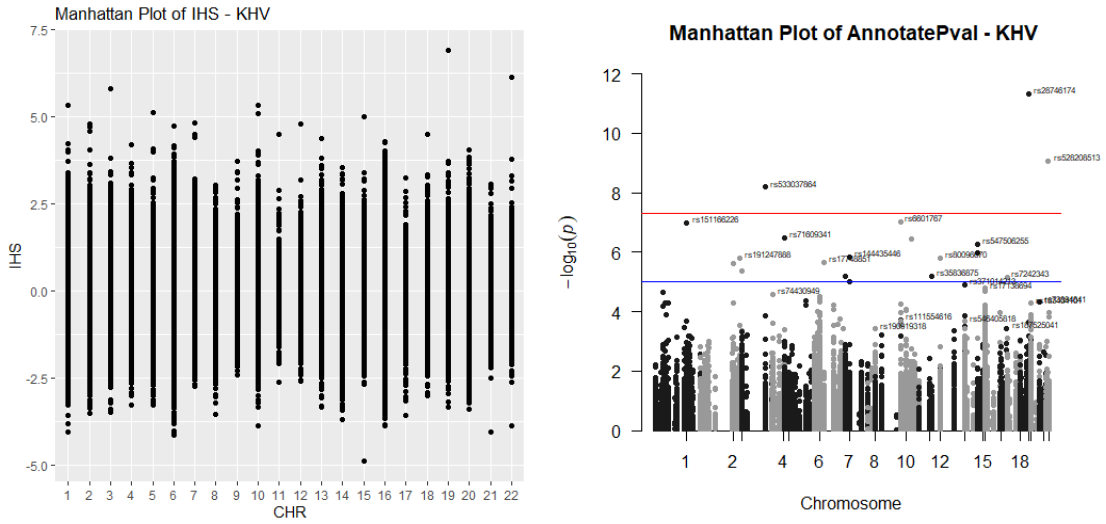

Figure 3.23 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the KHV subpopulation of the East Asia population.

IHS values and associated P values from IHS analysis for 196 genes in Africa populations were analyzed. In the GWD subpopulation, the *SLC22A23* (rs112606997) gene located on chromosome 6 has the highest positive IHS value of 5.03. *C17orf67* (rs563908371) gene located on chromosome 17 has the highest negative IHS value with

-3.77 (Figure 3.24). In the YRI subpopulation, the *MACROD2* (rs563938935) gene located on chromosome 20 has the highest positive IHS value of 5.94. *ZGPAT* (rs554831273) gene located on chromosome 20 has the highest negative IHS value with -5.97 (Figure 3.25). In the ESN subpopulation, the *IKZF3* (rs545219922) gene located on chromosome 17 has the highest positive IHS value of 6.93. *LOC107986537* (rs532430659) located on chromosome 6 and *LOC107986537* (rs191039753) genes located on chromosome 6 had the highest negative IHS value with -3.77 and -3.74, respectively (Figure 3.26). In the LWK subpopulation, the *SLAIN2* (rs182647815) gene located on chromosome 4 has the highest positive IHS value of 7.65. *RBFOX1* (rs565198015) gene located on chromosome 16 has the highest negative IHS value with -3.33 (Figure 3.27). The P values associated with IHS for these genes were significant ($p<0.001$) (Table 3.1).

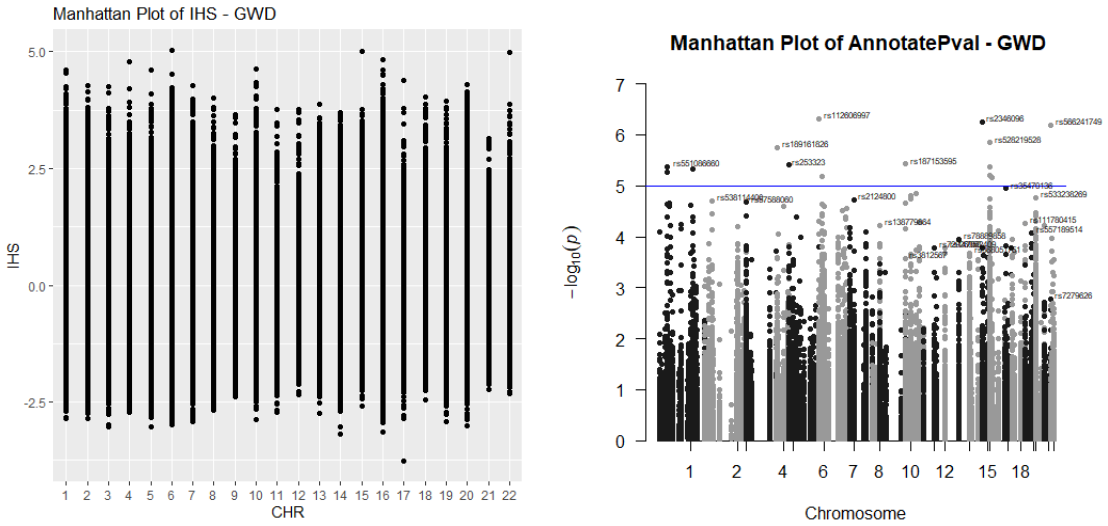

Figure 3.24 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the GWD subpopulation of the Africa population.

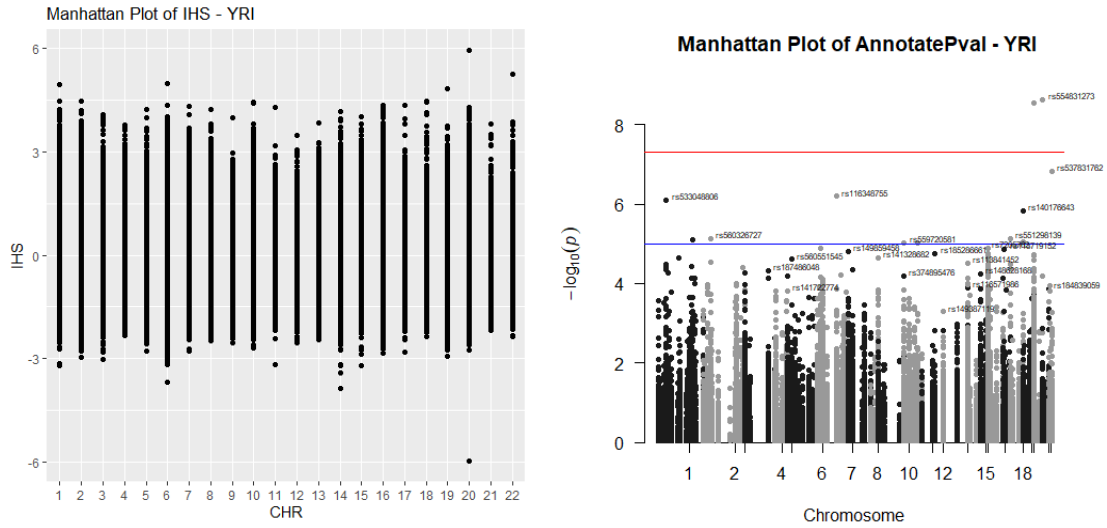

Figure 3.25 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the YRI subpopulation of the Africa population.

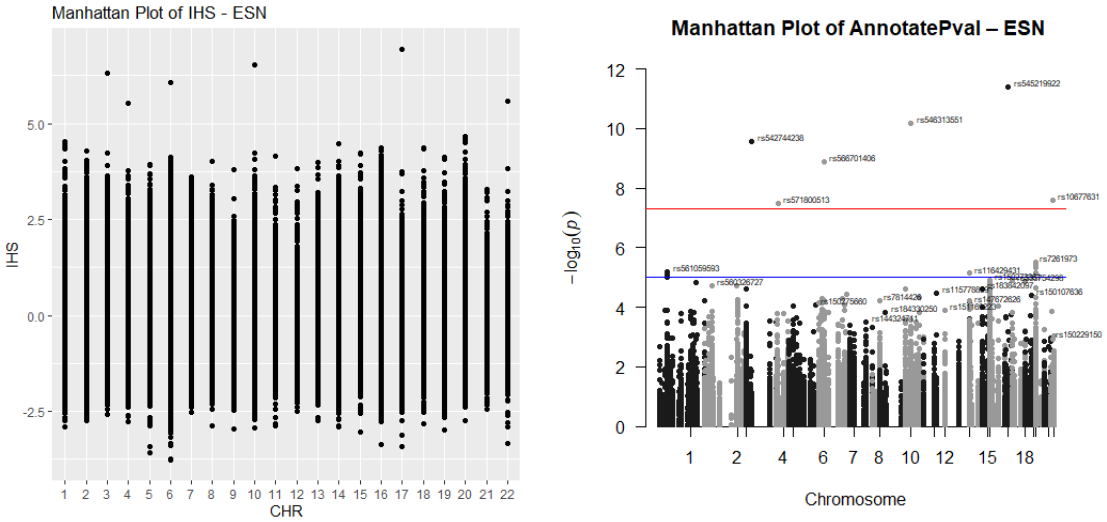

Figure 3.26 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the ESN subpopulation of the Africa population.

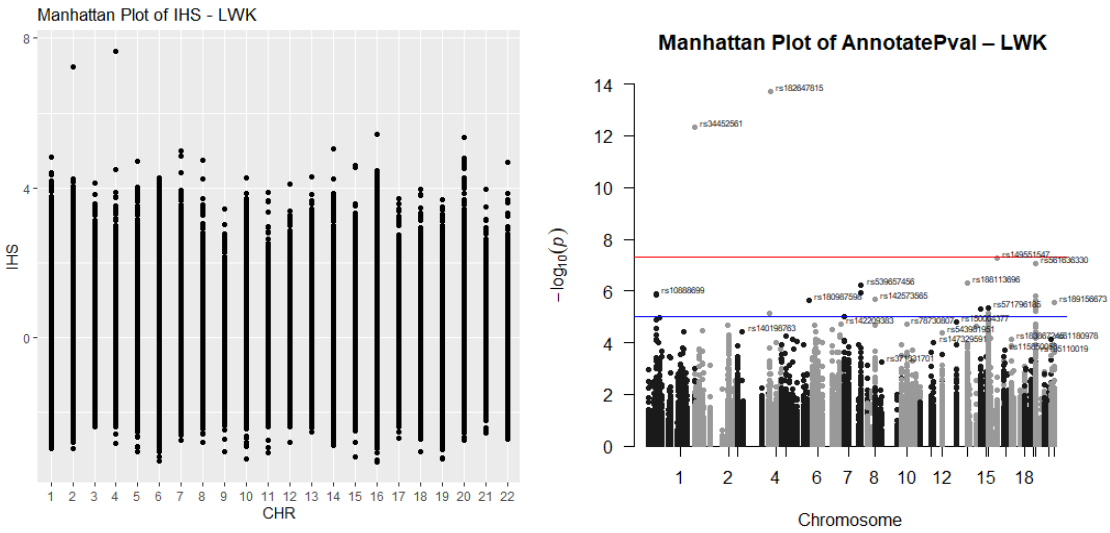

Figure 3.27 Plots of IHS values and associated P values obtained from IHS analysis results for 196 genes of the LWK subpopulation of the Africa population.

Table 3.1 Top 5 highest and lowest values of IHS analysis results for 196 genes of 14 populations.

POP	SNP	GENE	IHS	P	Variant Type
CEU	rs556200781	*ANKRD55*	7.89	2.8e-15	*ANKRD55* Upstream
	rs557147993	*MACROD2*	6.62	3.7e-11	Intron
	rs146230583	*LOC105374761_LOC105374764*	6.24	4.32e-10	*LOC105374764, TMEM17* Intron
	rs76917062	*LOC105378327_ALDH7A1P4*	5.56	2.67e-08	Intergenic
	rs16930451	*SVIL2P*	5.5	3.79e-08	*SVIL2P* Upstream / *LYZL2* 2KB Upstream
	rs549214602	*ITLN1*	-4.19	2.78e-05	*ITLN1* Downstream / *CD244* 2KB Upstream
	rs114990400	*ZBTB38*	-3.8	0.0001	*LOC124909441* Upstream
	rs78116042	*ZNF365*	-3.7	0.0002	Intron
	rs149518916	*MYO10*	-3.65	0.0002	Intron
	rs75315664	*PDGFB*	-3.62	0.0002	*PDGFB* Upstream
FIN	rs113848747	*TRIB1AL,LINC02964*	5.79	7.1e-09	Intron
	rs9562236	*ELF1*	5.78	7.34e-09	*SUGT1P3, TPTE2P5* Intron / *ELF1* Downstream
	rs542915899	*PRDM1_ LOC105377923*	5.3	1.18e-07	*PRDM1* 2KB Upstream
	rs76706325	*ZBTB38*	5.1	3.47e-07	Intron
	rs547084761	*SLAIN2*	5.01	5.45e-07	Intron
	rs57201378	*LOC101927745*	-4.79	1.66e-06	Intron
	rs138464066	*LINC01475*	-3.83	0.0001	*LINC01475* Downstream
	rs529797623	*DOCK7*	-3.71	0.0002	Intron
	rs189707864	*MACROD2*	-3.55	0.0003	Intron
	rs60699354	*ELF1*	-3.47	0.0005	*SUGT1P3, TPTE2P5* Intron / *ELF1* Downstream

(cont. on next page)

Table 3.1 (cont.)

GBR	rs551382470	*SPATA48_IKZF1*	5.94	2.91e-09	*SPATA48* Intron
	rs78391465	*LOC107984997_RNFT1P2*	5.3	1.19e-07	Intergenic
	rs527524959	*SATB1-AS1*	5.04	4.71e-07	*SATB1-AS1* Downstream
	rs527524959	*SATB1-AS1_ LOC107986066*	4.99	6.11e-07	*SATB1-AS1* Downstream
	rs535249657	*PHACTR2*	4.98	6.44e-07	Intron
	rs548894351	*TAGAP-AS1*	-3.83	0.0001	*LOC112267968* Intron
	rs573656789	*NLRP13*	-3.75	0.0001	*NLRP4* Intron / *NLRP13* Downstream
	rs62392933	*SLC22A23*	-3.73	0.0001	Intron
	rs188824258	*SLC22A23*	-3.68	0.0002	Intron
	rs143214265	*PNKD_TMBIM1*	-3.6	0.0003	*CATIP-AS2* Intron
IBS	rs111407155	*C17orf67*	6.78	1.17e-11	*LOC124904037* Intron
	rs9822628	*LINC01014*	6.78	1.24e-11	*LINC01014* Upstream
	rs531002475	*BANK1*	6.37	1.83e-10	Intron
	rs151333414	*HERC2*	6.33	2.52e-10	Intron
	rs149770768	*STAT4*	6.23	4.57e-10	Intron
	rs2694928	*FOXP2*	-4.16	3.17e-05	Intron
	rs281362	*AKAP10*	-3.99	6.74e-05	*AKAP10* Downstream
	rs538868878	*ITLN1*	-3.95	7.93e-05	*ITLN1* Downstream / *CD244* Upstream
	rs75695259	*TPPP2_NDRG2*	-3.93	8.43e-05	Intron
	rs112911768	*SVIL2P*	-3.71	0.0002	*SVIL2P* Upstream m
TSI	rs576516444	*RBX1_RPS9P2*	6.89	5.61e-12	Intergenic
	rs542144702	*ZNF365*	6.57	5.12e-11	Intron
	rs557147993	*MACROD2*	6.46	1.04e-10	Intron
	rs536913616	*MACROD2*	5.72	1.07e-08	Intron
	rs6975281	*C7orf33*	5.38	7.27e-08	Intron
	rs143637581	*SLC22A23*	-4.62	3.86e-06	Intron
	rs7523503	*MROH3P*	-4.62	4.13e-06	Intron
	rs150777522	*TAGAP-AS1*	-3.94	8.13e-05	*LOC112267968* Intron / *TAGAP-AS1* Downstream
	rs111250403	*TAGAP-AS1*	-3.88	0.0001	*LOC112267968* Intron / *TAGAP-AS1* Downstream
	rs74005028	*RBFOX1*	-3.8	0.0001	Intron
CDX	rs534404683	*LYRM4*	7.58	3.41e-14	Intron
	rs563956224	*LOC101928354*	7.54	4.58e-14	Intron
	rs6809353	*SATB1-AS1_ LOC107986066*	6.79	1.1e-11	Intergenic
	rs16829184	*PLCL1*	6.28	3.39e-10	*PLCL1* Downstream
	rs111802303	*PTPN2*	5.65	1.57e-08	*PTPN2* Upstream
	rs79577099	*PLSCR4*	-5.37	7.68e-08	Intron
	rs565484031	*GP2_UMOD*	-4.37	1.25e-05	*PDILT* Intron / *GP2_UMOD* Upstream
	rs534393924	*GP2_UMOD*	-4.37	1.25e-05	*PDILT* Intron / *GP2_UMOD* Upstream
	rs548949225	*GP2_UMOD*	-4.35	1.39e-05	*PDILT* Intron / *GP2_UMOD* Upstream
	rs147451109	*HERC2*	-4.25	2.12e-05	Intron
CHB	rs78351220	*LYRM4*	8.53	1.46e-17	Intron
	rs76414485	*APBB2*	7.05	1.73e-12	Intron
	rs191255390	*MUC19*	6.74	1.57e-11	Intron
	rs150060942	*CD244*	6.64	3.16e-11	*LY9* Intron / *CD244* Downstream
	rs576172846	*CPEB4*	6.37	1.92e-10	Intron
	rs191740154	*LOC107984997_ RNFT1P2*	-5.21	1.94e-07	Regulatory Region
	rs191998556	*ITLN1*	-4.77	1.85e-06	*ITLN1* Upstream
	rs1157874	*PLCL1*	-4.5	6.73e-06	*PLCL1* Downstream
	rs73256765	*RNASET2*	-4.49	6.96e-06	Intron
	rs139980316	*PRDM1*	-4.19	2.78e-05	*PRDM1* Downstream

(cont. on next page)

Table 3.1 (cont.)

CHS	rs11155322	*PHACTR2*	7.71	1.26e-14	Intron
	rs149304930	*TPPP2_NDRG2*	7.67	1.7e-14	*METTL17* Intron
	rs554349601	*KSR1*	6.66	2.74e-11	*LGALS9* 2KB Upstream
	rs16829184	*PLCL1*	6.4	1.51e-10	*PLCL1* Downstream
	rs183358230	*HERC2*	6.25	4.15e-10	Intron
	rs189716020	*SKAP2*	-6.05	1.43e-09	*SKAP2* Upstream
	rs184268581	*LOC107984997_ RNFT1P2*	-5.12	3.12e-07	Intergenic
	rs76410738	*PRDM1*	-4.93	8.22e-07	*PRDM1* Downstream
	rs148216333	*GPR35*	-4.91	9.11e-07	Intron
	rs117964613	*ITLN1*	-4.86	1.16e-06	*CD244* Intron / *ITLN1* Downstream
JPT	rs145618071	*TPPP2_NDRG2*	7.85	4.22e-15	*METTL17* Intron
	rs11155322	*PHACTR2*	7.7	1.36e-14	Intron
	rs17780048	*LINC02539_ WAKMAR2*	7.01	2.36e-12	WAKMAR2 Intron
	rs183358230	*HERC2*	6.86	6.66e-12	Intron
	rs553936332	*NFATC1*	6.8	1.03e-11	Intron
	rs1358748	*IL23R*	-4.69	2.72e-06	Intron
	rs6962370	*IKZF1*	-4.03	5.48e-05	Intron
	rs182990946	*NFATC1*	-3.85	0.0001	*NFATC1* Downstream
	rs192346909	*DOCK7*	-3.75	0.0001	Intron
	rs6962370	*SPATA48_IKZF1*	-3.72	0.0001	*IKZF1* Intron
KHV	rs28746174	*MAMSTR*	6.92	4.58e-12	*MAMSTR* Downstream
	rs528208513	*NCF4,NCF4-AS1*	6.13	8.71e-10	Intergenic
	rs533037864	*ZBTB38*	5.8	6.5e-09	Intron
	rs6601767	*LOC105376364_LINC02639*	5.33	9.76e-08	Regulatory Region
	rs151166226	*AIMP1P2_ TNFSF18*	5.32	1.02e-07	*TNFSF18* Downstream
	rs2346097	*HERC2*	-4.89	1.01e-06	Intron
	rs184243148	*HLA*	-4.15	3.35e-05	Intron
	rs184243148	*HLADQB1_ MTCO3P1*	-4.08	4.6e-05	Intron
	rs73894841	*LOC101927745*	-4.06	4.9e-05	Intron
	rs17101414	*LOC107984997_ RNFT1P2*	-4.05	5.15e-05	Intergenic
GWD	rs112606997	*SLC22A23*	5.03	4.86e-07	Intron
	rs2346096	*HERC2*	5.01	5.51e-07	Intron
	rs566241749	*UBE2L3*	4.98	6.4e-07	Intron
	rs532174800	*RBFOX1*	4.82	1.44e-06	Intron
	rs528219528	*RBFOX1*	4.82	1.44e-06	Intron
	rs563908371	*C17orf67*	-3.77	0.0001	*LOC124904037* Intron / *C17orf67* Upstream
	rs576242047	*TRA_TRAV12-2*	-3.2	0.0013	*TRA* Intron
	rs561094186	*RBFOX1*	-3.14	0.0016	Intron
	rs560287182	*TRA_TRAV12-2*	-3.03	0.0024	*TRA* Intron
	rs111880719	*LOC105379031*	-3.03	0.0024	Intron
YRI	rs563938935	*MACROD2*	5.94	2.78e-09	Intron
	rs537831762	*RBX1_RPS9P2*	5.25	1.49e-07	Intergenic
	rs116348755	*PRDM1_ LOC105377923*	4.98	6.37e-07	Intergenic
	rs533048806	*FAF1*	4.94	7.98e0-7	Intron
	rs140176643	*SBNO2*	4.81	1.47e-06	Intron
	rs554831273	*ZGPAT*	-5.97	2.34e-09	Intron
	rs576242047	*TRA_TRAV12-2*	-3.87	0.0001	*TRA* Intron
	rs142570222	*LOC112267902*	-3.67	0.0002	*HLA-C* Missense Variant
	rs148608512	*TRA_TRAV12-2*	-3.41	0.0006	*TRA* Intron
	rs568519785	*TRA_TRAV12-2*	-3.21	0.0013	*TRA* Intron

(cont. on next page)

Table 3.1 (cont.)

ESN	rs545219922	*IKZF3*	6.93	4.1e-12	Intron
	rs546313551	*CCNY*	6.53	6.45e-11	Intron
	rs542744238	*BSN*	6.31	2.78e-10	Intron
	rs566701406	*BTLN2*	6.07	1.3e-09	Intron
	rs9610589	*NCF4,NCF4-AS1*	5.57	2.48e-08	Intergenic
	rs532430659	*LOC107986537*	-3.77	0.0001	*LOC10798653* Downstream
	rs191039753	*LOC107986537*	-3.74	0.0001	Intron
	rs551613468	*LOC105379031*	-3.57	0.0003	Noncoding Transcript Exon
	rs190909134	*SMIM3*	-3.42	0.0006	*SMIM3* Upstream
	rs187648606	*IKZF3*	-3.41	0.0006	*IKZF3* Downstream
LWK	rs182647815	*SLAIN2*	7.65	1.94e-14	*SLAIN2* Upstream
	rs34452561	*DNMT3A*	7.23	4.76e-13	*DNMT3A* Downstream
	rs149551547	*NOD2*	5.43	5.5e-08	*SNX20* Intron / *NOD2* Upstream
	rs561636330	*MACROD2*	5.35	8.6e-08	Intron
	rs188113696	*TRA_TRAV12-2*	5.03	4.8e-07	*TRA* Intron
	rs565198015	*RBFOX1*	-3.33	0.0008	Intron
	rs186905057	*SLC22A23*	-3.29	0.001	*PSMG4* Intron
	rs149228109	*RBFOX1*	-3.27	0.001	Intron
	rs190610994	*SVIL2P*	-3.24	0.0011	*SVIL2P* Upstream /
	rs183271367	*SLC7A10_CEBPA*	-3.23	0.0012	Intergenic

(The "_" character indicates that the relevant SNP is between two genes. The "," character indicates that the relevant SNP is present in both genes. Red highlight represents duplicate variants.)

In the IHS analysis results of 196 genes from 14 subpopulations of European, East Asian and African populations, it was determined that there were common variants among populations within the highest and lowest 5 values (Table 3.1). No variants were identified in this data that mapped to SNPs (variants) associated with Crohn's Disease. The rs557147993 (*MACROD2*) variant was observed to be common in the CEU and TSI subpopulations of the European population. The rs527524959 (*SATB1-AS1*) variant was observed twice in the GBR subpopulation of the European population. The rs16829184 (*PLCL1*) variant was observed to be common in the CDX and CHS subpopulations of the East Asian population. The rs11155322 (*PHACTR2*) variant was observed to be common in the CHS and JPT subpopulation of the East Asian population. The rs183358230 (*HERC2*) variant was observed to be common in the CHS and JPT subpopulation of the East Asian population. The rs6962370 (*IKZF1*) variant was observed twice in the JPT subpopulation of the East Asian population. The rs184243148 (*HLA*) variant was observed twice in the KHV subpopulation belonging to the East Asian population. The rs576242047 (*TRA_TRAV12-2*) variant was observed to be common in the GWD and YRI subpopulation of the African population.

In the recent selection analysis of 14 subpopulations, genes with high (+/-) IHS values (Table 3.1) were frequently observed to be *HERC2*, *MACROD2*, *RBFOX1*, *TRA_TRAV12-2*, *SLC22A23*, *ITLN1*, *LOC107984997_RNFT1P2* genes. Assuming that these genes were subject to potential selection, the corresponding EHH plots were evaluated. In particular, large ancestral and derived haplotype structures were observed in the EHH plots of *HERC2*, *ITLN1* and *LOC107984997_RNFT1P2* genes (Appendix C).

Among the genes with high additive values in the IHS analyses given in Table 3.1, especially *ZBTB38* (rs76706325) (Figure C.2), *SLAIN2* (rs547084761) (Figure C. 2), *LOC107984997_RNFT1P2* (rs78391465) (Figure C.3), *CPEB4* (rs576172846) (Figure C.7), *MAMSTR* (rs28746174) (Figure C.10), *ZBTB38* (rs533037864) (Figure C.10) and *NCF4,NCF4-AS1* (rs528208513) (Figure C.10) genes showed rather large ancestral haplotype structures. Genes with high additive values in IHS analyses include *ITLN1* (rs549214602) (Figure C.1), *LOC101927745* (rs57201378) (Figure C.2), *MROH3P* (rs7523503) (Figure C.5), *PLSCR4* (rs79577099) (Figure C.6), *RNASET2* (rs73256765) (Figure C.7), *IL23R* (rs1358748) (Figure C. 9), *LOC101927745* (rs73894841) (Figure C.10) and *LOC107984997_RNFT1P2* (rs17101414) (Figure C.10) genes showed rather large derived haplotype structures.

Common genes with high additive values in both PCA and IHS analyses were examined and the corresponding EHH plots were evaluated assuming that these genes were subject to potential selection. The rs151166226 variant located around the *AIMP1P2* gene was observed to form a large ancestral haplotype structure (Figure C.10). The rs556200781 variant located around the *ANKRD55* gene was observed to form a large ancestral haplotype structure (Figure C.1). The rs533048806 variant in the *FAF1* gene was observed to form large ancestral haplotype structure (Figure C.12). The rs183358230 variant in the *HERC2* gene was observed to form a large ancestral haplotype structure (Figure C.8 and Figure C.9). The rs184243148 variant in the *HLA* gene was observed to form large derived haplotype structure (Figure C.10). The rs545219922 variant in the *IKZF3* gene was observed to form a large ancestral haplotype structure (Figure C.13). The rs1358748 variant in the *IL23R* gene was observed to form a large derived haplotype structure (Figure C.9). The rs76917062 variant around the *LOC105378327_ALDH7A1P4* gene was observed to form a large ancestral haplotype structure (Figure C.1). The rs191039753 variant in *LOC107986537* gene was observed to form large derived haplotype structure (Figure C.13).

The IHS analyses showed that not only European but also other world populations had significant IHS and P values, indicating that possible selection could be seen in these populations along with various genes. Thanks to gene-based analyses, it was revealed that there may be different variants that may be associated with CD, and that the signals coming from the environment of these variants and the various functions of the genes they are associated with may need to be investigated. While the genes supporting the highest population differentiation were especially *PUS10* and *PPBP_CXCL5*, *PPP5C*, *AIMP1P2*, *IRGM*, *OR2B11* and *IL10*, these genes were not frequently encountered as a result of gene-based analyses. In both SNP-based and gene-based analyses, genes with more than one variant, especially those obtained from the literature review and highly associated with CD, were not encountered very frequently as a result of these analyses. Genes and variants showing allele frequency variation and possible selection were frequently observed in genes less associated with CD.

3.4 Biological Pathway and Molecular Function Analyses

Molecular functions, biological processes, associated pathways and protein classes of genes associated with Crohn's Disease were examined using online tools such as GeneCard and PANTHER (Appendix B).[132,134] In particular, the structure and function of genes with common, frequently observed and significant values obtained from PCA and IHS analyses were examined and their interactions with each other were evaluated. Limited gene ontology data was obtained for some relevant genes.

AIMP1P2, *ALDH7A1P4* and *MROH3P* genes are associated with CD and are Pseudogenes.[134] Limited data about these genes could be found in various databases. *ANKRD55, CPEB4, FAF1, HERC2, HLA-DQB1, IKZF1, IKZF3, IL23R, ITLN1, MACROD2, MAMSTR, NCF4, PHACTR2, PLCL1, PLSCR4, RBFOX1, RNASET2, SLAIN2, SLC22A23, TRA, TRAV12-2* and *ZBTB38* genes are protein coding genes.[134] *SATB1-AS1, NCF4-AS1, LOC101927745, LOC105378327, LOC107984997* and *LOC107986537* genes are RNA genes and belong to the lncRNA class. Variants rs1736020 and rs2823286 in the *LOC101927745* gene have been associated with CD. The rs7076156 variant in the *LOC105378327* gene has been associated with CD.[134]

Diseases associated with *ANKRD55* adaptor protein include Juvenile Idiopathic Arthritis variants.[134] Diseases associated with *CPEB4* mRNA polyadenylation factor include Portal Hypertension and Retinitis Pigmentosa. The pathways involved include the glucocorticoid receptor pathway and the nuclear receptors meta-pathway. GO and STRING annotations for this gene reveal its role as a sequence-specific RNA binding protein. Specifically, it binds to the cytoplasmic polyadenylation element, a uridine-rich sequence element located within the mRNA 3'-UTR. RNA binding causes a distinct conformational change similar to the mechanism observed in the Venus flytrap. This protein is involved in regulating the activation of the unfolded protein response during adaptation to endoplasmic reticulum (ER) stress in the liver. It achieves this by maintaining continuous translation of mRNAs regulated by the cytoplasmic polyadenylation element under conditions where global protein synthesis is suppressed (Appendix B).[133,134] Diseases associated with the *FAF1* adaptor protein include Uterine Corpus Sarcoma and Joubert Syndrome 31 (JBTS31). Related pathways include apoptosis, survival FAS signaling cascades and the retinoblastoma gene in cancer. GO and STRING annotations for this gene target the DNA replication licensing factor CDT1 for degradation, allowing DNA replication forks to proceed. It potentiates but does not initiate FAS-induced apoptosis, protein kinase binding and protein domain-specific binding (Appendix B).[133,134] Diseases associated with the *HERC2* guanyl-nucleotide exchange factor include Intellectual Developmental Disorder Autosomal Recessive 38 (MRT38) and Skin/Hair/Eye Pigmentation Variation 1 (SHEP1). Pathways include homology-directed repair and MHC I-mediated antigen processing and presentation. GO annotations for this gene include ligase activity and ubiquitin protein ligase binding.[134] STRING annotations of this gene regulate ubiquitin binding, recruit sites of DNA damage, and facilitate recruitment of UBE2N and RNF8, which promote the formation of DNA damage-associated ubiquitin chains. It is involved in the binding specificity among UBE2N and RNF8. Plays a role in maintaining RNF168 levels.[133] Diseases associated with the *HLA-DQB1* major histocompatibility complex protein include Multiple Sclerosis and Creutzfeldt-Jakob Disease (CJD). Pathways involved include TCR signaling and phosphorylation of CD3 and TCR zeta chains. The GO and STRING annotations for this gene bind peptides derived from antigens that enter the endocytic pathway of antigen-presenting cells. These peptides are then presented on the cell surface for recognition by CD4 T cells. The peptide binding cleft accommodates peptides ranging from 10 to 30 residues in length. Peptides presented by MHC class II molecules are

produced predominantly through the degradation of proteins that reach the endocytic pathway, where they are subjected to processing by lysosomal proteases and other hydrolases (Appendix B).[133,134] The *IKZF1* and *IKZF3* genes are C2H2 zinc finger transcription factors and are important paralogs of each other. GO annotations associated with these two genes include DNA-binding transcription factor activity. *IKZF1*-related diseases include Common Variable Immunodeficiency-13 (CVID13) and Diamond-Blackfan Anemia-Like (DBAL). Pathways involved include *NOTCH3* and Pre-NOCH expression, processing and signaling (Appendix B).[134] STRING annotations for this gene include binding gamma-satellite DNA. It helps the development of both B and T lymphocytes. It binds to the enhancer of the CD3-delta gene and triggers its activation. It also acts as a repressor of the TDT gene during thymocyte differentiation. Its role in transcriptional regulation involves interaction with both HDAC-dependent and HDAC-independent complexes.[133] Diseases associated with *IKZF3* include Immunodeficiency 84 (IMD84) and Chronic Lymphocytic Leukemia (CLL). Pathways involved include IL2 signaling events mediated by STAT5 and NF-kappaB signaling (Appendix B).[134] In STRING annotations for *IKZF3* are involved in the regulation of lymphocyte differentiation, B cell differentiation, proliferation and maturation to the effector state. It is involved in modulating BCL2 expression and directing apoptosis in T cells in an IL2-dependent manner (Appendix B).[133] Diseases associated with the *IL23R* transmembrane signaling receptor include Inflammatory Bowel Disease 17 (IBD17) and Psoriasis 7. The pathways involved include Akt signaling and cytokine production by Th17 cells in cystic fibrosis using a mouse model. Among the GO and STRING annotations associated with this gene, it exhibits interleukin-12 receptor binding and interleukin-23 receptor activity. This gene aids in the activation of T cells, NK cells, and potentially specific macrophage/myeloid cells, potentially through activation of the Jak-Stat signaling cascade. IL23 plays important roles in both innate and adaptive immunity and may contribute to the immediate response to infections in peripheral tissues (Appendix B).[133,134] Diseases associated with *ITLN1* include Type 2 Diabetes Mellitus and Pleuropneumonia. Related pathways include the innate immune system and defensing. GO annotations related to this gene include carbohydrate binding.[134] STRING annotations related to this gene specifically recognize microbial carbohydrate chains in a calcium-dependent manner. It binds to microbial glycans containing a terminal acyclic 1,2-diol moiety, D-phosphoglycerol-modified glycans, including beta-linked D-galactofuranose, 3-deoxy-D-manno-oct-2-ulosonic acid and D-glycero-D-thalo-oct-2-

ulosonic acid. Binds to glycans of Gram-positive and Gram-negative bacteria, S.pneumonia, Y.pestis, P.mirabilis, including K.pneumoniae, and P.vulgaris. It does not bind human glycans.[133] Diseases associated with *MACROD2* include Hypogonadotropic Hypogonadism 21 (HH21) and Autism. GO annotations associated with this gene involve hydrolase activity targeting glucosyl bonds and deacetylase activity (Appendix B).[134] According to STRING annotations, this gene catalyzes the removal of ADP-ribose from aspartate and glutamate residues in proteins containing a single ADP-ribose moiety. However, it remains inactive against proteins carrying poly-ADP-ribose. Additionally, it deacetylates O-acetyl-ADP ribose, a signaling molecule generated through the deacetylation of acetylated lysine residues in histones and other proteins.[133] Diseases associated with the *MAMSTR* transcriptional coactivator include Mediastinum and Adult Teratoma. GO annotations related to this gene include nucleic acid binding and obsolete transcription factor activity, RNA polymerase II transcription factor binding.[134] Stimulates the transcriptional activity of MEF2C. Stimulates MYOD1 activity partially through MEF2, leading to increased skeletal muscle differentiation.[133] Diseases associated with *NCF4* adaptor protein include Autosomal Recessive Chronic Granulomatous Disease-3 (CGD3) and Chronic Granulomatous Disease (CGD). Among its related pathways are signaling by Rho GTPases and cellular responses to stimuli. GO annotations related to this gene include protein dimerization activity and phosphatidylinositol-3-phosphate binding (Appendix B).[134] It is a component of NADPH-oxidase, a complex enzyme system responsible for the oxidative burst, where electrons are transferred from NADPH to molecular oxygen, producing reactive oxidant intermediates. Its role for the assembly and/or activation of the NADPH-oxidase complex may be crucial.[133] Diseases associated with *PHACTR2* include Body Dysmorphic Disorder and Parkinson's Disease. One of its associated pathways involves the response to increased cytosolic calcium levels in platelets. GO annotations related to this gene include actin binding and protein phosphatase inhibitor activity.[134] Diseases associated with *PLCL1* include Glass Syndrome and Astigmatism. Included in its related pathways are Proton Pump Inhibitor and GPR40 Pathway. GO annotations related to this gene include calcium ion binding and phosphoric diester hydrolase activity.[134] *PLCL1* is involved in an intracellular signaling cascade that relies on inositol phospholipids. However, it does not show phospholipase C activity against phosphatidylinositol 4,5-bisphosphate and phosphatidylinositol. It plays a role as a component in the phospho-dependent endocytosis process of GABA A receptors. Additionally, it maintains GABA-

mediated synaptic inhibition by regulating receptor turnover. Abnormal expression of *PLCL1* may play a role in the development and progression of lung carcinoma. It also acts as an inhibitor of PPP1C.[133] Diseases associated with *PLSCR4* include Atypical Choroid Plexus Papilloma. Included in its related pathways is the complement cascade. GO annotations associated with this gene involve and binding to SH3 domains and calcium ion binding.[134] *PLSCR4* potentially facilitates ATP-independent bidirectional interlayer movement of phospholipids when calcium ions bind, leading to disruption of phospholipid asymmetry in the plasma membrane.[133] Diseases associated with *RBFOX1* RNA-binding protein include Colorectal Cancer and Benign Epilepsy With Centrotemporal Spikes. Included in its related pathways are transcription and transcriptional regulation mediated by MECP2. GO annotations associated with this gene include nucleic acid and nucleotide binding.[134] *RBFOX1* controls alternative splicing events by interacting with 5'-UGCAUGU-3' elements. It modulates alternative splicing of tissue-specific exons and differentially spliced exons during erythropoiesis.[133] Diseases associated with *RNASET2* endoribonuclease include Cystic Leukoencephalopathy Without Megalencephaly (LCWM) and Combined Oxidative Phosphorylation Deficiency 13. Among its related pathways are innate immune system and pathways of nucleic acid metabolism and innate immune sensing. GO annotations related to this gene include RNA binding and RNA endonuclease activity.[134] *RNASET2* displays ribonuclease activity, with increased activity under acidic pH conditions. It probably participates in the lysosomal degradation of ribosomal RNA and contributes to cellular RNA degradation processes.[133] GO annotations of *SLAIN2* include binding to the plus end of microtubules and regulating microtubule organization (Appendix B).[134] *SLAIN2* facilitates the initiation and extension of cytoplasmic microtubules. It is essential for maintaining the proper organization of the microtubule cytoskeleton throughout interphase.[133] Diseases associated with *SLC22A23* secondary carrier transporter include Wolf-Hirschhorn Syndrome (WHS) and Inflammatory Bowel Disease (IBD). GO annotations related to this gene include transmembrane transporter activity.[134] Diseases associated with *TRA* immunoglobulin receptor include Immunodeficiency 7 (IMD7) and Adult T-Cell Lymphoma Leukemia (T-ALL). Its related pathways include immune response mechanisms such as antigen presentation by MHC class II and NFAT involvement in the immune response. GO annotations related to this gene include enables signaling receptor activity, immune response and response to bacterium (Appendix B).[134] *TRAV12-2* immunoglobulin receptor involved in adaptive

immune response and immune system process.[134] Diseases associated with *ZBTB38* C2H2 zinc finger transcription factor include Kuru and Acromesomelic Dysplasia 2B. GO annotations related to this gene include protein homodimerization activity and nucleic acid binding (Appendix B).[134] STRING annotations associated with this gene indicate its role as a transcriptional regulator with dual DNA-binding specificity. It exhibits a higher affinity for methylated CpG dinucleotides in the consensus sequence 5'-CGCG-3', but it can also bind to E-box elements. Additionally, it can specifically bind to a single methyl-CpG pair and repress transcription in a methyl-CpG-dependent manner. This gene plays a critical role in regulating DNA replication and the stability of common fragile sites in a manner dependent on RBBP6 and MCM10. It represses the expression of MCM10, which is crucial for DNA replication.[133]

STRING analysis was used to understand protein-protein interactions between genes of significant value and commonalities that may have shown possible positive selection associated with Crohn's Disease, and the inter-protein interaction network was analyzed (Figure 3.28).

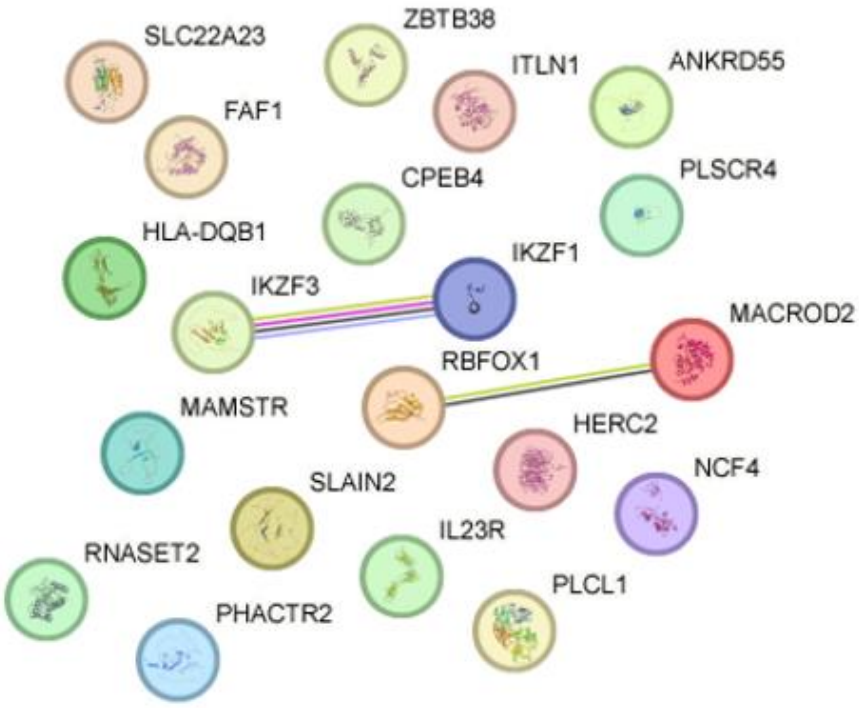

Figure 3.28 Protein-protein interaction network of genes showing possible positive selection associated with Crohn's Disease based on STRING analysis.

In this view, colored nodes represent the first shell of interactions and query proteins. The structures within the nodes are known and predicted as a 3D structure. A network was observed between *IKZF3* and *IKZF1* proteins. Edges in this relationship included experimentally determined (known interactions) (score 0.258), textmining (score 0.434), co-expression (score 0.258) and protein homology. Combined confidence of the functional interaction was 0.918 (very high). A network was observed between the *RBFOX1* and *MACROD2* proteins. The edges in this relationship included textmining (score 0.527) and co-expression (score 0.048). Combined confidence of the functional interaction was 0.539 (medium). Apart from these two interactions, no protein-protein interaction was observed between other genes. When the functional enrichments in the network were examined, they were associated with 3 diseases with significant rates of disease and genes. These are Arthritis, Autoimmune Disease and Primary Immunodeficiency Disease (false discovery rate was 0.0351).[133] With this association, it was observed that genes studied indeed had a significant interaction with autoimmune diseases.

The genetic diversity in Crohn's disease risk genes is influenced by a variety of evolutionary processes and it is undoubtedly difficult to understand the precise selection mechanisms underlying these processes. Future studies could improve the results obtained by using a larger set of sequence data with different intensities of CD prevalence, not only from groups with a high prevalence of Crohn's disease. Many of the genes analyzed in this study have been associated with other inflammatory diseases as risk factors. Thus, the findings not only apply to Crohn's disease but also to other inflammatory diseases. At the same time, in order to better understand and accurately assess the evolution of Crohn's disease from the past to the present, population sample size should be increased and more aDNA studies including genes and variants showing relevant selection should be conducted.

CHAPTER 4

CONCLUSION

Crohn’s Disease is a life-threatening health problem that is claimed to be related to modern diet, lifestyle and genetic predisposition. Crohn's-related genes can differentiate world populations and are frequently observed in Europe. CD associated alleles can also distinguish populations within Europe. CD-associated alleles, both derived and ancestral alleles (protective and susceptible), play a role in CD susceptibility.

When the variants associated with CD were analyzed, clear divergence was observed between modern metapopulations and their subpopulations, as well as between ancient and modern world populations. A large allele frequency variation was observed between ancient and modern populations on the European continent.

In the gene based recent selection analyses, it was observed that the genes with the highest IHS values in terms of 14 subpopulations were frequently *HERC2*, *MACROD2*, *RBFOX1*, *TRA* and *TRAV12-2* genes. Especially *HERC2*, *ITLN1* and *RNFT1P2* genes showed large haplotype structures. Gene-based analyzes include specific RNA and/or DNA binding (*CPEB4*, *MAMSTR*, *RBFOX1*, *RNASET2*, *ZBTB38*, *IKZF1* and *IKZF3*), protein binding activity (*FAF1*, *NCF4*, *SLC22A23*, *PHACTR2*, *IL23R*, *ZBTB38*, *ITLN1 and IKZF1*), ubiquitin and ligase activity (*HERC2*), adaptive immune response (*HLA-DQB1*, *IL23R* and *TRAV12-2*), innate immune system activity (*IL23R*, *ITLN1* and *RNASET2*), immune response (*TRA*), DNA-binding transcription factor activity (*ZBTB38*, *IKZF1* and *IKZF3*), hydrolase and/or deacetylase activity (*PLCL1* and *MACROD2*), calcium ion binding (*PLSCR4*, *ITLN1* and *PLCL1*), microtubule activity (*SLAIN2*), and other molecular processes.

Since these genes are related to the immune system and show a possible selection signature, they may be responsible for disease susceptibility and progression. Identifying specific genes and risk loci responsible for the disease and applying methods to compare the allele frequencies of these relevant variants in historical and current populations will provide insight into disease biology and potential treatment targets.

REFERENCES

1. Yeshi, K.; Ruscher, R.; Hunter, L.; Daly, N. L.; Loukas, A.; Wangchuk, P. Revisiting Inflammatory Bowel Disease: Pathology, Treatments, Challenges and Emerging Therapeutics Including Drug Leads from Natural Products. *J. Clin. Med.* **2020**, *9* (5), 1273. https://doi.org/10.3390/jcm9051273.

2. Crohn, B. B.; Ginzburg, L.; Oppenheimer, G. D. Regional Ileitis: A Pathologic and Clinical Entity. *J. Am. Med. Assoc.* **1932**, *99* (16), 1323–1329. https://doi.org/10.1001/jama.1984.03340250053024.

3. Gajendran, M.; Loganathan, P.; Catinella, A. P.; Hashash, J. G. A Comprehensive Review and Update on Crohn's Disease. *Disease-a-month* **2018**, *64* (2), 20–57. https://doi.org/10.1016/j.disamonth.2017.07.001.

4. Roda, G.; Chien Ng, S.; Kotze, P. G.; Argollo, M.; Panaccione, R.; Spinelli, A.; Kaser, A.; Peyrin-Biroulet, L.; Danese, S. Crohn's Disease. *Nat. Rev. Dis. Prim.* **2020**, *6* (1), 22. https://doi.org/10.1038/s41572-020-0156-2.

5. Molodecky, N. A.; Soon, S.; Rabi, D. M.; Ghali, W. A.; Ferris, M.; Chernoff, G.; Benchimol, E. I.; Panaccione, R.; Ghosh, S.; Barkema, H. W. Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases with Time, Based on Systematic Review. *Gastroenterology* **2012**, *142* (1), 46–54. https://doi.org/10.1053/j.gastro.2011.10.001.

6. Ng, S. C.; Shi, H. Y.; Hamidi, N.; Underwood, F. E.; Tang, W.; Benchimol, E. I.; Panaccione, R.; Ghosh, S.; Wu, J. C. Y.; Chan, F. K. L. Worldwide Incidence and Prevalence of Inflammatory Bowel Disease in the 21st Century: A Systematic Review of Population-Based Studies. *Lancet* **2017**, *390* (10114), 2769–2778. https://doi.org/10.1016/S0140-6736(17)32448-0.

7. Ananthakrishnan, A. N. Epidemiology and Risk Factors for IBD. *Nat. Rev. Gastroenterol. Hepatol.* **2015**, *12* (4), 205–217. https://doi.org/10.1038/nrgastro.2015.34.

8. Lichtenstein, G. R.; Loftus, E. V; Isaacs, K. L.; Regueiro, M. D.; Gerson, L. B.; Sands, B. E. ACG Clinical Guideline: Management of Crohn's Disease in Adults. *Off. J. Am. Coll. Gastroenterol. ACG* **2018**, *113* (4), 481–517. https://doi.org/10.1038/ajg.2018.27.

9. Farmer, R. G.; Easley, K. A.; Rankin, G. B. Clinical Patterns, Natural History, and Progression of Ulcerative Colitis: A Long-Term Follow-up of 1116 Patients. *Dig. Dis. Sci.* **1993**, *38*, 1137–1146. https://doi.org/10.1007/BF01295733.

10. Kefalas, C. H. Gastroduodenal Crohn's Disease. In *Baylor University Medical Center Proceedings*; Taylor & Francis, 2003; Vol. 16, pp 147–151. https://doi.org/10.1080/08998280.2003.11927896.

11. Poggioli, G.; Stocchi, L.; Laureti, S.; Selleri, S.; Marra, C.; Salone, M. C.; Cavallari, A. Duodenal Involvement of Crohn's Disease: Three Different Clinicopathologic Patterns. *Dis. colon rectum* **1997**, *40*, 179–183. https://doi.org/10.1007/BF02054984.

12. Higuero, T.; Merle, C.; Thiéfin, G.; Coussinet, S.; Jolly, D.; Diebold, M.-D.; Zeitoun, P.; Cadiot, G. Jejunoileal Crohn's Disease: A Case-Control Study. *Gastroentérologie Clin. Biol.* **2004**, *28* (2), 160–166. https://doi.org/10.1016/S0399-8320(04)94871-3.

13. Farmer, M.; Petras, R. E.; Hunt, L. E.; Janosky, J. E.; Galandiuk, S. The Importance of Diagnostic Accuracy in Colonic Inflammatory Bowel Disease. *Am. J. Gastroenterol.* **2000**, *95* (11), 3184–3188. https://doi.org/10.1111/j.1572-0241.2000.03199.x.

14. Yantiss, R. K.; Odze, R. D. Diagnostic Difficulties in Inflammatory Bowel Disease Pathology. *Histopathology* **2006**, *48* (2), 116–132. https://doi.org/10.1111/j.1365-2559.2005.02248.x.

15. Ye, B. D.; Yang, S.-K.; Cho, Y. K.; Park, S. H.; Yang, D.-H.; Yoon, S. M.; Kim, K. J.; Byeon, J.-S.; Myung, S.-J.; Yu, C. S. Clinical Features and Long-Term Prognosis of Crohn's Disease in Korea. *Scand. J. Gastroenterol.* **2010**, *45* (10), 1178–1185. https://doi.org/10.3109/00365521.2010.497936.

16. Lee, K.-M.; Lee, J. M. Crohn's Disease in Korea: Past, Present, and Future. *Korean J. Intern. Med.* **2014**, *29* (5), 558. http://dx.doi.org/10.3904/kjim.2014.29.5.558.

17. Veauthier, B.; Hornecker, J. R. Crohn's Disease: Diagnosis and Management. *Am. Fam. Physician* **2018**, *98* (11), 661–669.

18. Torres, J.; Mehandru, S.; Colombel, J.-F.; Peyrin-Biroulet, L. Crohn's Disease. *Lancet* **2017**, *389* (10080), 1741–1755. https://doi.org/10.1016/S0140-6736(16)31711-1.

19. Soon, I. S.; Molodecky, N. A.; Rabi, D. M.; Ghali, W. A.; Barkema, H. W.; Kaplan, G. G. The Relationship between Urban Environment and the Inflammatory Bowel Diseases: A Systematic Review and Meta-Analysis. *BMC Gastroenterol.* **2012**, *12* (1), 1–14. https://doi.org/10.1186/1471-230X-12-51.

20. Cholapranee, A.; Ananthakrishnan, A. N. Environmental Hygiene and Risk of Inflammatory Bowel Diseases: A Systematic Review and Meta-Analysis. *Inflamm. Bowel Dis.* **2016**, *22* (9), 2191–2199. https://doi.org/10.1097/MIB.0000000000000852.

21. Mahid, S. S.; Minor, K. S.; Soto, R. E.; Hornung, C. A.; Galandiuk, S. Smoking and Inflammatory Bowel Disease: A Meta-Analysis. In *Mayo Clinic Proceedings*; Elsevier, 2006; Vol. 81, pp 1462–1471. https://doi.org/10.4065/81.11.1462.

22. Cosnes, J.; Nion-larmurier, I.; Afchain, P.; Beaugerie, L.; Gendre, J. Gender Differences in the Response of Colitis to Smoking. *Clin. Gastroenterol. Hepatol.* **2004**, *2* (1), 41–48. https://doi.org/10.1016/S1542-3565(03)00290-8.

23. Hatoum, O. A.; Heidemann, J. A. N.; Binion, D. G. The Intestinal Microvasculature as a Therapeutic Target in Inflammatory Bowel Disease. *Ann. N. Y. Acad. Sci.* **2006**, *1072* (1), 78–97. https://doi.org/10.1196/annals.1326.003.

24. Biedermann, L.; Brülisauer, K.; Zeitz, J.; Frei, P.; Scharl, M.; Vavricka, S. R.; Fried, M.; Loessner, M. J.; Rogler, G.; Schuppler, M. Smoking Cessation Alters Intestinal Microbiota: Insights from Quantitative Investigations on Human Fecal Samples Using FISH. *Inflamm. Bowel Dis.* **2014**, *20* (9), 1496–1501. https://doi.org/10.1097/MIB.0000000000000129.

25. Piovani, D.; Danese, S.; Peyrin-Biroulet, L.; Nikolopoulos, G. K.; Lytras, T.; Bonovas, S. Environmental Risk Factors for Inflammatory Bowel Diseases: An Umbrella Review of Meta-Analyses. *Gastroenterology* **2019**, *157* (3), 647–659. https://doi.org/10.1053/j.gastro.2019.04.016.

26. Odes, H. S.; Fich, A.; Reif, S.; Halak, A.; Lavy, A.; Keter, D.; Eliakim, R.; Paz, J.; Broide, E.; Niv, Y.; Ron, Y.; Villa, Y.; Arber, N.; Gilat, T. Effects of Current Cigarette Smoking on Clinical Course of Crohn's Disease and Ulcerative Colitis. *Dig. Dis. Sci.* **2001**, *46* (8), 1717–1721. https://doi.org/10.1023/A:1010609722315.

27. Ng, S. C.; Tang, W.; Leong, R. W.; Chen, M.; Ko, Y.; Studd, C.; Niewiadomski, O.; Bell, S.; Kamm, M. A.; de Silva, H. J. Environmental Risk Factors in Inflammatory Bowel Disease: A Population-Based Case-Control Study in Asia-

Pacific. *Gut* **2015**, *64* (7), 1063–1071. https://doi.org/10.1136/gutjnl-2014-307410.

28. Bernstein, C. N.; Shanahan, F. Disorders of a Modern Lifestyle: Reconciling the Epidemiology of Inflammatory Bowel Diseases. *Gut* **2008**, *57* (9), 1185–1191. https://doi.org/10.1136/gut.2007.122143.

29. Hashash, J. G.; Chintamaneni, P.; Ramos Rivers, C. M.; Koutroubakis, I. E.; Regueiro, M. D.; Baidoo, L.; Swoger, J. M.; Barrie, A.; Schwartz, M.; Dunn, M. A.; Binion, D. G. Patterns of Antibiotic Exposure and Clinical Disease Activity in Inflammatory Bowel Disease: A 4-Year Prospective Study. *Inflamm. Bowel Dis.* **2015**, *21* (11), 2576–2582. https://doi.org/10.1097/MIB.0000000000000534.

30. Shaw, S. Y.; Blanchard, J. F.; Bernstein, C. N. Association Between the Use of Antibiotics in the First Year of Life and Pediatric Inflammatory Bowel Disease. *Off. J. Am. Coll. Gastroenterol. | ACG* **2010**, *105* (12). https://doi.org/10.1038/ajg.2010.398.

31. Penders, J.; Thijs, C.; Vink, C.; Stelma, F. F.; Snijders, B.; Kummeling, I.; van den Brandt, P. A.; Stobberingh, E. E. Factors Influencing the Composition of the Intestinal Microbiota in Early Infancy. *Pediatrics* **2006**, *118* (2), 511–521. https://doi.org/10.1542/peds.2005-2824.

32. Ungaro, R.; Bernstein, C. N.; Gearry, R.; Hviid, A.; Kolho, K.-L.; Kronman, M. P.; Shaw, S.; Van Kruiningen, H.; Colombel, J.-F.; Atreja, A. Antibiotics Associated with Increased Risk of New-Onset Crohn's Disease but Not Ulcerative Colitis: A Meta-Analysis. *Off. J. Am. Coll. Gastroenterol. ACG* **2014**, *109* (11), 1728–1738. https://doi.org/10.1038/ajg.2014.246.

33. Ungaro, R.; Chang, H. L.; Cote-Daigneaut, J.; Mehandru, S.; Atreja, A.; Colombel, J.-F. Statins Associated with Decreased Risk of New Onset Inflammatory Bowel Disease. *Off. J. Am. Coll. Gastroenterol. ACG* **2016**, *111* (10), 1416–1423. https://doi.org/10.1038/ajg.2016.233.

34. Gevers, D.; Kugathasan, S.; Denson, L. A.; Vázquez-Baeza, Y.; Van Treuren, W.; Ren, B.; Schwager, E.; Knights, D.; Song, S. J.; Yassour, M.; Morgan, X. C.; Kostic, A. D.; Luo, C.; González, A.; McDonald, D.; Haberman, Y.; Walters, T.; Baker, S.; Rosh, J.; Stephens, M.; Heyman, M.; Markowitz, J.; Baldassano, R.; Griffiths, A.; Sylvester, F.; Mack, D.; Kim, S.; Crandall, W.; Hyams, J.; Huttenhower, C.; Knight, R.; Xavier, R. J. The Treatment-Naive Microbiome in New-Onset Crohn’s Disease. *Cell Host Microbe* **2014**, *15* (3), 382–392. https://doi.org/10.1016/j.chom.2014.02.005.

35. Singh, S.; Graff, L. A.; Bernstein, C. N. Do NSAIDs, Antibiotics, Infections, or Stress Trigger Flares in IBD? *Off. J. Am. Coll. Gastroenterol. | ACG* **2009**, *104* (5), 1298-1313. https://doi.org/0.1038/ajg.2009.15.

36. Radford-Smith, G. L. What Is the Importance of Appendectomy in the Natural History of IBD? *Inflamm. Bowel Dis.* **2008**, *14* (suppl_2), S72–S74. https://doi.org/10.1002/ibd.20623.

37. Radford-Smith, G. L.; Edwards, J. E.; Purdie, D. M.; Pandeya, N.; Watson, M.; Martin, N. G.; Green, A.; Newman, B.; Florin, T. H. J. Protective Role of Appendicectomy on Onset and Severity of Ulcerative Colitis and Crohn's Disease. *Gut* **2002**, *51* (6), 808–813. https://doi.org/10.1136/gut.51.6.808.

38. Andersson, R. E.; Olaison, G.; Tysk, C.; Ekbom, A. Appendectomy Is Followed by Increased Risk of Crohn's Disease. *Gastroenterology* **2003**, *124* (1), 40–46. https://doi.org/10.1053/gast.2003.50021.

39. Ananthakrishnan, A. N.; Khalili, H.; Higuchi, L. M.; Bao, Y.; Korzenik, J. R.; Giovannucci, E. L.; Richter, J. M.; Fuchs, C. S.; Chan, A. T. Higher Predicted Vitamin D Status Is Associated with Reduced Risk of Crohn's Disease. *Gastroenterology* **2012**, *142* (3), 482–489. https://doi.org/10.1053/j.gastro.2011.11.040.

40. Froicu, M.; Cantorna, M. T. Vitamin D and the Vitamin D Receptor Are Critical for Control of the Innate Immune Response to Colonic Injury. *BMC Immunol.* **2007**, *8* (1), 1–11. https://doi.org/10.1186/1471-2172-8-5.

41. Cantorna, M. T.; Munsick, C.; Bemiss, C.; Mahon, B. D. 1, 25-Dihydroxycholecalciferol Prevents and Ameliorates Symptoms of Experimental Murine Inflammatory Bowel Disease. *J. Nutr.* **2000**, *130* (11), 2648–2652. https://doi.org/10.1093/jn/130.11.2648.

42. Froicu, M.; Weaver, V.; Wynn, T. A.; McDowell, M. A.; Welsh, J. E.; Cantorna, M. T. A Crucial Role for the Vitamin D Receptor in Experimental Inflammatory Bowel Diseases. *Mol. Endocrinol.* **2003**, *17* (12), 2386–2392. https://doi.org/10.1210/me.2003-0281.

43. Ananthakrishnan, A. N.; Cagan, A.; Gainer, V. S.; Cai, T.; Cheng, S.-C.; Savova, G.; Chen, P.; Szolovits, P.; Xia, Z.; De Jager, P. L. Normalization of Plasma 25-Hydroxy Vitamin D Is Associated with Reduced Risk of Surgery in Crohn's Disease. *Inflamm. Bowel Dis.* **2013**, *19* (9), 1921–1927. https://doi.org/10.1097/MIB.0b013e3182902ad9.

44. Haase, H.; Rink, L. Zinc Signals and Immune Function. *Biofactors* **2014**, *40* (1), 27–40. https://doi.org/10.1002/biof.1114.

45. El-Tawil, A. M. Zinc Supplementation Tightens Leaky Gut in Crohn's Disease. *Inflamm. Bowel Dis.* **2012**, *18* (2), E399–E399. https://doi.org/10.1002/ibd.21926.

46. Aamodt, G.; Bukholm, G.; Jahnsen, J.; Moum, B.; Vatn, M. H.; Group, I. S. The Association between Water Supply and Inflammatory Bowel Disease Based on a 1990–1993 Cohort Study in Southeastern Norway. *Am. J. Epidemiol.* **2008**, *168* (9), 1065–1072. https://doi.org/10.1093/aje/kwn218.

47. Ananthakrishnan, A. N.; Long, M. D.; Martin, C. F.; Sandler, R. S.; Kappelman, M. D. Sleep Disturbance and Risk of Active Disease in Patients with Crohn's Disease and Ulcerative Colitis. *Clin. Gastroenterol. Hepatol.* **2013**, *11* (8), 965–971. https://doi.org/10.1016/j.cgh.2013.01.021.

48. Ananthakrishnan, A. N.; Khalili, H.; Konijeti, G. G.; Higuchi, L. M.; de Silva, P.; Fuchs, C. S.; Richter, J. M.; Schernhammer, E. S.; Chan, A. T. Sleep Duration Affects Risk for Ulcerative Colitis: A Prospective Cohort Study. *Clin. Gastroenterol. Hepatol.* **2014**, *12* (11), 1879–1886. https://doi.org/10.1016/j.cgh.2014.04.021.

49. Uemura, R.; Fujiwara, Y.; Iwakura, N.; Shiba, M.; Watanabe, K.; Kamata, N.; Yamagami, H.; Tanigawa, T.; Watanabe, T.; Tominaga, K.; Arakawa, T. Sleep Disturbances in Japanese Patients with Inflammatory Bowel Disease and Their Impact on Disease Flare. *Springerplus* **2016**, *5* (1), 1792. https://doi.org/10.1186/s40064-016-3408-6.

50. Ananthakrishnan, A. N.; Khalili, H.; Pan, A.; Higuchi, L. M.; de Silva, P.; Richter, J. M.; Fuchs, C. S.; Chan, A. T. Association between Depressive Symptoms and Incidence of Crohn's Disease and Ulcerative Colitis: Results from the Nurses' Health Study. *Clin. Gastroenterol. Hepatol.* **2013**, *11* (1), 57–62. https://doi.org/10.1016/j.cgh.2012.08.032.

51. Bonaz, B. L.; Bernstein, C. N. Brain-Gut Interactions in Inflammatory Bowel Disease. *Gastroenterology* **2013**, *144* (1), 36–49. https://doi.org/10.1053/j.gastro.2012.10.003.

52. Bitton, A.; Dobkin, P. L.; Edwardes, M. D.; Sewitch, M. J.; Meddings, J. B.; Rawal, S.; Cohen, A.; Vermeire, S.; Dufresne, L.; Franchimont, D. Predicting Relapse in Crohn's Disease: A Biopsychosocial Model. *Gut* **2008**, *57* (10), 1386–1392. https://doi.org/10.1136/gut.2007.134817.

53. Sonnenberg, A. Occupational Distribution of Inflammatory Bowel Disease among German Employees. *Gut* **1990**, *31* (9), 1037–1040. https://doi.org/10.1136/gut.31.9.1037.

54. Khalili, H.; Ananthakrishnan, A. N.; Konijeti, G. G.; Liao, X.; Higuchi, L. M.; Fuchs, C. S.; Spiegelman, D.; Richter, J. M.; Korzenik, J. R.; Chan, A. T. Physical Activity and Risk of Inflammatory Bowel Disease: Prospective Study from the Nurses' Health Study Cohorts. *Bmj* **2013**, *347*. https://doi.org/10.1136/bmj.f6633.

55. Kostic, A. D.; Xavier, R. J.; Gevers, D. The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead. *Gastroenterology* **2014**, *146* (6), 1489–1499. https://doi.org/10.1053/j.gastro.2014.02.009.

56. Darfeuille-Michaud, A.; Boudeau, J.; Bulois, P.; Neut, C.; Glasser, A.-L.; Barnich, N.; Bringer, M.-A.; Swidsinski, A.; Beaugerie, L.; Colombel, J.-F. High Prevalence of Adherent-Invasive Escherichia Coli Associated with Ileal Mucosa in Crohn's Disease. *Gastroenterology* **2004**, *127* (2), 412–421. https://doi.org/10.1053/j.gastro.2004.04.061.

57. Lapaquette, P.; Glasser, A.; Huett, A.; Xavier, R. J.; Darfeuille-Michaud, A. Crohn's Disease-associated Adherent-invasive E. Coli Are Selectively Favoured by Impaired Autophagy to Replicate Intracellularly. *Cell. Microbiol.* **2010**, *12* (1), 99–113. https://doi.org/10.1111/j.1462-5822.2009.01381.x.

58. Sokol, H.; Pigneur, B.; Watterlot, L.; Lakhdari, O.; Bermúdez-Humarán, L. G.; Gratadoux, J.-J.; Blugeon, S.; Bridonneau, C.; Furet, J.-P.; Corthier, G. Faecalibacterium Prausnitzii Is an Anti-Inflammatory Commensal Bacterium Identified by Gut Microbiota Analysis of Crohn Disease Patients. *Proc. Natl. Acad. Sci.* **2008**, *105* (43), 16731–16736. https://doi.org/10.1073/pnas.0804812105.

59. Quévrain, E.; Maubert, M. A.; Michon, C.; Chain, F.; Marquant, R.; Tailhades, J.; Miquel, S.; Carlier, L.; Bermúdez-Humarán, L. G.; Pigneur, B. Identification of an Anti-Inflammatory Protein from Faecalibacterium Prausnitzii, a Commensal Bacterium Deficient in Crohn's Disease. *Gut* **2016**, *65* (3), 415–425. https://doi.org/10.1136/gutjnl-2014-307649.

60. Morgan, X. C.; Tickle, T. L.; Sokol, H.; Gevers, D.; Devaney, K. L.; Ward, D. V; Reyes, J. A.; Shah, S. A.; LeLeiko, N.; Snapper, S. B. Dysfunction of the Intestinal Microbiome in Inflammatory Bowel Disease and Treatment. *Genome Biol.* **2012**, *13* (9), 1–18. https://doi.org/10.1186/gb-2012-13-9-r79.

61. David, L. A.; Maurice, C. F.; Carmody, R. N.; Gootenberg, D. B.; Button, J. E.;

Wolfe, B. E.; Ling, A. V; Devlin, A. S.; Varma, Y.; Fischbach, M. A. Diet Rapidly and Reproducibly Alters the Human Gut Microbiome. *Nature* **2014**, *505* (7484), 559–563. https://doi.org/10.1038/nature12820.

62. Loftus, E. V; Silverstein, M. D.; Sandborn, W. J.; Tremaine, W. J.; Harmsen, W. S.; Zinsmeister, A. R. Ulcerative Colitis in Olmsted County, Minnesota, 1940–1993: Incidence, Prevalence, and Survival. *Gut* **2000**, *46* (3), 336–343. https://doi.org/10.1136/gut.46.3.336.

63. Ananthakrishnan, A. N.; Khalili, H.; Konijeti, G. G.; Higuchi, L. M.; De Silva, P.; Korzenik, J. R.; Fuchs, C. S.; Willett, W. C.; Richter, J. M.; Chan, A. T. A Prospective Study of Long-Term Intake of Dietary Fiber and Risk of Crohn's Disease and Ulcerative Colitis. *Gastroenterology* **2013**, *145* (5), 970–977. https://doi.org/10.1053/j.gastro.2013.07.050.

64. Amre, D. K.; D'souza, S.; Morgan, K.; Seidman, G.; Lambrette, P.; Grimard, G.; Israel, D.; Mack, D.; Ghadirian, P.; Deslandres, C. Imbalances in Dietary Consumption of Fatty Acids, Vegetables, and Fruits Are Associated with Risk for Crohn's Disease in Children. *Off. J. Am. Coll. Gastroenterol. ACG* **2007**, *102* (9), 2016–2025. https://doi.org/10.1111/j.1572-0241.2007.01411.x.

65. Das, B.; Ghosh, T. S.; Kedia, S.; Rampal, R.; Saxena, S.; Bag, S.; Mitra, R.; Dayal, M.; Mehta, O.; Surendranath, A. Analysis of the Gut Microbiome of Rural and Urban Healthy Indians Living in Sea Level and High Altitude Areas. *Sci. Rep.* **2018**, *8* (1), 10104. https://doi.org/10.1038/s41598-018-28550-3.

66. Zoetendal, E. G.; Puylaert, P. G. B.; Ou, J.; Vipperla, K.; Brouard, F. M.; Ruder, E. H.; Newton, K.; Carbonero, F.; Gaskins, H. R.; de Vos, W. M. Distinct Microbiotas Are Present in Urban and Rural Native South Africans, and in African Americans. *Gastroenterology* **2013**, *144* (5), 347. https://doi.org/10.1016/S0016-5085(13)61277-9.

67. Tsianos, E. V; Katsanos, K. H.; Tsianos, V. E. Role of Genetics in the Diagnosis and Prognosis of Crohn's Disease. *World J. Gastroenterol. WJG* **2012**, *18* (2), 105. https://doi.org/10.3748/wjg.v18.i2.105.

68. Halme, L.; Paavola-Sakki, P.; Turunen, U.; Lappalainen, M.; Färkkilä, M.; Kontula, K. Family and Twin Studies in Inflammatory Bowel Disease. *World J. Gastroenterol. WJG* **2006**, *12* (23), 3668. https://doi.org/10.3748/wjg.v12.i23.3668.

69. Yang, H.; McElree, C.; Roth, M. P.; Shanahan, F.; Targan, S. R.; Rotter, J. I.

Familial Empirical Risks for Inflammatory Bowel Disease: Differences between Jews and Non-Jews. *Gut* **1993**, *34* (4), 517–524. https://doi.org/10.1136/gut.34.4.517.

70. Abraham, C.; Cho, J. H. Mechanisms of Disease. *N Engl J Med* **2009**, *361* (21), 2066–2078.

71. Moller, F. T.; Andersen, V.; Wohlfahrt, J.; Jess, T. Familial Risk of Inflammatory Bowel Disease: A Population-Based Cohort Study 1977–2011. *Off. J. Am. Coll. Gastroenterol. ACG* **2015**, *110* (4), 564–571. https://doi.org/10.1038/ajg.2015.50.

72. Brant, S. R. Update on the Heritability of Inflammatory Bowel Disease: The Importance of Twin Studies. *Inflammatory bowel diseases*. Oxford University Press Oxford, UK 2011, pp 1–5. https://doi.org/10.1002/ibd.21385.

73. Orholm, M.; Binder, V.; Sørensen, T. I. A.; Rasmussen, L. P.; Kyvik, K. O. Concordance of Inflammatory Bowel Disease among Danish Twins: Results of a Nationwide Study. *Scand. J. Gastroenterol.* **2000**, *35* (10), 1075–1081. https://doi.org/10.1080/003655200451207.

74. Thompson, N. P.; Driscoll, R.; Pounder, R. E.; Wakefield, A. J. Genetics versus Environment in Inflammatory Bowel Disease: Results of a British Twin Study. *Bmj* **1996**, *312* (7023), 95–96. https://doi.org/10.1136/bmj.312.7023.95.

75. Halfvarson, J.; Bodin, L.; Tysk, C.; Lindberg, E. V. A.; Järnerot, G. Inflammatory Bowel Disease in a Swedish Twin Cohort: A Long-Term Follow-up of Concordance and Clinical Characteristics. *Gastroenterology* **2003**, *124* (7), 1767–1773. https://doi.org/10.1016/S0016-5085(03)00385-8.

76. Huang, H.; Fang, M.; Jostins, L.; Umićević Mirkov, M.; Boucher, G.; Anderson, C. A.; Andersen, V.; Cleynen, I.; Cortes, A.; Crins, F. Fine-Mapping Inflammatory Bowel Disease Loci to Single-Variant Resolution. *Nature* **2017**, *547* (7662), 173–178. https://doi.org/10.1038/nature22969.

77. Hugot, J.-P.; Laurent-Puig, P.; Gower-Rousseau, C.; Olson, J. M.; Lee, J. C.; Beaugerie, L.; Naom, I.; Dupas, J.-L.; Van Gossum, A.; Digestives, G. d'Etude T. des A. I. Mapping of a Susceptibility Locus for Crohn's Disease on Chromosome 16. *Nature* **1996**, *379* (6568), 821–823. https://doi.org/10.1038/379821a0.

78. Hugot, J.-P.; Chamaillard, M.; Zouali, H.; Lesage, S.; Cézard, J.-P.; Belaiche, J.; Almer, S.; Tysk, C.; O'Morain, C. A.; Gassull, M. Association of NOD2 Leucine-

Rich Repeat Variants with Susceptibility to Crohn's Disease. *Nature* **2001**, *411* (6837), 599–603. https://doi.org/10.1038/35079107.

79. Ogura, Y.; Bonen, D. K.; Inohara, N.; Nicolae, D. L.; Chen, F. F.; Ramos, R.; Britton, H.; Moran, T.; Karaliuskas, R.; Duerr, R. H. A Frameshift Mutation in NOD2 Associated with Susceptibility to Crohn's Disease. *Nature* **2001**, *411* (6837), 603–606. https://doi.org/10.1038/35079114.

80. Jostins, L.; Ripke, S.; Weersma, R. K.; Duerr, R. H.; McGovern, D. P.; Hui, K. Y.; Lee, J. C.; Philip Schumm, L.; Sharma, Y.; Anderson, C. A. Host–Microbe Interactions Have Shaped the Genetic Architecture of Inflammatory Bowel Disease. *Nature* **2012**, *491* (7422), 119–124. https://doi.org/10.1038/nature11582.

81. Vermeire, S. NOD2/CARD15: Relevance in Clinical Practice. *Best Pract. Res. Clin. Gastroenterol.* **2004**, *18* (3), 569–575. https://doi.org/10.1016/j.bpg.2003.12.008.

82. Duerr, R. H.; Taylor, K. D.; Brant, S. R.; Rioux, J. D.; Silverberg, M. S.; Daly, M. J.; Steinhart, A. H.; Abraham, C.; Regueiro, M.; Griffiths, A. A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science (80-.).* **2006**, *314* (5804), 1461–1463. https://doi.org/10.1126/science.11352.

83. 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), 1152–1167. https://doi.org/10.1136/gut.2008.163667.

84. Qiu, J.; Guo, X.; Zong-ming, E. C.; He, L.; Sonnenberg, G. F.; Artis, D.; Fu, Y.-X.; Zhou, L. Group 3 Innate Lymphoid Cells Inhibit T-Cell-Mediated Intestinal Inflammation through Aryl Hydrocarbon Receptor Signaling and Regulation of Microflora. *Immunity* **2013**, *39* (2), 386–399. https://doi.org/10.1016/j.immuni.2013.08.002.

85. Qiu, J.; Heller, J. J.; Guo, X.; Zong-ming, E. C.; Fish, K.; Fu, Y.-X.; Zhou, L. The Aryl Hydrocarbon Receptor Regulates Gut Immunity through Modulation of Innate Lymphoid Cells. *Immunity* **2012**, *36* (1), 92–104. https://doi.org/10.1016/j.immuni.2011.11.011.

86. Geremia, A.; Arancibia-Cárcamo, C. V; Fleming, M. P. P.; Rust, N.; Singh, B.; Mortensen, N. J.; Travis, S. P. L.; Powrie, F. IL-23–Responsive Innate Lymphoid Cells Are Increased in Inflammatory Bowel Disease. *J. Exp. Med.* **2011**, *208* (6),

1127–1133. https://doi.org/10.1084/jem.20101712.

87. Khor, B.; Gardet, A.; Xavier, R. J. Genetics and Pathogenesis of Inflammatory Bowel Disease. *Nature* **2011**, *474* (7351), 307–317. https://doi.org/10.1038/nature10209.

88. Chassaing, B.; Koren, O.; Goodrich, J. K.; Poole, A. C.; Srinivasan, S.; Ley, R. E.; Gewirtz, A. T. Dietary Emulsifiers Impact the Mouse Gut Microbiota Promoting Colitis and Metabolic Syndrome. *Nature* **2015**, *519* (7541), 92–96. https://doi.org/10.1038/nature14232.

89. Boltin, D.; Perets, T. T.; Vilkin, A.; Niv, Y. Mucin Function in Inflammatory Bowel Disease: An Update. *J. Clin. Gastroenterol.* **2013**, *47* (2), 106–111. https://doi.org/10.1097/MCG.0b013e3182688e73.

90. Liu, J. Z.; Anderson, C. A. Genetic Studies of Crohn's Disease: Past, Present and Future. *Best Pract. Res. Clin. Gastroenterol.* **2014**, *28* (3), 373–386. https://doi.org/10.1016/j.bpg.2014.04.009.

91. Stoll, M.; Corneliussen, B.; Costello, C. M.; Waetzig, G. H.; Mellgard, B.; Koch, W. A.; Rosenstiel, P.; Albrecht, M.; Croucher, P. J. P.; Seegert, D. Genetic Variation in DLG5 Is Associated with Inflammatory Bowel Disease. *Nat. Genet.* **2004**, *36* (5), 476–480. https://doi.org/10.1038/ng1345.

92. Brand, S.; Konrad, A.; Crispin, A.; Göke, B.; Lohse, P.; Ochsenkühn, T. The Role of Toll-like Receptor 4 Asp299Gly and Thr399Ile Polymorphisms and CARD15/NOD2 Mutations in the Susceptibility and Phenotype of Crohn's Disease. *Inflamm. Bowel Dis.* **2005**, *11* (7), 645–652. https://doi.org/10.1097/01.MIB.0000168372.94907.d2.

93. Ioannidis, J.; Ntzani, E. E.; Trikalinos, T. A.; Contopoulos-Ioannidis, D. G. Replication Validity of Genetic Association Studies. *Nat. Genet.* **2001**, *29* (3), 306–309. https://doi.org/10.1038/ng749.

94. Colhoun, H. M.; McKeigue, P. M.; Smith, G. D. Problems of Reporting Genetic Associations with Complex Outcomes. *Lancet* **2003**, *361* (9360), 865–872. https://doi.org/10.1016/S0140-6736(03)12715-8.

95. Sherry, S. T.; Ward, M.; Sirotkin, K. DbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.* **1999**, *9* (8), 677–679. https://doi.org/10.1101/gr.9.8.677.

96. 1, International S N P Map Working Group Cold Spring Harbor Laboratories: Sachidanandam Ravi 1 Weissman David 1 Schmidt Steven C 1 Kakol Jerzy M 1 Stein Lincoln D 2, National Center for Biotechnology Information: Marth Gabor 2 Sherry Steve M, Sanger Centre: Mullikin James C 3 Mortimore Beverley J 3 Willey David L 3 Hunt Sarah E 3 Cole Charlotte G 3 Coggill Penny C 3 Rice Catherine M 3 Ning Zemin 3 Rogers Jane 3 Bentley David R drb@ sanger. ac. uk 3; 4, Washington University in St. Louis: Kwok Pui-Yan 4 Mardis Elaine R 4 Yeh Raymond T 4 Schultz Brian 4 Cook Lisa 4 Davenport Ruth 4 Dante Michael 4 Fulton Lucinda 4 Hillier LaDeana 4 Waterston Robert H 4 McPherson John D Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms. *Nature* **2001**, *409* (6822), 928–933. https://doi.org/10.1038/35057149.

97. Consortium, I. H. A Second Generation Human Haplotype Map of over 3.1 Million SNPs. *Nature* **2007**, *449* (7164), 851. https://doi.org/10.1038/nature06258.

98. Syvänen, A.-C. Toward Genome-Wide SNP Genotyping. *Nat. Genet.* **2005**, *37* (Suppl 6), S5–S10. https://doi.org/10.1038/ng1558.

99. Barrett, J. C.; Cardon, L. R. Evaluating Coverage of Genome-Wide Association Studies. *Nat. Genet.* **2006**, *38* (6), 659–662. https://doi.org/10.1038/ng1801.

100. Yamazaki, K.; McGovern, D.; Ragoussis, J.; Paolucci, M.; Butler, H.; Jewell, D.; Cardon, L.; Takazoe, M.; Tanaka, T.; Ichimori, T. Single Nucleotide Polymorphisms in TNFSF15 Confer Susceptibility to Crohn's Disease. *Hum. Mol. Genet.* **2005**, *14* (22), 3499–3506. https://doi.org/10.1093/hmg/ddi379.

101. Hampe, J.; Franke, A.; Rosenstiel, P.; Till, A.; Teuber, M.; Huse, K.; Albrecht, M.; Mayr, G.; De La Vega, F. M.; Briggs, J. A Genome-Wide Association Scan of Nonsynonymous SNPs Identifies a Susceptibility Variant for Crohn Disease in ATG16L1. *Nat. Genet.* **2007**, *39* (2), 207–211. https://doi.org/10.1038/ng1954.

102. 20, 1958 Birth Cohort Controls Jones Richard W 18 McArdle Wendy L 18 Ring Susan M 18 Strachan David P 19 Pembrey Marcus 18; 2, Type 1 Diabetes Clayton David G 2 Dunger David B 2 41 Nutland Sarah 2 Stevens Helen E 2 Walker Neil M 2 Widmer Barry 2 41 Todd John A. Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature* **2007**, *447* (7145), 661–678. https://doi.org/10.1038/nature05911.

103. Huang, C.; Haritunians, T.; Okou, D. T.; Cutler, D. J.; Zwick, M. E.; Taylor, K. D.; Datta, L. W.; Maranville, J. C.; Liu, Z.; Ellis, S. Characterization of Genetic Loci That Affect Susceptibility to Inflammatory Bowel Diseases in African Americans. *Gastroenterology* **2015**, *149* (6), 1575–1586.

https://doi.org/10.1053/j.gastro.2015.07.065.

104. Liu, J. Z.; Van Sommeren, S.; Huang, H.; Ng, S. C.; Alberts, R.; Takahashi, A.; Ripke, S.; Lee, J. C.; Jostins, L.; Shah, T. Association Analyses Identify 38 Susceptibility Loci for Inflammatory Bowel Disease and Highlight Shared Genetic Risk across Populations. *Nat. Genet.* **2015**, *47* (9), 979–986. https://doi.org/10.1038/ng.3359.

105. Parkes, M.; Barrett, J. C.; Prescott, N. J.; Tremelling, M.; Anderson, C. A.; Fisher, S. A.; Roberts, R. G.; Nimmo, E. R.; Cummings, F. R.; Soars, D. Sequence Variants in the Autophagy Gene IRGM and Multiple Other Replicating Loci Contribute to Crohn's Disease Susceptibility. *Nat. Genet.* **2007**, *39* (7), 830–832. https://doi.org/10.1038/ng2061.

106. Van Limbergen, J.; Wilson, D. C.; Satsangi, J. The Genetics of Crohn's Disease. *Annu. Rev. Genomics Hum. Genet.* **2009**, *10*, 89–116. https://doi.org/10.1146/annurev-genom-082908-150013.

107. Zhernakova, A.; van Diemen, C. C.; Wijmenga, C. Detecting Shared Pathogenesis from the Shared Genetics of Immune-Related Diseases. *Nat. Rev. Genet.* **2009**, *10* (1), 43–55. https://doi.org/10.1038/nrg2489. https://doi.org/10.1038/nrg2489.

108. Economou, M.; Pappas, G. New Global Map of Crohn's Disease: Genetic, Environmental, and Socioeconomic Correlations. *Inflamm. Bowel Dis.* **2008**, *14* (5), 709–720. https://doi.org/10.1002/ibd.20352.

109. Feinberg, A. P.; Fallin, M. D. Epigenetics at the Crossroads of Genes and the Environment. *Jama* **2015**, *314* (11), 1129–1130. https://doi.org/10.1001/jama.2015.10414.

110. Hill, A. V. S. Evolution, Revolution and Heresy in the Genetics of Infectious Disease Susceptibility. *Philos. Trans. R. Soc. B Biol. Sci.* **2012**, *367* (1590), 840–849. https://doi.org/10.1098/rstb.2011.0275.

111. Azad, A. K.; Sadee, W.; Schlesinger, L. S. Innate Immune Gene Polymorphisms in Tuberculosis. *Infect. Immun.* **2012**, *80* (10), 3343–3359. https://doi.org/10.1128/iai.00443-12**.**

112. Chiodini, R. J.; Chamberlin, W. M.; Sarosiek, J.; McCallum, R. W. Crohn's Disease and the Mycobacterioses: A Quarter Century Later. Causation or Simple Association? *Crit. Rev. Microbiol.* **2012**, *38* (1), 52–93.

https://doi.org/10.3109/1040841X.2011.638273.

113. Naser, S. A.; Sagramsingh, S. R.; Naser, A. S.; Thanigachalam, S. Mycobacterium Avium Subspecies Paratuberculosis Causes Crohn's Disease in Some Inflammatory Bowel Disease Patients. *World J. Gastroenterol. WJG* **2014**, *20* (23), 7403. https://doi.org/10.3748/wjg.v20.i23.7403.

114. Coulombe, F.; Divangahi, M.; Veyrier, F.; de Léséleuc, L.; Gleason, J. L.; Yang, Y.; Kelliher, M. A.; Pandey, A. K.; Sassetti, C. M.; Reed, M. B. Increased NOD2-Mediated Recognition of N-Glycolyl Muramyl Dipeptide. *J. Exp. Med.* **2009**, *206* (8), 1709–1716. https://doi.org/10.1084/jem.20081779.

115. Higuchi, R.; Bowman, B.; Freiberger, M.; Ryder, O. A.; Wilson, A. C. DNA Sequences from the Quagga, an Extinct Member of the Horse Family. *Nature* **1984**, *312* (5991), 282–284. https://doi.org/10.1038/312282a0.

116. Pääbo, S. Molecular Cloning of Ancient Egyptian Mummy DNA. *Nature* **1985**, *314* (6012), 644–645. Pääbo, S. Molecular cloning of Ancient Egyptian mummy DNA. *Nature* **314**, 644–645 (1985). https://doi.org/10.1038/314644a0.

117. Rizzi, E.; Lari, M.; Gigli, E.; De Bellis, G.; Caramelli, D. Ancient DNA Studies: New Perspectives on Old Samples. *Genet. Sel. Evol.* **2012**, *44*, 1–19. https://doi.org/10.1186/1297-9686-44-21.

118. Willerslev, E.; Cooper, A. Ancient Dna. *Proc. R. Soc. B Biol. Sci.* **2005**, *272* (1558), 3–16. https://doi.org/10.1098/rspb.2004.2813.

119. Pääbo, S.; Wilson, A. C. Polymerase Chain Reaction Reveals Cloning Artefacts. *Nature* **1988**, *334* (6181), 387–388. https://doi.org/10.1038/334387b0.

120. Pääbo, S. Ancient DNA: Extraction, Characterization, Molecular Cloning, and Enzymatic Amplification. *Proc. Natl. Acad. Sci.* **1989**, *86* (6), 1939–1943. https://doi.org/10.1073/pnas.86.6.1939.

121. Pääbo, S.; Higuchi, R. G.; Wilson, A. C. Ancient DNA and the Polymerase Chain Reaction: The Emerging Field of Molecular Archaeology (Minireview). *J. Biol. Chem.* **1989**, *264* (17), 9709–9712.

122. Thomas, R. H.; Schaffner, W.; Wilson, A. C.; Pääbo, S. DNA Phylogeny of the Extinct Marsupial Wolf. *Nature* **1989**, *340* (6233), 465–467.

https://doi.org/10.1038/340465a0.

123. Irving-Pease, E. K.; Muktupavela, R.; Dannemann, M.; Racimo, F. Quantitative Human Paleogenetics: What Can Ancient DNA Tell Us about Complex Trait Evolution? *Front. Genet.* **2021**, *12*, 703541. https://doi.org/10.3389/fgene.2021.703541.

124. Harrison, P. W.; Amode, M. R.; Austine-Orimoloye, O.; Azov, A. G.; Barba, M.; Barnes, I.; Becker, A.; Bennett, R.; Berry, A.; Bhai, J.; Bhurji, S. K.; Boddu, S.; Branco Lins, P. R.; Brooks, L.; Ramaraju, S. B.; Campbell, L. I.; Martinez, M. C.; Charkhchi, M.; Chougule, K.; Cockburn, A.; Davidson, C.; De Silva, N. H.; Dodiya, K.; Donaldson, S.; El Houdaigui, B.; Naboulsi, T. E.; Fatima, R.; Giron, C. G.; Genez, T.; Grigoriadis, D.; Ghattaoraya, G. S.; Martinez, J. G.; Gurbich, T. A.; Hardy, M.; Hollis, Z.; Hourlier, T.; Hunt, T.; Kay, M.; Kaykala, V.; Le, T.; Lemos, D.; Lodha, D.; Marques-Coelho, D.; Maslen, G.; Merino, G. A.; Mirabueno, L. P.; Mushtaq, A.; Hossain, S. N.; Ogeh, D. N.; Sakthivel, M. P.; Parker, A.; Perry, M.; Piližota, I.; Poppleton, D.; Prosovetskaia, I.; Raj, S.; Pérez-Silva, J. G.; Salam, A. I. A.; Saraf, S.; Saraiva-Agostinho, N.; Sheppard, D.; Sinha, S.; Sipos, B.; Sitnik, V.; Stark, W.; Steed, E.; Suner, M.-M.; Surapaneni, L.; Sutinen, K.; Tricomi, F. F.; Urbina-Gómez, D.; Veidenberg, A.; Walsh, T. A.; Ware, D.; Wass, E.; Willhoft, N. L.; Allen, J.; Alvarez-Jarreta, J.; Chakiachvili, M.; Flint, B.; Giorgetti, S.; Haggerty, L.; Ilsley, G. R.; Keatley, J.; Loveland, J. E.; Moore, B.; Mudge, J. M.; Naamati, G.; Tate, J.; Trevanion, S. J.; Winterbottom, A.; Frankish, A.; Hunt, S. E.; Cunningham, F.; Dyer, S.; Finn, R. D.; Martin, F. J.; Yates, A. D. Ensembl 2024. *Nucleic Acids Res.* **2024**, *52* (D1), D891–D899. https://doi.org/10.1093/nar/gkad1049.

125. Sayers, E. W.; Bolton, E. E.; Brister, J. R.; Canese, K.; Chan, J.; Comeau, D. C.; Connor, R.; Funk, K.; Kelly, C.; Kim, S.; Madej, T.; Marchler-Bauer, A.; Lanczycki, C.; Lathrop, S.; Lu, Z.; Thibaud-Nissen, F.; Murphy, T.; Phan, L.; Skripchenko, Y.; Tse, T.; Wang, J.; Williams, R.; Trawick, B. W.; Pruitt, K. D.; Sherry, S. T. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2022**, *50* (D1), D20–D26. https://doi.org/10.1093/nar/gkab1112.

126. Mallick, S.; Micco, A.; Mah, M.; Ringbauer, H.; Lazaridis, I.; Olalde, I.; Patterson, N.; Reich, D. The Allen Ancient DNA Resource (AADR) a Curated Compendium of Ancient Human Genomes. *Sci. Data* **2024**, *11* (1), 182. https://doi.org/10.1038/s41597-024-03031-7.

127. Kars, M. E.; Başak, A. N.; Onat, O. E.; Bilguvar, K.; Choi, J.; Itan, Y.; Çağlar, C.; Palvadeau, R.; Casanova, J.-L.; Cooper, D. N. The Genetic Structure of the Turkish Population Reveals High Levels of Variation and Admixture. *Proc. Natl. Acad. Sci.* **2021**, *118* (36), e2026076118. https://doi.org/10.1073/pnas.2026076118.

128. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M. A. R.; Bender, D.; Maller, J.; Sklar, P.; De Bakker, P. I. W.; Daly, M. J. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **2007**, *81* (3), 559–575. https://doi.org/10.1086/519795.

129. Team, R. C. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. *(No Title)* **2013**.

130. Voight, B. F.; Kudaravalli, S.; Wen, X.; Pritchard, J. K. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol.* **2006**, *4* (3), e72. https://doi.org/10.1371/journal.pbio.0040072.

131. Gautier, M.; Vitalis, R. Rehh: An R Package to Detect Footprints of Selection in Genome-Wide SNP Data from Haplotype Structure. *Bioinformatics* **2012**, *28* (8), 1176–1177. https://doi.org/10.1093/bioinformatics/bts115.

132. Mi, H.; Thomas, P. PANTHER Pathway: An Ontology-Based Pathway Database Coupled with Data Analysis Tools. *Protein networks Pathw. Anal.* **2009**, 123–140. https://doi.org/10.1007/978-1-60761-175-2_7.

133. Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; Jensen, L. J.; Mering, C. von. STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47* (D1), D607–D613. https://doi.org/10.1093/nar/gky1131. https://doi.org/10.1093/nar/gky1131.

134. Stelzer, G.; Rosen, N.; Plaschkes, I.; Zimmerman, S.; Twik, M.; Fishilevich, S.; Stein, T. I.; Nudel, R.; Lieder, I.; Mazor, Y. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinforma.* **2016**, *54* (1), 1–30. https://doi.org/10.1002/cpbi.5.

135. Hirano, A.; Yamazaki, K.; Umeno, J.; Ashikawa, K.; Aoki, M.; Matsumoto, T.; Nakamura, S.; Ninomiya, T.; Matsui, T.; Hirai, F. Association Study of 71 European Crohn's Disease Susceptibility Loci in a Japanese Population. *Inflamm. Bowel Dis.* **2013**, *19* (3), 526–533. https://doi.org/10.1097/MIB.0b013e31828075e7.

136. Franke, A.; McGovern, D. P. B.; Barrett, J. C.; Wang, K.; Radford-Smith, G. L.; Ahmad, T.; Lees, C. W.; Balschun, T.; Lee, J.; Roberts, R. Genome-Wide Meta-Analysis Increases to 71 the Number of Confirmed Crohn's Disease Susceptibility Loci. *Nat. Genet.* **2010**, *42* (12), 1118–1125. https://doi.org/10.1038/ng.717.

137. Barrett, J. C.; Hansoul, S.; Nicolae, D. L.; Cho, J. H.; Duerr, R. H.; Rioux, J. D.; Brant, S. R.; Silverberg, M. S.; Taylor, K. D.; Barmada, M. M. Genome-Wide Association Defines More than 30 Distinct Susceptibility Loci for Crohn's Disease. *Nat. Genet.* **2008**, *40* (8), 955–962. https://doi.org/10.1038/ng.175.

138. Fuyuno, Y.; Yamazaki, K.; Takahashi, A.; Esaki, M.; Kawaguchi, T.; Takazoe, M.; Matsumoto, T.; Matsui, T.; Tanaka, H.; Motoya, S. Genetic Characteristics of Inflammatory Bowel Disease in a Japanese Population. *J. Gastroenterol.* **2016**, *51*, 672–681. https://doi.org/10.1007/s00535-015-1135-3.

139. Wang, K.; Baldassano, R.; Zhang, H.; Qu, H.-Q.; Imielinski, M.; Kugathasan, S.; Annese, V.; Dubinsky, M.; Rotter, J. I.; Russell, R. K. Comparative Genetic Analysis of Inflammatory Bowel Disease and Type 1 Diabetes Implicates Multiple Loci with Opposite Effects. *Hum. Mol. Genet.* **2010**, *19* (10), 2059–2067. https://doi.org/10.1093/hmg/ddq078.

140. Franke, A.; Balschun, T.; Karlsen, T. H.; Hedderich, J.; May, S.; Lu, T.; Schuldt, D.; Nikolaus, S.; Rosenstiel, P.; Krawczak, M. Replication of Signals from Recent Studies of Crohn's Disease Identifies Previously Unknown Disease Loci for Ulcerative Colitis. *Nat. Genet.* **2008**, *40* (6), 713–715. https://doi.org/10.1038/ng.148.

141. Rioux, J. D.; Xavier, R. J.; Taylor, K. D.; Silverberg, M. S.; Goyette, P.; Huett, A.; Green, T.; Kuballa, P.; Barmada, M. M.; Datta, L. W. Genome-Wide Association Study Identifies New Susceptibility Loci for Crohn Disease and Implicates Autophagy in Disease Pathogenesis. *Nat. Genet.* **2007**, *39* (5), 596–604. https://doi.org/10.1038/ng2032.

142. Yamazaki, K.; Umeno, J.; Takahashi, A.; Hirano, A.; Johnson, T. A.; Kumasaka, N.; Morizono, T.; Hosono, N.; Kawaguchi, T.; Takazoe, M. A Genome-Wide Association Study Identifies 2 Susceptibility Loci for Crohn's Disease in a Japanese Population. *Gastroenterology* **2013**, *144* (4), 781–788. https://doi.org/10.1053/j.gastro.2012.12.021.

143. Julià, A.; Domènech, E.; Ricart, E.; Tortosa, R.; García-Sánchez, V.; Gisbert, J. P.; Mateu, P. N.; Gutiérrez, A.; Gomollón, F.; Mendoza, J. L. A Genome-Wide Association Study on a Southern European Population Identifies a New Crohn's Disease Susceptibility Locus at RBX1-EP300. *Gut* **2013**, *62* (10), 1440–1445. https://doi.org/10.1136/gutjnl-2012-302865.

144. González-Serna, D.; Ochoa, E.; López-Isac, E.; Julià, A.; Degenhardt, F.; Ortego-Centeno, N.; Radstake, T. R. D. J.; Franke, A.; Marsal, S.; Mayes, M. D. A Cross-Disease Meta-GWAS Identifies Four New Susceptibility Loci Shared between

Systemic Sclerosis and Crohn's Disease. *Sci. Rep.* **2020**, *10* (1), 1862. https://doi.org/10.1038/s41598-020-58741-w.

145. Krause-Kyora, B.; Torres, G.; da Silva, N.; Kolbe, D.; Dose, J.; Schade-Lindig, S.; Wahl, J.; Berszin, C.; Francken, M.; Görner, I. Ancient Implications for Today's Precision Medicine: How the First Near East Farmers Shaped the European Genetic Risk Architecture for IBD. **2022**. https://doi.org/10.21203/rs.3.rs-2075746/v1.

146. Lakner, L.; Csöngei, V.; Sarlós, P.; Járomi, L.; Sáfrány, E.; Varga, M.; Orosz, P.; Magyari, L.; Bene, J.; Miheller, P. IGR2096a_1 T and IGR2198a_1 C Alleles on IBD5 Locus of Chromosome 5q31 Region Confer Risk for Crohn's Disease in Hungarian Patients. *Int. J. Colorectal Dis.* **2009**, *24*, 503–507. https://doi.org/10.1007/s00384-009-0670-x.

147. Baskaran, K.; Pugazhendhi, S.; Ramakrishna, B. S. Association of IRGM Gene Mutations with Inflammatory Bowel Disease in the Indian Population. *PLoS One* **2014**, *9* (9), e106863. https://doi.org/10.1371/journal.pone.0106863.

148. Jung, S.; Ye, B. D.; Lee, H.-S.; Baek, J.; Kim, G.; Park, D.; Park, S. H.; Yang, S.-K.; Han, B.; Liu, J. Identification of Three Novel Susceptibility Loci for Inflammatory Bowel Disease in Koreans in an Extended Genome-Wide Association Study. *J. Crohn's Colitis* **2021**, *15* (11), 1898–1907. https://doi.org/10.1093/ecco-jcc/jjab060.

149. Villani, A.-C.; Lemire, M.; Fortin, G.; Louis, E.; Silverberg, M. S.; Collette, C.; Baba, N.; Libioulle, C.; Belaiche, J.; Bitton, A. Common Variants in the NLRP3 Region Contribute to Crohn's Disease Susceptibility. *Nat. Genet.* **2009**, *41* (1), 71–76. https://doi.org/10.1038/ng.285.

APPENDIX A. GENES AND VARIANTS ASSOCIATED WITH CROHN'S DISEASE

Gene	SNP	Type	Chromosome	Ancestral	Derived	RiskAllel	Effect	Pvalue	OR	Crohn Pop.	References
ACO2	rs727563	Intron	22:41471373	C	T	C	Susceptible	1.88E-10	1.1	European, East Asian, Indian, Iranian	104
ADAM10, HSP90AB4P	rs4774310	Intron, Noncoding Transcript	15:58692965	T	G	T		0.005		Germany	101
AIMP1P2	rs2157453	Intron	1:172894808	A	G	A		0.003		Germany	101
AIMP1P2	rs7517810	Intron	1:172884320	C	T	T	Susceptible	1.51E-15	1.22	European / Japanese	135 136
AIMP1P2	rs9286879	Intron	1:172893094	A	G	G	Susceptible	1.53E-9	1.19	European	80 137 138 139
AIMP1P2_ TNFSF18	rs12035082	Intron	1:172929237	C	T	C	Susceptible	4.00E-4	1.2	German	101 140
AKAP10	rs203462	Missense	17:19909228	C	T	C		0.004		Germany	101
ANKRD34B	rs32857	Stopgained	5:80559553	G	A	A		0.003		Germany	101
ANKRD55	rs10065637	Intron	5:56143024	C	T	C	Susceptible	3.68E-12	1.12	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103
APBB2	rs4861358	Missense	4:41013882	C	T	C		0.0009		Germany	101
AQP9	rs1867380	Missense	15:58184082	G	A	A		0.003		Germany	101
ARFRP1	rs6011040	Intron	20:63706054	A	G	G	Protective	0.0047	0.85	German	140
ATG16L1	rs10210302	Intron	2:233250193	T	C	T	Susceptible	5.26E-14	1.85	European British	102

(cont. on the next page)

Appendix A (cont.)

ATG16L1	rs12994997	Intron	2:233264857	G	A	A	Susceptible	4.14E-56	1.23	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	[80] [103]
ATG16L1	rs2241880	Missense	2:233274722	A	G	G	Protective	5.82E-8	0.7	German	[140] [141]
ATG16L1	rs3828309	Intron	2:233271764	A	G	G	Susceptible	2.36E-32	1.28	European (US-Canadian, Belgian-French, British)	[137]
ATG16L1, SCARNA5	rs3792109	Intron, Noncoding Transcript Exon	2:233275771	A	G	A	Susceptible	6.76E-41	1.34	European	[136]
ATG16L2	rs11235604	Missense	11:72822491	C	T	T	Susceptible	1.00E-8	1.47	Japanese	[138]
ATP23	rs3751325	Missense	12:57941843	A	T	T		0.003		Germany	[101]
BACH2	rs1847472	Intron	6:90263440	C	T	C	Susceptible	5.10E-9	1.07	European	[136]
BANK1	rs13126505	Intron	4:101944147	G	A	A	Susceptible	2.33E-10	1.17	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Netherlands, New Zealand, Norway, Sweden, UK, USA	[80] [103]
BSN	rs9858542	Synonymous	3:49664550	G	A	A	Susceptible	5.06E-7	1.32	German	[102] [140] [139]
BTNL2	rs3763313	Upstream	6:32408694	A	C	C	Susceptible	5.20E-9	1.19	European (US-Canadian, Belgian-French, British)	[137]

(cont. on the next page)

Appendix A (cont.)

C17orf67	rs3853824	Intron	17:56803632	T	C	T	Protective	1.17E-10	0.92	European, East Asian, Indian, Iranian	[104]
C7orf33	rs7807268	Intron	7:148560956	C	G	G	Susceptible	4.42E-06	1.47	European British	[102]
CACNA1E	rs704326	Missense	1:181790521	G	A	A		0.004		Germany	[101]
CAPSL	rs1445898	Missense	5:35910427	C	T	C		0.005		Germany	[101]
CARD9	rs4077515	Missense	9:136372044	C	T	T	Susceptible	1.30E-3	1.18	European	[136]
CAVIN1	rs11871801	Intron	17:42418754	A	C	A	Susceptible	2.51E-8	1.15	European / Japanese	[135] [136]
CCL2	rs991804	Downstreame	17:34260706	C	T	C	Susceptible	1.07E-6	1.1	European (US-Canadian, Belgian-French, British)	[137]
CCL7	rs3091315	Upstream	17:34266646	A	G	A	Susceptible	1.70E-13	1.2	European	[136]
CCNY	rs12242110	Upstream	10:35246767	G	A	G	Susceptible	1.10E-9	1.15	European / Japanese	[135] [136]
CCNY	rs3936503	Intron	10:35260329	G	A	A	Susceptible	5.76E-5	1.2	German	[140] [139]
CD244	rs4656940	Intron	1:160860478	A	G	A	Susceptible	6.17E-7	1.15	European / Japanese	[135] [136]
CD40	rs1569723	Upstream	20:46113425	A	C	C	Susceptible	0.0176	1.11	Japanese (European)	[138]
CDKAL1	rs6908425	Intron	6:20728500	C	T	C	Susceptible	8.96E-10	1.21	European (US-Canadian, Belgian-French, British)	[137] [140] [136]
CENPW	rs9388489	Intron	6:126377573	A	G	G	Protective	0.002	0.89	European	[139]
CEP43	rs2301436	Intron	6:167024500	T	C	T	Susceptible	1.04E-12	1.21	European (US-Canadian, Belgian-French, British)	[137]
CEP43	rs415890	Intron	6:166993145	C	G	C	Susceptible	2.51E-12	1.17	European / Japanese	[135] [136]
CIBAR2	rs8050910	Intron	16:85105567	T	G	G	Protective	0.0085	0.84	North American NIDDK	[141]
CLEC2D	rs3764022	Missense	12:9680928	C	G	G		0.009		Germany	[101]

(cont. on the next page)

Appendix A (cont.)

CLN3	rs151181	Intron	16:28479196	C	T	C	Susceptible	1.50E-11	1.07	European / Japanese	[135] [136]
CPEB4	rs17695092	Intron	5:173910850	G	T	T	Susceptible	4.68E-9	1.1	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	[80]
CPEB4	rs359457	Intergenic	5:173852839	T	C	T	Susceptible	2.50E-12	1.08	European	[136]
CUL2	rs11010067	Downstream	10:35006503	G	C	C	Susceptible	0.000841	1.17	Japanese (European)	[138]
CUL2	rs17582416	Regulatory	10:34998722	G	T	G	Susceptible	1.79E-9	1.16	European	[137] [139]
CXCR5	rs630923	Upstream	11:118883644	C	A	A	Protective	7.78E-14	0.91	European	[76]
DCP1B	rs12423058	Missense	12:1955500	T	C	C		5.8E-14		Germany	[101]
DENND1B	rs1998598	Intron	1:197758512	A	G	G	Susceptible	8.70E-9	1.04	European	[136]
DENND2B	rs3812762	Missense	11:8730093	G	C	C		0.003		Germany	[101]
DHX34	rs12984558	Missense	19:47353079	C	T	T		0.003		Germany	[101]
DNMT3A	rs13428812	Intron	2:25269598	A	G	G	Susceptible	8.50E-10	1.06	European	[136]
DOCK7	rs1748195	Intron	1:62583922	G	C	G	Susceptible	7.13E-8	1.07	European, East Asian, Indian, Iranian	[104]
ELF1	rs61300271	Intron	13:41028326	C	T	C	Susceptible	2.08E-6	1.25	Japanese	[138]
ELF1	rs7329174	Intron	13:40983974	A	G	G	Susceptible	7.96E-9	1.27	Japanese	[142]
EMSY	rs7927894	Upstream	11:76590272	C	T	T	Susceptible	1.32E-9	1.16	European (US-Canadian, Belgian-French, British)	[137] [139]
EMSY	rs7927997	Upstream	11:76590331	C	T	T	Susceptible	5.62E-13	1.17	European	[136]

(cont. on the next page)

Appendix A (cont.)

ENSAP3	rs10427252	Noncoding Transcript Exon	2:214697853	G	A	G		0.007		Germany	[101]
EPO	rs1734907	Upstream	7:100717894	G	A	A	Susceptible	0.0183	1.28	Japanese (European)	[138]
ERAP1,ERAP2	rs2549794	Intron	5:96908845	T	C	C	Susceptible	1.10E-10	1.05	European	[136]
FAF1	rs11205760	Intron	1:50708658	C	T	C	Protective	0.043	0.92	German	[140]
FAM187B	rs541169	Stopgained	19:35228117	C	T	T		0.003		Germany	[101]
FGL1_ MTUS1-DT	rs2157650	Intron	8:17847863	G	C	C		0.007		Germany	[101]
FIBP-CCDC85B	rs2231884	Upstream	11:65889093	C	T	T	Susceptible	0.0205	1.11	Japanese (European)	[138]
FOXP2	rs1869839	Intergenic	7:114717488	A	G	G	Protective	7.3E-6	0.83	Southern European (Spain)	[143]
FUCA1	rs13551	Missense	1:23854472	T	C	C		0.002		Germany	[101]
FUT2	rs516246	Intron	19:48702915	C	T	T	Susceptible	1.00E-15	1.11	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	[80] [103]
GAL3ST2	rs35320439	Intron	2:241797926	C	T	C	Susceptible	9.89E-10	1.09	European, East Asian, Indian, Iranian	[104]
GCKR	rs780093	Intron	2:27519736	C	T	T	Susceptible	4.70E-11	1.15	European / Japanese	[135] [136]
GCKR	rs780094	Intron	2:27518370	C	T	T	Susceptible	3.14E-7	1.08	European (US-Canadian, Belgian-French, British)	[137]

(cont. on the next page)

Appendix A (cont.)

GP2_UMOD	rs12444268	Upstream	16:20331250	T	A	A	Protective	0.011	0.9	European	139
GPR35	rs3749171	Missense	2:240630275	C	T	T	Susceptible	0.00756	1.23	Japanese (European)	138
GPR65	rs8005161	Intron	14:88006251	C	T	T	Susceptible	4.20E-18	1.23	European / Japanese	135 136 138
GRP	rs9319943	Intergenic	18:59212595	C	T	C	Susceptible	9.05E-7	1.08	European, East Asian, Indian, Iranian	104
GSTA9P	rs10948733	Noncoding Transcript Exon	6:52955584	G	A	A		0.0009		Germany	101
HEATR5A	rs7157977	Missense	14:31389003	C	T	T		0.004		Germany	101
HERC2	rs916977	Intron	15:28268218	T	C	T	Susceptible	0.0024	1.26	German	140
HLA-DQA1	rs2187668	Intron	6:32638107	C	T	T	Protective	1.14E-6	0.73	European	139
HLA-DQB1, HLA-DQB1-AS1	rs1063355	3PrimeUTR	6:32659937	T	G	T	Susceptible	3.95E-6	1.2	European	139
HLA-DQB1_ MTCO3P1	rs9469220	TF Binding Site	6:32690533	A	G	A	Susceptible	2.28E-6	1.52	European British	102
HLA-DRA	rs3129871	Upstream	6:32438565	C	A	A	Susceptible	5.76E-7	1.22	European	139
HORMAD2	rs5753037	Intron	22:30185733	C	T	T	Protective	0.0014	0.88	European	139
HORMAD2	rs713875	Intron	22:30196498	C	G	C	Susceptible	7.30E-12	1.08	European	135 136
HS6ST3	rs2282135	3PrimeUTR	3:96837651	C	T	T		0.001		Germany	101
ICOSLG	rs2838519	Intron	21:44195140	G	A	G	Susceptible	2.09E-14	1.18	European / Japanese	135 136
ICOSLG	rs7282490	Intron	21:44195858	G	A	G	Susceptible	5.85E-7	1.25	Japanese	138
ICOSLG	rs762421	Intron	21:44195678	A	G	G	Susceptible	1.41E-9	1.13	European (US-Canadian, Belgian-French, British)	137 139
IFI44L	rs3820093	Missense	1:78629841	G	A	A		0.005		Germany	101

(cont. on the next page)

Appendix A (cont.)

IFNGR2	rs2284553	Intron	21:33404389	T	G	G	Susceptible	2.13E-16	1.12	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103
IFNGR2	rs2834215	Intron	21:33424579	G	A	A	Protective	2.89E-7	0.82	Southern European (Spain)	143
IGHMBP2	rs17612126	Missense	11:68938206	C	A	A		0.009		Germany	101
IKZF1	rs1456896	Upstream	7:50264865	T	C	T	Susceptible	1.20E-8	1.14	European	136
IKZF3	rs12946510	Downstream	17:39756124	C	T	T	Susceptible	9.31E-5	1.31	African Americans - Caucasian	103
IL10	rs3024505	Upstream	1:206766559	G	A	A	Susceptible	1.60E-14	1.12	European	136 139
IL12B	rs10045431	Intron	5:159387525	C	A	C	Susceptible	3.86E-13	1.11	European (US-Canadian, Belgian-French, British)	137 139
IL12B	rs6887695	Intron	5:159395637	G	C	C	Susceptible	7.99E-6	1.36	German	140
IL12RB2	rs6659932	Intron	1:67336688	C	A	A	Protective	0.000133	0.79	European(UK)	144
IL18RAP	rs2058660	Intron	2:102437989	A	G	G	Susceptible	1.58E-12	1.19	European	136
IL18RAP	rs917997	Downstream	2:102454108	C	T	T	Susceptible	2.22E-5	1.05	European (US-Canadian, Belgian-French, British)	137 139
IL23R	rs11209026	Missense	1:67240275	G	A	G	Susceptible	1.00E-64	2.66	European	145 140 136 139

(cont. on the next page)

Appendix A (cont.)

IL23R	rs11465804	Intron	1:67236843	T	G	T	Susceptible	6.66E-63	2.5	European (US-Canadian, Belgian-French, British)	[137]
IL23R	rs11465810	Intron	1:67237502	A	C	C	Susceptible	6.19E-14	1.14	European	[76]
IL23R	rs11805303	Intron	1:67209833	C	T	T	Susceptible	5.85E-12	1.86	European British	[102]
IL23R	rs41313262	Missense	1:67240217	G	A	A	Protective	1.01E-14	0.69	European	[76]
IL23R	rs76418789	Missense	1:67182913	G	A	A	Protective	8.18E-14	0.55	European	[76]
IL27	rs26528	Intron	16:28506388	C	T	C	Susceptible	0.0338	1.1	Japanese (European)	[138]
IL27	rs4788084	Upstream	16:28528527	C	T	T	Susceptible	1.4E-7	1.23	European	[139]
IL2RA	rs12722489	Intron	10:6060049	C	T	C	Susceptible	2.90E-9	1.11	European	[136]
IL2RA	rs61839660	Intron	10:6052734	C	T	T	Susceptible	2.31E-14	1.28	European	[76]
IL6R	rs2228145	Missense	1:154454494	A	C	C	Protective	0.00003	0.88		[145]
IL7R	rs6897932	Missense	5:35874473	C	T	T		0.0004		Germany	[101]
INAVA	rs7554511	Intron	1:200908434	C	A	C	Susceptible	1.58E-7	1.14	European	[136]
IPMK	rs1819658	Regulatory	10:58153390	T	C	C	Susceptible	9.10E-17	1.19	European / Japanese	[135] [136]
IPMK	rs2790216	Intron	10:58238165	G	A	G	Susceptible	0.0479	1.09	Japanese (European)	[138]
IRF1	rs2548998	Intron	5:132496822	G	A	G	Protective	3.09E-7	0.79	European(UK)	[144]
IRF1-AS1	rs10077785	Intron	5:132465466	C	T	C	Protective	0.0004	0.78	German	[140]
IRF1-AS1	rs12521868	Intron	5:132448701	G	T	T	Susceptible	1.41E-20	1.23	European	[136] [146]
IRF1-AS1	rs2188962	Intron	5:132435113	C	T	T	Susceptible	2.32E-18	1.25	European (US-Canadian, Belgian-French, British)	[76] [137] [139]
IRF1-AS1	rs6596075	Upstream	5:132406536	G	C	C	Susceptible	3.19E-06	2.06	European British	[102]
IRF4	rs7773324	Intergenic	6:382559	G	A	G	Protective	1.06E-9	0.92	European, East Asian, Indian, Iranian	[104]

(cont. on the next page)

Appendix A (cont.)

IRGM	rs10065172	Missense	5:150848436	C	T	C	Susceptible	7.94E-7	1.52	German	[140]
IRGM	rs11747270	Intron	5:150879305	G	A	G	Susceptible	3.40E-16	1.33	European (US-Canadian, Belgian-French, British)	[137 140 139]
IRGM	rs13361189	Upstream	5:150843825	C	T	C	Susceptible	1.27E-5	1.44	German	[140 147]
IRGM	rs4958847	Intron	5:150860025	A	G	A	Susceptible	3.95E-7	1.46	German	[140]
IRGM	rs7714584	Intron	5:150890858	A	G	G	Susceptible	7.76E-19	1.37	European	[136]
IRGM	rs1000113	Intron	5:150860514	C	T	T	Susceptible	3.15E-7	1.92	European British	[102 140 147]
IRGM	rs10041072	Intron	5:150880080	C	T	C	Susceptible	6.55E-7	1.52	German	[140]
IRGM	rs11949556	Intron	5:150850239	C	A	A	Susceptible	6.83E-7	1.44	German	[140]
IRGM	rs11957134	Intron	5:150851388	A	G	A	Susceptible	9.25E-6	1.47	German	[140]
IRGM	rs1428555	Intron	5:150877829	C	T	T	Susceptible	4.09E-7	1.53	German	[140]
IRGM	rs17111376	Intron	5:150847337	C	T	C	Susceptible	1.38E-6	1.51	German	[140]
IRGM	rs180802994	Missense	5:150848174	G	C	C	Susceptible	0.0046	1.8	Indian	[147]
IRGM	rs6579806	Intron	5:150891522	C	T	C	Susceptible	1.31E-6	1.5	German	[140]
IRGM	rs9637876	5primeUTR	5:150847863	C	T	T	Susceptible	0.0446	1.25	Indian	[147]
IRGM,ZNF300	rs4958427	Intron	5:150899025	T	C	C	Susceptible	5.24E-7	1.45	German	[140]
ITLN1	rs2274910	Noncoding Transcript Exon	1:160882256	T	C	C	Susceptible	1.46E-9	1.14	European (US-Canadian, Belgian-French, British)	[137]
ITLN1	rs4656958	Upstream	1:160887174	G	A	G	Susceptible	0.00338	1.15	Japanese (European)	[138]
JAK2	rs10758669	Upstream	9:4981602	A	C	C	Susceptible	3.46E-9	1.12	European (US-Canadian, Belgian-French, British)	[137]

(cont. on the next page)

Appendix A (cont.)

JAZF1	rs864745	Intron	7:28140937	T	C	T	Susceptible	3.64E-9	1.09	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103 138
KIF15	rs3804583	Missense	3:44843155	C	A	C		0.004		Germany	101
KSR1	rs2945412	Intron	17:27516617	G	A	A	Susceptible	8.68E-03	1.14	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103138
LACC1	rs3764147	Missense	13:43883789	A	G	G	Susceptible	2.08E-13	1.25	European (US-Canadian, Belgian-French, British) / Japanese	137 135 145 138 139
LINC00581	rs12663356	Intergenic	6:21430497	T	C	C	Susceptible	4.00E-12	1.1	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103
LINC00824	rs6651252	Intron	8:128554935	T	C	T	Susceptible	3.90E-18	1.23	European	80 103 136

(cont. on the next page)

Appendix A (cont.)

LINC01014	rs2291479	Noncoding Transcript Exon	3:178457156	C	A	A		0.001		Germany	[101]
LINC01475	rs10883365	Noncoding Transcript Exon	10:99528007	A	G	G	Susceptible	4.03E-5	1.21	German	[102] [140]
LINC01475	rs1548964	Intron	10:99529896	G	C	C	Susceptible	8.09E-6	1.23	German	[140]
LINC02181	rs3829486	Intron	16:87066700	C	T	T		0.006		Germany	[101]
LINC02341	rs2062305	Intron	13:42478744	G	A	G	Susceptible	4.90E-10	1.1	European	[136]
LINC02341	rs9525625	Intron	13:42443894	T	C	T	Susceptible	1.41E-9	1.08	European, East Asian, Indian, Iranian	[104]
LINC02513	rs73243351	Intergenic	4:38333446	G	A	A	Susceptible	6.31E-11	1.38	Japanese	[138]
LINC02513_ LOC105374409	rs1487630	Intergenic	4:38334202	C	T	T	Susceptible	1.29E-11	1.33	Japanese	[142]
LINC02539_ WAKMAR2	rs7753394	Intron	6:137764111	C	T	C	Susceptible	2.59E-05	1.48	European British	[102]
LOC101927668 (RPL21P75)	rs1558043	Intron	7:19940150	G	C	C	Protective	0.0769	0.85	German	[140]
LOC101927745	rs1736020	Intron	21:15440233	C	A	C	Susceptible	9.33E-12	1.16	European / Japanese	[135] [136]
LOC101927745	rs2823286	Upstream	21:15445619	A	G	G	Susceptible	0.000858	1.23	Japanese (European)	[138]
LOC101927745 (CYCSP42)	rs1736135	Intron	21:15432901	T	C	T	Susceptible	7.40E-9	1.18	European (US-Canadian, Belgian-French, British)	[137] [139]
LOC101927745 (CYCSP42)	rs1736137	Intron	21:15434376	A	G	G	Protective	7.92E-14	0.92	European	[76]
LOC101928354	rs17119	Intron	6:14719265	G	A	A	Susceptible	0.0198	1.24	Japanese (European)	[138]

(cont. on the next page)

Appendix A (cont.)

LOC101929305 (KIF21B_ CACNA1S)	rs17419032	Upstream	1:201029403	C	T	C	Protective	0.0039	0.81	German	140
LOC102723878 (PRDX5)	rs694739	Intron	11:64329761	A	G	A	Susceptible	6.00E-10	1.1	European	136
LOC105371313 (CDH11)	rs1864147	Regulatory	16:64910710	C	G	G		0.003		Germany	101
LOC105373724	rs2111485	Regulatory	2:162254026	A	G	A	Susceptible	0.0354	1.13	Japanese (European)	138
LOC105374736	rs7711427	Intron	5:40414784	A	C	C	Susceptible	3.05E-14	1.2	European	76
LOC105374761_ LOC105374764	rs10865331	Intron	2:62324337	G	A	A	Susceptible	9.77E-10	1.1	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103
LOC105374780	rs6740462	Intron	2:65440138	A	C	A	Susceptible	0.00147	1.22	Japanese (European)	138
LOC105375005, OR14J1	rs9257694	Missense	6:29306709	T	C	C		0.002		Germany	101
LOC105376364_ LINC02639	rs6601764	Downstream	10:3820350	T	C	C	Susceptible	8.95E-6	1.52	European British	102
LOC105376856 (ZBTB40)	rs12568930	Intergenic	1:22375738	T	C	T	Susceptible	0.0469	1.11	Japanese (European)	138
LOC105378204 (NDFIP1)	rs11167764	Regulatory	5:142099500	C	A	C	Susceptible	2.00E-9	1.06	European	136
LOC105378327_ ALDH7A1P4	rs10761659	Intron	10:62685804	A	G	G	Protective	8.25E-5	0.75	German / Japanese	102 135 140 136 138

(cont. on the next page)

Appendix A (cont.)

LOC105378327_ ALDH7A1P4	rs224136	Intron	10:62710915	T	C	C	Protective	0.013	0.83	German	140 141
LOC105379031 (LINC02230)	rs7702331	Intron	5:73255307	G	A	A	Susceptible	5.90E-12	1.12	European / Japanese	103 135 136 138
LOC107984647 (RNU6-921P)	rs4902642	Regulatory	14:68743482	G	A	G	Susceptible	1.60E-10	1.07	European	136
LOC107984997_ RNFT1P2	rs17391694	Intergenic	1:78157942	C	T	C	Susceptible	2.95E-9	1.13	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103
LOC107986537	rs68191	Regulatory	6:33512961	T	C	C	Susceptible	8.7E-6	1.39	European(UK)	144
LOC112267902 (LINC02571)	rs9264942	Intron	6:31306603	T	C	C	Susceptible	4.96E-14	1.15	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103
LOC124900967_ LOC105374737 (PTGER4)	rs10043340	Intron	5:40494865	A	C	C	Protective	3.25E-6	0.75	African Americans - Caucasian	103
LOC124900967_ LOC105374737 (PTGER4)	rs1505994	Intron	5:40526922	C	T	T	Protective	3.65E-5	0.78	African Americans - Caucasian	103

(cont. on the next page)

Appendix A (cont.)

LOC124900967_ LOC105374737 (PTGER4)	rs17234657	Intergenic	5:40401407	T	G	G	Susceptible	1.33E-9	1.49	German	[137 140]
LOC124900967_ LOC105374737 (PTGER4)	rs1876141	Intron	5:40525633	G	A	G	Protective	0.00004	0.78	African Americans - Caucasian	[103]
LOC124900967_ LOC105374737 (PTGER4)	rs4286721	Intron	5:40497502	A	G	G	Protective	2.97E-6	0.74	African Americans - Caucasian	[103]
LOC124900967_ LOC105374737 (PTGER4)	rs6866402	Intron	5:40517229	T	C	T	Protective	4.05E-5	0.78	African Americans - Caucasian	[103]
LOC124900967_ LOC105374737 (PTGER4)	rs9292777	Regulatory	5:40437846	C	T	T	Protective	8.67E-6	0.8	German	[140]
LOC124904259	rs8098673	Intron	18:22093370	A	C	C	Susceptible	2.88E-5	1.05	European (US-Canadian, Belgian-French, British)	[137]
LOC285626 (IL12B)	rs1363670	Intron	5:159357103	G	C	C	Protective	0.024	0.84	German	[140]
LOC285626 (IL12B)	rs6556412	Intron	5:159360377	A	G	A	Susceptible	5.37E-14	1.18	European / Japanese	[135 136]
LOC285626_ LINC01845	rs6871626	Intron	5:159399784	C	A	A	Susceptible	1.43E-9	1.3	Japanese (European)	[138]
LRRK2	rs3761863	Missense	12:40364850	C	T	T		0.007		Germany	[101]
LRRK2	rs7307562	Intron	12:40331158	G	T	T	Protective	1.36E-10	0.9	European	[76]
LTA-TNF	rs1799964	Upstream	6:31574531	C	T	C	Susceptible	3.98E-11	1.19	European / Japanese	[135 136]
LTBR	rs7954567	Intron	12:6381959	G	A	A	Susceptible	1.30E-9	1.09	European, East Asian, Indian, Iranian	[104]

(cont. on the next page)

Appendix A (cont.)

LYRM4	rs12529198	Intron	6:5151013	G	A	G	Susceptible	6.96E-7	1.12	European (US-Canadian, Belgian-French, British)	137 140
MACROD2	rs6105269	Intron	20:14419519	G	A	A	Susceptible	1.99E-5	1.19	Southern European (Spain)	143
MAMSTR	rs281379	Downstream	19:48711017	A	G	A	Susceptible	7.40E-12	1.07	European	136
METTL4	rs3810071	Regulatory	18:2518698	C	T	T		0.003		Germany	101
MFSD12	rs2240751	Missense	19:3548233	A	G	G	Susceptible	3.03E-8	1.25	Korean	148
MROH3P	rs11584383	Downstream	1:200966738	T	C	T	Susceptible	1.43E-11	1.18	European (US-Canadian, Belgian-French, British)	137
MST1	rs3197999	Missense	3:49684099	G	A	A	Susceptible	1.15E-12	1.2	European (US-Canadian, Belgian-French, British)	137 136 139
MTCO3P1	rs9275383	Regulatory	6:32701069	T	G	T	Protective	3.85E-6	0.73	European	139
MUC19	rs11175593	Noncoding Transcript Exon	12:40208138	C	T	T	Susceptible	3.08E-1	1.54	European (US-Canadian, Belgian-French, British)	137 139
MYO10	rs27431	Intron	5:16670509	C	T	C		0.005		Germany	101
NCF4, NCF4-AS1	rs4821544	Intron	22:36862461	C	T	C	Susceptible	0.031	1.07	German	140 141
NCR3	rs2844480	Upstream	6:31597044	C	T	T	Susceptible	3.36E-7	1.27	European	139
NFATC1	rs7236492	Intron	18:79460616	C	T	T	Protective	9.09E-9	0.91	European, East Asian, Indian, Iranian	104
NHSL1-AS1	rs1129180	Noncoding Transcript	6:138696995	A	G	G		0.004		Germany	101
NKX2-3	rs10786559	Intergenic	10:99549133	G	C	C	Protective	0.0006	0.77	German	140
NKX2-3	rs10883353	Intergenic	10:99496351	G	A	A	Susceptible	0.0046	1.3	German	140

(cont. on the next page)

Appendix A (cont.)

NKX2-3	rs11190140	Upstream	10:99531836	C	T	T	Susceptible	3.06E-16	1.2	European (US-Canadian, Belgian-French, British)	137 140 139
NKX2-3	rs12414093	Intergenic	10:99574477	A	G	A	Protective	0.004	0.8	German	140
NKX2-3	rs1360522	Intergenic	10:99566040	T	C	T	Protective	0.0001	0.76	German	140
NKX2-3	rs4409764	Upstream	10:99524480	T	G	T	Susceptible	2.29E-20	1.22	European / Japanese	135 136 138
NKX2-3	rs4601693	Intergenic	10:99572282	T	A	A	Susceptible	0.016	1.28	German	140
NKX2-3	rs4919345	Intergenic	10:99544560	G	C	C	Protective	2.75E-5	0.74	German	140
NKX2-3	rs7091572	Intergenic	10:99568094	T	C	C	Protective	2.44E-5	0.73	German	140
NLRP13	rs303997	Missense	19:55913077	T	C	T		0.001		Germany	101
NLRP8	rs306481	Missense	19:55976237	A	G	A		0.008		Germany	101
NOD2	rs104895444	Missense	16:50712288	G	A	A	Susceptible	8.18E-14	1.97	European	76
NOD2	rs104895467	Missense	16:50716899	A	G	G	Susceptible	5.9E-14	1.99	European	76
NOD2	rs17221417	Intron	16:50705671	C	G	G	Susceptible	3.98E-11	1.92	European British	102
NOD2	rs2066843	Synonymous	16:50711288	C	T	T	Susceptible	3.22E-20	1.46	European	139
NOD2	rs2066844	Missense	16:50712015	C	T	T	Susceptible	4.02E-21	2.35	German	76 140
NOD2	rs2066845	Missense	16:50722629	G	C	C	Susceptible	9.16E-14	2.23	European	76 101 140
NOD2	rs2066847	Frameshift	16:50729868-50729870	CCC	CCCC	C	Protective	3.59E-14	0.31	European	76 137 140
NOD2	rs2076756	Intron	16:50722970	A	G	G	Susceptible	3.98E-69	1.53	European	136
NOD2	rs5743271	Missense	16:50710777	A	G	G	Susceptible	8.02E14	1.52	European	76
NOD2	rs5743289	Intron	16:50722863	T	C	T	Susceptible	6.12E-5	1.76	African Americans - Caucasian	103
NOD2	rs72796367	Intron	16:50728860	T	C	C	Susceptible	2.10E-14	1.47	European	76

(cont. on the next page)

Appendix A (cont.)

NOTCH2 (ADAM30)	rs3897478	Downstream	1:119908567	T	C	T	Susceptible	1.97E-11	1.16	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	[80] [103]
NUDCD1	rs2980618	Missense	8:109289769	T	C	C		0.006		Germany	[101]
OAS2	rs15895	Stoplost	12:113010483	A	G	A		0.004		Germany	[101]
OR10A4	rs2595453	Missense	11:6877264	T	C	C		0.00005		Germany	[101]
OR2B11	rs10733113	Upstream	1:247459055	A	G	G	Susceptible	3.49E-9	1.53	European	[149]
OR2B11	rs4266924	Intron	1:247453832	G	A	A	Susceptible	6.01E-7	1.69	European	[149]
OR2B11	rs4353135	Intron	1:247453734	G	T	T	Susceptible	8.36E-3	1.21	European	[149]
OR2B11	rs55646866	Intron	1:247457083	C	T	C	Susceptible	7.20E-7	1.69	European	[149]
OR2B11	rs6672995	Noncoding Transcript	1:247457731	G	A	G	Susceptible	2.91E-6	1.53	European	[149]
OR2H4P	rs3129096	Noncoding Transcript Exon	6:29215569	G	A	A		0.002		Germany	[101]
OR2J4P	rs3116817	Noncoding Transcript Exon	6:29181784	A	G	A		0.002		Germany	[101]
OR8I1P	rs17613241	Noncoding Transcript Exon	11:56296687	G	A	G		2.2E-09		Germany	[101]

(cont. on the next page)

Appendix A (cont.)

OSGIN2	rs7015630	ıntergenic	8:89863690	T	C	T	Susceptible	1.4E-11	1.08	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 138
PDE4A	rs145530718	Intron	19:10458207	G	C	C	Susceptible	7.30E-14	1.24	European	76
PDGFB	rs2413583	Intron	22:39263768	T	C	C	Susceptible	1.10E-26	1.23	European	136
PER3	rs2797685	Intron	1:7819003	C	T	T	Susceptible	7.10E-9	1.05	European / Japanese	135 136
PHACTR2	rs12199775	Intron	6:143577757	A	G	A	Susceptible	0.045	1.23	Japanese (European)	138
PHOX2B-AS1	rs16853571	Intron	4:41751113	A	C	C	Protective	0.0084	0.69	North American NIDDK	141
PHTF1	rs6679677	Upstream	1:113761186	C	A	C	Susceptible	2.02E-15	1.2	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103
PKD1L2	rs1869348	Synonymous	16:81108652	C	T	T		0.002		Germany	101
PLA2G4A	rs10798069	Intron	1:186906327	G	T	T	Protective	4.25E-9	0.93	European, East Asian, Indian, Iranian	104
PLAU,C10orf55	rs2227564	Missense, Intron	10:73913343	C	T	C	Susceptible	0.00621	1.15	Japanese (European)	138 145
PLCL1	rs6738825	Intron	2:198032171	A	G	A	Susceptible	3.50E-9	1.06	European / Japanese	135 136

(cont. on the next page)

Appendix A (cont.)

PLSCR2	rs13092702	Noncoding Transcript	3:146401508	C	A	C		0.005		Germany	[101]
PLSCR4	rs3762685	Missense	3:146220832	T	C	C		0.002		Germany	[101]
PLSCR4	rs1061409	Missense	3:146199974	T	C	C		0.003		Germany	[101]
PNKD,TMBIM1	rs2382817	Intron	2:218286495	A	C	A	Susceptible	0.0163	1.11	Japanese (European)	[138]
PPBP_CXCL5	rs2472649	Upstream	4:73991991	A	G	G	Susceptible	0.0142	1.19	Japanese (European)	[138]
PPM1G	rs1728918	Upstream	2:27412596	G	A	A	Susceptible	4.85E-16	1.12	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	[80 103 138]
PPP5C	rs4802307	Upstream	19:46346549	G	T	G	Susceptible	1.99E-10	1.1	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	[80 103 138]
PRDM1	rs28701841	Intron	6:106082455	G	A	A	Susceptible	7.99E-14	1.12	European	[76]
PRDM1	rs6568421	Regulatory	6:105987150	A	G	G	Susceptible	4.37E-8	1.13	European	[136]
PRDM1_ LOC105377923	rs7746082	Regulatory	6:105987394	G	C	C	Susceptible	2.44E-10	1.17	European (US-Canadian, Belgian-French, British)	[137]
PTGER4	rs11742570	Upstream	5:40410482	C	T	C	Susceptible	7.08E-36	1.33	European	[136]

(cont. on the next page)

Appendix A (cont.)

PTGER4	rs4613763	Regulatory	5:40392626	T	C	C	Susceptible	6.82E-27	1.32	European (US-Canadian, Belgian-French, British)	137 139
PTPN2	rs10460003	Intron	18:12747013	C	T	C	Susceptible	0.012	1.23	German	140
PTPN2	rs1893217	Intron	18:12809341	A	G	G	Susceptible	1.29E-14	1.25	European	136 139
PTPN2	rs2542151	Upstream	18:12779948	T	G	G	Susceptible	5.10E-17	1.35	European (US-Canadian, Belgian-French, British)	102 110 140
PTPN2	rs2542170	Intron	18:12800821	T	C	T	Protective	0.01	0.89	German	140
PTPN2	rs487273	Intron	18:12853459	G	T	G	Susceptible	0.038	1.13	German	140
PTPN22	rs2476601	Missense	1:113834946	G	A	G	Susceptible	1.46E-8	1.31	European (US-Canadian, Belgian-French, British)	137 145 136 139
PTPRC	rs7555082	Intergenic	1:198629533	G	A	A	Susceptible	1.47E-10	1.13	European, East Asian, Indian, Iranian	104
PUS10	rs13003464	Intron	2:60959694	G	A	G	Susceptible	4.60E-6	1.16	European (US-Canadian, Belgian-French, British)	137
PUS10	rs7608910	Intron	2:60977721	A	G	G	Susceptible	1.07E-5	1.39	Japanese	138
PUS10	rs10181042	Intron	2:60997124	C	T	T	Susceptible	6.61E-9	1.14	European / Japanese	136 135
RASGRP1	rs16967103	Intergenic	15:38606989	C	T	C	Susceptible	3.88E-9	1.09	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103 138

(cont. on the next page)

Appendix A (cont.)

RBFOX1	rs2191423	Intron	16:6337347	C	A	A		0.006		Germany	[101]
RBX1_RPS9P2	rs4820425	Intergenic	22:41035338	C	A	A	Susceptible	3.42E-8	1.27	Southern European (Spain)	[143]
RFX6	rs6936629	Intron	6:116917978	T	C	C	Susceptible	3.63E-8	1.25	Korean	[148]
RIT1	rs670523	Intron	1:155908941	A	G	A	Susceptible	0.042	1.13	Japanese (European)	[138]
RNASET2	rs1819333	Upstream	6:166960059	T	G	T	Susceptible	0.000113	1.18	Japanese	[138]
RPL31P17	rs6730351	Noncoding Transcript Exon	2:222726702	C	G	G		0.0007		Germany	[101]
RSPH6A	rs8111071	Intron	19:45804148	A	G	G	Susceptible	1.75E-05	1.28	European British	[102]
RSPO3	rs9491697	Intron	6:127134977	A	G	G	Susceptible	3.78E-10	1.08	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	[80] [103]
S100Z	rs1320308	Missense	5:76875427	A	C	A		0.0003		Germany	[101]
SATB1-AS1	rs13073817	Intron	3:18665366	G	A	A	Susceptible	6.70E-9	1.08	European	[136]
SATB1-AS1_ LOC107986066	rs4256159	Intron	3:18725912	C	T	T	Susceptible	0.00117	1.19	Japanese (European)	[138]

(cont. on the next page)

Appendix A (cont.)

SBNO2	rs2024092	Intron	19:1124032	G	A	A	Susceptible	8.26E-8	1.16	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103
SBNO2	rs4807569	Intron	19:1123379	A	C	C	Susceptible	2.12E-9	1.02	European (US-Canadian, Belgian-French, British)	137
SCAMP3	rs1142287	Synonymous	1:155260340	C	T	T	Susceptible	2.30E-13	1.13	European	136
SKAP2	rs10486483	Intron	7:26852821	A	G	A	Susceptible	2.55E-8	1.09	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80
SLAIN2	rs6837335	Intron	4:48361966	G	A	G	Susceptible	1.75E-8	1.09	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80
SLC16A4	rs2271885	Missense	1:110379093	T	G	G		0.003		Germany	101

(cont. on the next page)

Appendix A (cont.)

SLC17A3	rs1165165	Missense	6:25862238	C	T	T		0.0009		Germany	[101]
SLC1A4	rs759458	Missense	2:65018231	G	A	A		0.004		Germany	[101]
SLC22A23	rs17309827	Intron	6:3433084	T	G	T	Susceptible	2.74E-6	1.1	European / Japanese	[137] [135] [140] [136]
SLC22A23	rs13204048	Intron	6:3420172	T	C	C	Protective	2.89E-8	0.93	European, East Asian, Indian, Iranian	[104]
SLC22A4, MIR3936HG	rs1050152	Missense, Intron	5:132340627	C	T	T		0.003		Germany	[101]
SLC22A5	rs11739135	Regulatory	5:132397705	G	C	C	Susceptible	0.01	1.65	Hungarian	[146]
SLC44A5	rs211716	Intron	1:75641654	G	A	A		0.0001		Germany	[101]
SLC44A5	rs211715	Intron	1:75641788	G	C	C		0.0002		Germany	[101]
SLC7A10_ CEBPA	rs736289	Intergenic	19:33266156	C	T	T	Susceptible	8.70E-9	1.06	European	[136]
SMAD3	rs17293632	Intron	15:67150258	C	T	T	Susceptible	2.70E-19	1.12	European	[136]
SMIM3	rs12653083	Upstream	5:150778385	C	T	C	Susceptible	0.001	1.52	German	[140]
SP140	rs6716753	Intron	2:230232414	C	T	C	Susceptible	1.16E-16	1.13	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	[80] [103]
SP140	rs7423615	Intron	2:230252159	C	T	T	Susceptible	3.10E-13	1.12	European	[136]
SPATA48_IKZF1	rs1456893	Intron	7:50230076	A	G	A	Susceptible	4.60E-9	1.2	European (US-Canadian, Belgian-French, British)	[137] [139]
SPEF2	rs7710284	Missense	5:35692673	A	T	A		0.008		Germany	[101]

(cont. on the next page)

Appendix A (cont.)

STAT3	rs4796791	Intron	17:42378745	T	C	T	Protective	0.000152	0.84	European(UK)	144
STAT3	rs744166	Intron	17:42362183	G	A	A	Susceptible	6.82E-12	1.18	European (US-Canadian, Belgian-French, British)	137 140
STAT3	rs12942547	Intron	17:42375526	G	A	A	Susceptible	1.57E-9	1.3	Japanese (European)	138
STAT4	rs1517352	Intron	2:191066738	A	C	C	Susceptible	0.0221	1.1	Japanese (European)	138
STAT5B	rs7220367	Intron	17:42283095	C	G	G	Protective	2.23E-5	0.75	African Americans - Caucasian	103
STX8	rs9895062	Intron	17:9467306	A	G	A	Protective	0.048	0.85	German	140
SVIL2P	rs1826619	Noncoding Transcript	10:30712565	A	C	A		0.008		Germany	101
TAB2, TAB2-AS1	rs7758080	Intron	6:149255943	A	G	G	Susceptible	7.27E-9	1.08	European, East Asian, Indian, Iranian	104
TAGAP-AS1	rs212388	Intron	6:159069404	C	T	C	Susceptible	2.30E-11	1.1	European	80 103 136
TAP2	rs241427	Intron	6:32836637	G	A	A	Susceptible	0.00616	1.12	Japanese	138
TEX41	rs11681525	Intron	2:144734815	C	G	C	Protective	4.08E-11	0.86	European, East Asian, Indian, Iranian	104
THADA	rs10495903	Intron	2:43579779	C	T	T	Susceptible	1.60E-14	1.14	European	136
THEMIS	rs13204742	Intergenic	6:127924620	G	T	T	Susceptible	8.38E-15	1.17	Australia, Belgium, France, Canada, Denmark, Germany, Italy, Lithuania/Baltic, Netherlands, New Zealand, Norway, Slovenia, Spain, Sweden, UK, USA	80 103
THRAP3	rs6425977	Missense	1:36286832	C	T	C		0.004		Germany	101
TINAG	rs1058768	Missense	6:54321349	C	T	T		1.7E-12		Germany	101

(cont. on the next page)

Appendix A (cont.)

TMEM258	rs102275	Noncoding Transcript Exon	11:61790331	C	T	C	Susceptible	2.30E-11	1.08	European	[136]
TNFSF15	rs4246905	Intron	9:114790969	T	C	C	Susceptible	7.25E-28	1.81	Japanese	[138]
TNFSF15	rs4263839	Intron	9:114804160	G	A	G	Susceptible	2.60E-10	1.22	European (US-Canadian, Belgian-French, British)	[137] [139]
TNFSF15	rs6478108	Intron	9:114796423	T	C	T	Susceptible	8.19E-42	1.95	Japanese	[138]
TNFSF15	rs7869487	Intergenic	9:114818634	T	C	T	Protective	0.01	0.87	German	[140]
TNFSF15	rs3810936	Synonymous	9:114790605	C	T	C	Susceptible	1.00E-15	1.21	European / Japanese	[135] [136]
TNFSF18	rs12037606	Intron	1:172929262	A	G	A	Susceptible	1.09E-05	1.52	European British	[102]
TPPP2,NDRG2	rs9624	Missense, Intron	14:21031962	G	T	T		0.008		Germany	[101]
TRA,TRAV12-2	rs10483261	Missense	14:21888371	C	T	T		0.001		Germany	[101]
TRIB1AL, LINC02964	rs1551398	Intron	8:125527809	G	A	A	Susceptible	4.50E-9	1.08	European (US-Canadian, Belgian-French, British)	[137] [139]
TRIB1AL, LINC02964	rs4871611	Intron	8:125525328	G	A	A	Susceptible	1.51E-12	1.17	European	[136]
TTN,TTN-AS1	rs10497517	Synonymous, Intron	2:178578813	A	G	G		3.4E-07		Germany	[101]
TUBD1	rs1292053	Missense	17:59886176	G	A	G	Susceptible	0.000201	1.18	Japanese	[138]
TXNDC11	rs3190321	Missense	16:11679806	G	C	G		0.002		Germany	[101]
TYK2	rs12720356	Missense	19:10359299	A	C	C	Susceptible	1.40E-12	1.12	European	[136]
UBAP2	rs1785506	Missense	9:34017108	C	T	C		0.006		Germany	[101]
UBE2L3	rs181359	Intron	22:21574352	G	A	A	Susceptible	4.80E-16	1.1	European	[136]
UBQLN4	rs2297792	Missense	1:156041653	C	T	C		0.002		Germany	[101]

(cont. on the next page)

Appendix A (cont.)

UQCRHP1 (AIF1)	rs9348876	Intergenic	6:31607499	T	C	T	Susceptible	2.6E-6	1.41	Southern European (Spain)	[143]
USP16	rs2274802	Missense	21:29036349	A	T	T		0.009		Germany	[101]
WARS2-AS1	rs2003365	Noncoding Transcript	1:119219414	T	C	C		0.008		Germany	[101]
ZBTB38	rs724016	5primeUTR	3:141386728	A	G	G	Susceptible	3.36E-6	1.06	European, East Asian, Indian, Iranian	[104]
ZGPAT	rs4809330	Intron	20:63718234	A	G	G	Susceptible	2.70E-15	1.12	European / Japanese	[135] [136]
ZGPAT	rs6062504	Intron	20:63717555	A	G	G	Susceptible	0.0133	1.11	Japanese (European)	[138]
ZMIZ1	rs1250546	Intron	10:79272775	G	A	A	Susceptible	7.23E-5	1.19	Japanese	[138]
ZMIZ1	rs1250550	Intron	10:79300560	C	A	C	Susceptible	1.10E-30	1.19	European / Japanese	[135] [136]
ZNF300	rs7724036	Upstream	5:150905246	T	C	C	Susceptible	3.47E-10	1.39	German	[140]
ZNF300,IRGM	rs11741861	Intron	5:150898347	A	G	G	Susceptible	0.00875	1.12	Japanese (European)	[138]
ZNF300P1	rs17800886	Downstream	5:150929977	C	G	C	Susceptible	0.0003	1.35	German	[140]
ZNF365	rs10995271	Downstream	10:62678726	G	C	C	Susceptible	4.46E-20	1.25	European (US-Canadian, Belgian-French, British)	[137] [139]
ZPBP2	rs2872507	Intergenic	17:39884510	G	A	A	Susceptible	5.00E-9	1.12	European (US-Canadian, Belgian-French, British)	[137] [139]

(The "_" character indicates that the SNP of interest is between two genes. The "," character indicates that the SNP of interest is present in both genes. The gene within the "()" character represents that it is related to or located near the SNP of interest .)

APPENDIX B. GENE ONTOLOGY ANALYSIS OF CROHN'S DISEASE RELATED GENES

Gene Symbol	Panther Family Name	Panther Protein Class	Panther GO-Slim Molecular Function	PANTHER GO-Slim Biological Process	Panther GO-Slim Cellular Component	Pathway
ADAM10	DISINTEGRIN AND METALLOPROTEINASE DOMAIN-CONTAINING PROTEIN 10 (PTHR45702:SF4)	protease (PC00190)	metalloendopeptidase activity (GO:0004222)	Notch signaling pathway (GO:0007219); membrane protein ectodomain proteolysis (GO:0006509)	synaptic membrane (GO:0097060)	Alzheimer disease-amyloid secretase pathway->A disintegrase and metalloprotease 10 Notch signaling pathway->TNF-alpha converting enzyme
ADAM30	DISINTEGRIN AND METALLOPROTEINASE DOMAIN-CONTAINING PROTEIN 30 (PTHR11905:SF148)	metalloprotease (PC00153)		male gonad development (GO:0008584)	leaflet of membrane bilayer (GO:0097478); external side of plasma membrane (GO:0009897); plasma membrane region (GO:0098590)	
AIF1	ALLOGRAFT INFLAMMATORY FACTOR 1 (PTHR10356:SF4)	calmodulin-related (PC00061)	actin filament binding (GO:0051015); calcium ion binding (GO:0005509)	plasma membrane bounded cell projection assembly (GO:0120031); actin filament bundle assembly (GO:0051017)	ruffle membrane (GO:0032587); actin filament (GO:0005884)	
ANKRD55	ANKYRIN REPEAT DOMAIN-CONTAINING PROTEIN 55 (PTHR24123:SF85)	scaffold/adaptor protein (PC00226)				
AQP9	AQUAPORIN-9 (PTHR43829:SF6)	transporter (PC00227)	channel activity (GO:0015267); carbohydrate transmembrane transporter activity (GO:0015144);	organic hydroxy compound transport (GO:0015850); carbohydrate transport (GO:0008643)	basolateral plasma membrane (GO:0016323)	

(cont. on the next page)

Appendix B (cont.)

ARFRP1	ADP-RIBOSYLATION FACTOR-RELATED PROTEIN 1 (PTHR45909:SF1)			Golgi to plasma membrane protein transport (GO:0043001); protein localization to organelle (GO:0033365); intracellular protein transport (GO:0006886)	Golgi apparatus (GO:0005794); vacuole (GO:0005773); plasma membrane (GO:0005886)	
ATG16L1	AUTOPHAGY-RELATED PROTEIN 16-1 (PTHR19878:SF6)			autophagosome assembly (GO:0000045); proteolysis (GO:0006508)	vacuolar membrane (GO:0005774); autophagosome (GO:0005776)	
ATG16L2	PROTEIN ATG16L2 (PTHR19878:SF7)			autophagosome assembly (GO:0000045); proteolysis (GO:0006508)	vacuolar membrane (GO:0005774); autophagosome (GO:0005776)	
ATP23	MITOCHONDRIAL INNER MEMBRANE PROTEASE ATP23 HOMOLOG (PTHR21711:SF0)	metalloprotease (PC00153)		mitochondrial proton-transporting ATP synthase complex assembly (GO:0033615); mitochondrial protein processing (GO:0034982)	extrinsic component of membrane (GO:0019898); mitochondrial inner membrane (GO:0005743)	
BACH2	TRANSCRIPTION REGULATOR PROTEIN BACH2 (PTHR46105:SF8)	C2H2 zinc finger transcription factor (PC00248)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978); DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)	transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)		

(cont. on the next page)

Appendix B (cont.)

BSN	PROTEIN BASSOON (PTHR14113:SF1)		structural molecule activity (GO:0005198)	synapse assembly (GO:0007416); protein localization to cell junction (GO:1902414)	presynaptic active zone (GO:0048786); cytoplasmic region (GO:0099568); plasma membrane region (GO:0098590); glutamatergic synapse (GO:0098978); axon(GO:0030424)	
BTNL2	BTNL2-RELATED (PTHR24100:SF105)	immunoglobulin receptor superfamily (PC00124)	signaling receptor binding (GO:0005102)	immune response (GO:0006955); cytokine production (GO:0001816); regulation of cytokine production (GO:0001817); T cell receptor signaling pathway (GO:0050852)	leaflet of membrane bilayer (GO:0097478); external side of plasma membrane (GO:0009897)	
CAVIN1	CAVEOLAE-ASSOCIATED PROTEIN 1 (PTHR15240:SF3)	membrane traffic protein (PC00150)	rRNA binding (GO:0019843)	transcription by RNA polymerase I (GO:0006360); DNA-templated transcription, initiation (GO:0006352)	caveola (GO:0005901); cytoplasm (GO:0005737)	General transcription by RNA polymerase I->PTRF
CL2	C-C MOTIF CHEMOKINE 2 (PTHR12015:SF98)	cytokine (PC00083)	signaling receptor activity (GO:0038023); cytokine activity (GO:0005125); chemokine receptor binding (GO:0042379)	ERK1 and ERK2 cascade (GO:0070371); neutrophil migration (GO:1990266); inflammatory response (GO:0006954); G protein-coupled receptor signaling pathway (GO:0007186); cytokine-mediated signaling pathway (GO:0019221); innate immune response (GO:0045087); positive regulation of ERK1 and ERK2 cascade (GO:0070374); response to interleukin-1 (GO:0070555); cellular response to tumor necrosis factor (GO:0071356); granulocyte chemotaxis (GO:0071621); lymphocyte migration (GO:0072676); positive regulation of GTPase activity (GO:0043547)	extracellular space (GO:0005615)	Inflammation mediated by chemokine and cytokine signaling pathway->Chemokine

(cont. on the next page)

Appendix B (cont.)

CCL7	C-C MOTIF CHEMOKINE 7 (PTHR12015:SF161)	cytokine (PC00083)	signaling receptor activity (GO:0038023); cytokine activity (GO:0005125); chemokine receptor binding (GO:0042379)	ERK1 and ERK2 cascade (GO:0070371); neutrophil migration (GO:1990266); inflammatory response (GO:0006954); G protein-coupled receptor signaling pathway (GO:0007186); cytokine-mediated signaling pathway (GO:0019221); innate immune response (GO:0045087); positive regulation of ERK1 and ERK2 cascade (GO:0070374); response to interleukin-1 (GO:0070555); cellular response to tumor necrosis factor (GO:0071356); granulocyte chemotaxis (GO:0071621); lymphocyte migration (GO:0072676)	extracellular space (GO:0005615)	Inflammation mediated by chemokine and cytokine signaling pathway->Chemokine
CCNY	CYCLIN-Y (PTHR14248:SF33)		cyclin binding (GO:0030332); protein kinase binding (GO:0019901); protein serine/threonine kinase activator activity (GO:0043539); cyclin-dependent protein serine/threonine kinase activity (GO:0004693)	regulation of protein kinase activity (GO:0045859); canonical Wnt signaling pathway (GO:0060070); regulation of canonical Wnt signaling pathway (GO:0060828); protein phosphorylation (GO:0006468)	cytoplasm (GO:0005737); plasma membrane (GO:0005886);	
CDKAL1	THREONYLCARBAMOYLADENOSINE TRNA METHYLTHIOTRANSFERASE (PTHR11918:SF45)		transferase activity (GO:0016740)	tRNA modification (GO:0006400)		

(cont. on the next page)

Appendix B (cont.)

CEBPA	CCAAT/ENHANCER-BINDING PROTEIN ALPHA (PTHR23334:SF5)	basic leucine zipper transcription factor (PC00056)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978); DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)	myeloid cell differentiation (GO:0030099); transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)		
CPEB4	CYTOPLASMIC POLYADENYLATION ELEMENT-BINDING PROTEIN 4 (PTHR12566:SF2)	mRNA polyadenylation factor (PC00146)	ribosome binding (GO:0043022); translation factor activity, RNA binding (GO:0008135); mRNA 3'-UTR binding (GO:0003730)	cytoplasmic translation (GO:0002181); negative regulation of translation (GO:0017148); translational elongation (GO:0006414)	nucleus (GO:0005634); cytoplasm (GO:0005737); plasma membrane region (GO:0098590); synapse (GO:0045202); ribonucleoprotein complex (GO:1990904); neuron projection (GO:0043005)	
CUL2	CULLIN-2 (PTHR11932:SF126)	ubiquitin-protein ligase (PC00234)	ubiquitin protein ligase binding (GO:0031625)	protein ubiquitination (GO:0016567)	cullin-RING ubiquitin ligase complex (GO:0031461)	Parkinson disease->F-box and WD-40 domain protein 7
CXCL5	CHEMOKINE 5 (PTHR10179:SF86)	chemokine (PC00074)	signaling receptor activity (GO:0038023); cytokine activity (GO:0005125); chemokine receptor binding (GO:0042379)	neutrophil migration (GO:1990266); inflammatory response (GO:0006954); cellular response to lipopolysaccharide (GO:0071222); cytokine-mediated signaling pathway (GO:0019221); granulocyte chemotaxis (GO:0071621); antimicrobial humoral immune response mediated by antimicrobial peptide (GO:0061844)	extracellular space (GO:0005615)	
CXCR5	C-X-C CHEMOKINE RECEPTOR TYPE 5 (PTHR10489:SF618)	cell adhesion molecule (PC00069)		immune response (GO:0006955)		Inflammation mediated by chemokine and cytokine signaling pathway->Chemokine receptor

(cont. on the next page)

Appendix B (cont.)

DCP1B	MRNA-DECAPPING ENZYME 1B (PTHR16290:SF5)	mRNA capping factor (PC00145)	mRNA binding (GO:0003729)	nucleic acid phosphodiester bond hydrolysis (GO:0090305); RNA decapping (GO:0110154); translational elongation (GO:0006414); nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay (GO:0000288)	P-body (GO:0000932)	TGF-beta signaling pathway->Co-activators or corepressors
DENND1B	DENN DOMAIN-CONTAINING PROTEIN 1B (PTHR13196:SF24)		phosphatidylinositol phosphate binding (GO:1901981)	vesicle budding from membrane (GO:0006900); membrane invagination (GO:0010324); endocytosis (GO:0006897); endocytic recycling (GO:0032456)	nuclear speck (GO:0016607); membrane (GO:0016020); cytosol (GO:0005829)	
DHX34	ATP-DEPENDENT RNA HELICASE DHX34-RELATED (PTHR18934:SF148)	RNA helicase (PC00032)	RNA binding (GO:0003723)	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (GO:0000184); gene expression (GO:0010467)	intracellular anatomical structure (GO:0005622)	
DNMT3A	DNA (CYTOSINE-5)-METHYLTRANSFERASE 3A (PTHR23068:SF10)	DNA methyltransferase (PC00013)		negative regulation of transcription, DNA-templated (GO:0045892); transcription, DNA-templated (GO:0006351); DNA metabolic process (GO:0006259); macromolecule methylation (GO:0043414)	nucleus (GO:0005634); cytoplasm (GO:0005737); membrane (GO:0016020)	
DOCK7	DEDICATOR OF CYTOKINESIS PROTEIN 7 (PTHR23317:SF78)	guanyl-nucleotide exchange factor (PC00113)	GTP binding (GO:0005525); protein binding (GO:0005515); GTPase activity (GO:0003924); guanyl-nucleotide exchange factor activity (GO:0005085); GDP binding (GO:0019003)	axonogenesis (GO:0007409)		

(cont. on the next page)

Appendix B (cont.)

ELF1	ETS-RELATED TRANSCRIPTION FACTOR ELF-1 (PTHR11849:SF156)	winged helix/forkhead transcription factor (PC00246)	RNA polymerase II transcription regulatory region DNA binding (GO:0000977); DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)	cell differentiation (GO:0030154); transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)	nucleus (GO:0005634); membrane (GO:0016020)	PDGF signaling pathway->Ets
EMSY	BRCA2-INTERACTING TRANSCRIPTIONAL REPRESSOR EMSY (PTHR16500:SF3)	DNA-binding transcription factor (PC00218)			membrane (GO:0016020); nucleoplasm (GO:0005654)	
EPO	ERYTHROPOIETIN (PTHR10370:SF0)		signaling receptor activity (GO:0038023); cytokine activity (GO:0005125)		extracellular space (GO:0005615)	
ERAP1	ENDOPLASMIC RETICULUM AMINOPEPTIDASE 1 (PTHR11533:SF156)	metalloprotease (PC00153)	zinc ion binding (GO:0008270); aminopeptidase activity (GO:0004177); peptide binding (GO:0042277); metallopeptidase activity (GO:0008237)	peptide metabolic process (GO:0006518); proteolysis (GO:0006508); organonitrogen compound catabolic process (GO:1901565); cellular catabolic process (GO:0044248)	cytoplasm (GO:0005737)	
ERAP2	ENDOPLASMIC RETICULUM AMINOPEPTIDASE 2 (PTHR11533:SF239)	metalloprotease (PC00153)	zinc ion binding (GO:0008270); aminopeptidase activity (GO:0004177); peptide binding (GO:0042277); metallopeptidase activity (GO:0008237)	peptide metabolic process (GO:0006518); proteolysis (GO:0006508); organonitrogen compound catabolic process (GO:1901565); cellular catabolic process (GO:0044248)	cytoplasm (GO:0005737)	

(cont. on the next page)

Appendix B (cont.)

FAF1	FAS-ASSOCIATED FACTOR 1 (PTHR23322:SF56)	scaffold/adaptor protein (PC00226)	NF-kappaB binding (GO:0051059); ubiquitin binding (GO:0043130)	cell death (GO:0008219); positive regulation of cell death (GO:0010942); ubiquitin-dependent ERAD pathway (GO:0030433)	nucleus (GO:0005634); endoplasmic reticulum (GO:0005783); vacuole (GO:0005773); plasma membrane (GO:0005886)	FAS signaling pathway->FAF1
FGL1	FIBRINOGEN-LIKE PROTEIN 1 (PTHR19143:SF263)	intercellular signal molecule (PC00207)		immune response (GO:0006955); T cell activation (GO:0042110); negative regulation of T cell activation (GO:0050868); leukocyte cell-cell adhesion (GO:0007159)	extracellular space (GO:0005615); collagen-containing extracellular matrix (GO:0062023)	
FIBP	ACIDIC FIBROBLAST GROWTH FACTOR INTRACELLULAR-BINDING PROTEIN (PTHR13223:SF2)				nucleus (GO:0005634); membrane (GO:0016020)	
FOXP2	FORKHEAD BOX PROTEIN P2 (PTHR45796:SF9)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978); RNA polymerase II-specific (GO:0001227)	transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)	nucleus (GO:0005634); membrane (GO:0016020)		
FUCA1	TISSUE ALPHA-L-FUCOSIDASE (PTHR10030:SF2)	glycosidase (PC00110)	hydrolase activity, hydrolyzing O-glycosyl compounds (GO:0004553)	hexose metabolic process (GO:0019318); glycosyl compound catabolic process (GO:1901658)	membrane (GO:0016020); lysosome (GO:0005764)	

(cont. on the next page)

Appendix B (cont.)

FUT2	GALACTOSIDE ALPHA-(1,2)-FUCOSYLTRANSFERASE 2 (PTHR11927:SF2)	glycosyltransferase (PC00111)	fucosyltransferase activity (GO:0008417)	protein glycosylation (GO:0006486)		
GAL3ST2	GALACTOSE-3-O-SULFOTRANSFERASE 2 (PTHR14647:SF55)	transferase (PC00220)	sulfotransferase activity (GO:0008146)	glycoprotein biosynthetic process (GO:0009101)		
GATA6	TRANSCRIPTION FACTOR GATA-6 (PTHR10071:SF23)	DNA-binding transcription factor (PC00218)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978); DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)	epithelial cell differentiation (GO:0030855); negative regulation of transcription by RNA polymerase II (GO:0000122); cell fate commitment (GO:0045165); transcription by RNA polymerase II (GO:0006366); positive regulation of transcription by RNA polymerase II (GO:0045944)		
GP2	PANCREATIC SECRETORY GRANULE MEMBRANE MAJOR GLYCOPROTEIN GP2 (PTHR14002:SF16)	transmembrane signal receptor (PC00197)		neutrophil migration (GO:1990266)	extracellular space (GO:0005615); apical plasma membrane (GO:0016324); cell surface (GO:0009986)	
GPR35	G-PROTEIN COUPLED RECEPTOR 35 (PTHR24232:SF54)	G-protein coupled receptor (PC00021)	G protein-coupled receptor activity (GO:0004930)	regulation of Rho protein signal transduction (GO:0035023); inositol phosphate-mediated signaling (GO:0048016); activation of phospholipase C activity (GO:0007202); positive regulation of Ras protein signal transduction (GO:0046579); Rho protein signal transduction (GO:0007266)	integral component of plasma membrane (GO:0005887)	

(cont. on the next page)

Appendix B (cont.)

GRP	GASTRIN-RELEASING PEPTIDE (PTHR16866:SF2)		neuropeptide hormone activity (GO:0005184); signaling receptor activity (GO:0038023)	neuropeptide signaling pathway (GO:0007218)	extracellular space (GO:0005615)	
HERC2	E3 UBIQUITIN-PROTEIN LIGASE HERC2 (PTHR22870:SF398)	guanyl-nucleotide exchange factor (PC00113)				
HLA-DQB1	HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, DQ BETA 1 CHAIN (PTHR19944:SF101)	major histocompatibility complex protein (PC00149)	protein-containing complex binding (GO:0044877)	positive regulation of T cell activation (GO:0050870); immunoglobulin mediated immune response (GO:0016064); T cell activation (GO:0042110); cellular protein-containing complex assembly (GO:0034622); antigen processing and presentation (GO:0019882); production of molecular mediator of immune response (GO:0002440); leukocyte cell-cell adhesion (GO:0007159)	plasma membrane protein complex (GO:0098797)	
HLA-DRA	HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, DR ALPHA CHAIN-RELATED (PTHR19944:SF86)	major histocompatibility complex protein (PC00149)	protein-containing complex binding (GO:0044877)	positive regulation of T cell activation (GO:0050870); immunoglobulin mediated immune response (GO:0016064); T cell activation (GO:0042110); cellular protein-containing complex assembly (GO:0034622); antigen processing and presentation (GO:0019882); production of molecular mediator of immune response (GO:0002440); eukocyte cell-cell adhesion (GO:0007159)	plasma membrane protein complex (GO:0098797)	T cell activation->MHC-antigen
HSP90AB4P	HEAT SHOCK PROTEIN HSP 90-BETA 2-RELATED (PTHR11528:SF78)	Hsp90 family chaperone (PC00028)	ATP binding (GO:0005524); unfolded protein binding (GO:0051082)	cellular response to heat (GO:0034605); protein folding (GO:0006457); protein stabilization (GO:0050821)	perinuclear region of cytoplasm (GO:0048471); cytosol (GO:0005829); plasma membrane (GO:0005886)	

(cont. on the next page)

Appendix B (cont.)

ICOSLG	ICOS LIGAND (PTHR24100:SF55)	immunoglobulin receptor superfamily (PC00124)				
IFNGR2	INTERFERON GAMMA RECEPTOR 2 (PTHR20859:SF46)	transmembrane signal receptor (PC00197)				Inflammation mediated by chemokine and cytokine signaling pathway->Cytokine receptor; Interferon-gamma signaling pathway->Interferon-gamma receptor; Interferon-gamma signaling pathway
IKZF1	DNA-BINDING PROTEIN IKAROS (PTHR24404:SF36)	C2H2 zinc finger transcription factor (PC00248)	DNA-binding transcription factor activity (GO:0003700); RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978)	transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)		
IKZF3	ZINC FINGER PROTEIN AIOLOS (PTHR24404:SF23)	C2H2 zinc finger transcription factor (PC00248)	DNA-binding transcription factor activity (GO:0003700); RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978)	transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)		
IL12B	INTERLEUKIN-12 SUBUNIT BETA (PTHR23036:SF156)	transmembrane signal receptor (PC00197)	cytokine binding (GO:0019955); cytokine receptor activity (GO:0004896)	cytokine-mediated signaling pathway (GO:0019221)	leaflet of membrane bilayer (GO:0097478); external side of plasma membrane (GO:0009897)	

(cont. on the next page)

Appendix B (cont.)

IL12RB2	INTERLEUKIN-12 RECEPTOR SUBUNIT BETA-2 (PTHR23036:SF79)	transmembrane signal receptor (PC00197)	cytokine binding (GO:0019955); cytokine receptor activity (GO:0004896)	cytokine-mediated signaling pathway (GO:0019221)	leaflet of membrane bilayer (GO:0097478); external side of plasma membrane (GO:0009897)	Interleukin signaling pathway->Interleukin receptor beta subunit
IL18RAP	INTERLEUKIN-18 RECEPTOR ACCESSORY PROTEIN (PTHR11890:SF23)	transmembrane signal receptor (PC00197)	cytokine binding (GO:0019955); cytokine receptor activity (GO:0004896)	transcription, DNA-templated (GO:0006351); cytokine-mediated signaling pathway (GO:0019221); positive regulation of NF-kappaB transcription factor activity (GO:0051092)		
IL23R	INTERLEUKIN-23 RECEPTOR (PTHR23036:SF112)	transmembrane signal receptor (PC00197)	cytokine binding (GO:0019955); cytokine receptor activity (GO:0004896)	cytokine-mediated signaling pathway (GO:0019221)	leaflet of membrane bilayer (GO:0097478); external side of plasma membrane (GO:0009897); receptor complex (GO:0043235)	
IL2RA	INTERLEUKIN-2 RECEPTOR SUBUNIT ALPHA (PTHR10573:SF0)	transmembrane signal receptor (PC00197)	cytokine binding (GO:0019955); cytokine receptor activity (GO:0004896); growth factor binding (GO:0019838)	inflammatory response (GO:0006954)		Interleukin signaling pathway->Interleukin receptor beta subunit
IL7R	INTERLEUKIN-7 RECEPTOR SUBUNIT ALPHA (PTHR23037:SF27)	transmembrane signal receptor (PC00197)	cytokine binding (GO:0019955); cytokine receptor activity (GO:0004896)	positive regulation of receptor signaling pathway via JAK-STAT (GO:0046427); cytokine-mediated signaling pathway (GO:0019221); hemopoiesis (GO:0030097); receptor signaling pathway via JAK-STAT (GO:0007259)	external side of plasma membrane (GO:0009897); leaflet of membrane bilayer (GO:0097478)	
INAVA	INNATE IMMUNITY ACTIVATOR PROTEIN (PTHR16093:SF4)				adherens junction organization (GO:0034332); positive regulation of protein ubiquitination (GO:0031398); protein ubiquitination (GO:0016567)	

(cont. on the next page)

Appendix B (cont.)

IPMK	INOSITOL POLYPHOSPHATE MULTIKINASE (PTHR12400:SF80)	kinase (PC00137)	phosphotransferase activity, alcohol group as acceptor (GO:0016773); kinase activity (GO:0016301)	alcohol biosynthetic process (GO:0046165); phosphate-containing compound metabolic process (GO:0006796); organophosphate biosynthetic process (GO:0090407)	membrane (GO:0016020); nucleus (GO:0005634); cytoplasm (GO:0005737)	
IRF1	INTERFERON REGULATORY FACTOR 1 (PTHR11949:SF3)	winged helix/forkhead transcription factor (PC00246)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978)	transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)	membrane (GO:0016020); nucleus (GO:0005634)	
IRF4	INTERFERON REGULATORY FACTOR 4 (PTHR11949:SF6)	winged helix/forkhead transcription factor (PC00246)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978)	transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)	membrane (GO:0016020); nucleus (GO:0005634)	
IRGM	IMMUNITY-RELATED GTPASE FAMILY M PROTEIN (PTHR32341:SF9)		GTPase activity (GO:0003924)	cellular response to cytokine stimulus (GO:0071345); defense response (GO:0006952)	endoplasmic reticulum membrane (GO:0005789); vacuole (GO:0005773); plasma membrane (GO:0005886)	
JAK2	TYROSINE-PROTEIN KINASE JAK2 (PTHR45807:SF1)	non-receptor tyrosine protein kinase (PC00168)	non-membrane spanning protein tyrosine kinase activity (GO:0004715); cytokine receptor binding (GO:0005126)	peptidyl-tyrosine phosphorylation (GO:0018108); cytokine-mediated signaling pathway (GO:0019221); hemopoiesis (GO:0030097); cell differentiation (GO:0030154); intracellular signal transduction (GO:0035556); receptor signaling pathway via JAK-STAT (GO:0007259)	cytosol (GO:0005829)	Inflammation mediated by chemokine and cytokine signaling pathway->Janus kinase; PI3 kinase pathway->JAK; PDGF signaling pathway->Jak; JAK/STAT signaling pathway->Jak; Interferon-gamma signaling pathway->Janus kinase 2; CCKR signaling map->JAK2

(cont. on the next page)

Appendix B (cont.)

JAZF1	JUXTAPOSED WITH ANOTHER ZINC FINGER PROTEIN 1 (PTHR23057:SF0)				nucleus (GO:0005634); membrane (GO:0016020)	
KSR1	KINASE SUPPRESSOR OF RAS 1 (PTHR23257:SF716)	non-receptor serine/threonine protein kinase (PC00167)	protein kinase activity (GO:0004672)	Ras protein signal transduction (GO:0007265)	cytoplasm (GO:0005737)	Gonadotropin-releasing hormone receptor pathway->KSR-1
LACC1	PURINE NUCLEOSIDE PHOSPHORYLASE LACC1 (PTHR30616:SF2)		copper ion binding (GO:0005507)			
LRRK2	LEUCINE-RICH REPEAT SERINE/THREONINE-PROTEIN KINASE 2 (PTHR45752:SF7)	scaffold/adaptor protein (PC00226)	signaling receptor binding (GO:0005102)	negative regulation of NF-kappaB transcription factor activity (GO:0032088); transcription, DNA-templated (GO:0006351); cellular response to tumor necrosis factor (GO:0071356)	basolateral plasma membrane (GO:0016323)	
LTA	LYMPHOTOXIN-ALPHA (PTHR11471:SF31)	intercellular signal molecule (PC00207)				Inflammation mediated by chemokine and cytokine signaling pathway->Interleukin 2; Apoptosis signaling pathway->Tumor necrosis factor
LYRM4	LYR MOTIF-CONTAINING PROTEIN 4 (PTHR13166:SF7)			iron-sulfur cluster assembly (GO:0016226)	membrane (GO:0016020); mitochondrion (GO:0005739); cytosol (GO:0005829); protein-containing complex (GO:0032991)	

(cont. on the next page)

Appendix B (cont.)

MACROD2	ADP-RIBOSE GLYCOHYDROLASE MACROD2 (PTHR11106:SF104)		hydrolase activity, acting on glycosyl bonds (GO:0016798)	cellular protein modification process (GO:0006464); purine nucleoside metabolic process (GO:0042278); cellular response to DNA damage stimulus (GO:0006974)	membrane (GO:0016020); nucleoplasm (GO:0005654)	
MAMSTR	LYSINE-RICH CEACAM1 CO-ISOLATED PROTEIN LYRIC PROTEIN (PTHR23251)		transcription coregulator activity (GO:0003712)	regulation of transcription by RNA polymerase II (GO:0006357)	nucleus (GO:0005634)	
METTL4	N(6)-ADENINE-SPECIFIC METHYLTRANSFERASE METTL4 (PTHR12829:SF4)	RNA methyltransferase (PC00033)	methyltransferase activity (GO:0008168)	RNA methylation (GO:0001510); mRNA metabolic process (GO:0016071)	nucleus (GO:0005634); membrane (GO:0016020); methyltransferase complex (GO:0034708)	
MFSD12	MAJOR FACILITATOR SUPERFAMILY DOMAIN-CONTAINING PROTEIN 12 (PTHR11328:SF28)			organic substance transport (GO:0071702)	integral component of plasma membrane (GO:0005887)	
MST1	HEPATOCYTE GROWTH FACTOR-LIKE PROTEIN-RELATED (PTHR24261:SF12)	serine protease (PC00203)	endopeptidase activity (GO:0004175); receptor tyrosine kinase binding (GO:0030971)	regulation of receptor signaling pathway via JAK-STAT (GO:0046425); receptor signaling pathway via JAK-STAT (GO:0007259)	extracellular space (GO:0005615)	

(cont. on the next page)

Appendix B (cont.)

NCF4	NEUTROPHIL CYTOSOL FACTOR 4 (PTHR15706:SF20)	scaffold/adaptor protein (PC00226)	protein binding (GO:0005515); enzyme activator activity (GO:0008047); oxidoreductase activity, acting on NAD(P)H (GO:0016651)	cellular metabolic process (GO:0044237)		
NDRG2	PROTEIN NDRG2 (PTHR11034:SF17)	serine protease (PC00203)		signal transduction (GO:0007165)	cytoplasm (GO:0005737)	
NKX2-3	HOMEOBOX PROTEIN NKX-2.3 (PTHR24340:SF32)	homeodomain transcription factor (PC00119)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978); DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)	cell differentiation (GO:0030154); transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)	nucleus (GO:0005634); membrane (GO:0016020)	
OAS2	2'-5'-OLIGOADENYLATE SYNTHASE 2 (PTHR11258:SF3)	nucleotidyltransferase (PC00174)	double-stranded RNA binding (GO:0003725); adenylyltransferase activity (GO:0070566)	RNA phosphodiester bond hydrolysis (GO:0090501); regulation of RNA metabolic process (GO:0051252); defense response to virus (GO:0051607); negative regulation of biological process (GO:0048519)	membrane (GO:0016020); cytosol (GO:0005829); nucleoplasm (GO:0005654)	
OR5U1	OLFACTORY RECEPTOR 14J1 (PTHR26452:SF42)	G-protein coupled receptor (PC00021)	odorant binding (GO:0005549); olfactory receptor activity (GO:0004984)			
PDE4A	CAMP-SPECIFIC 3',5'-CYCLIC PHOSPHODIESTERASE 4A (PTHR11347:SF74)	phosphodiesterase (PC00185)	3',5'-cyclic-AMP phosphodiesterase activity (GO:0004115)		nucleus (GO:0005634); perinuclear region of cytoplasm (GO:0048471); membrane (GO:0016020); cytosol (GO:0005829)	

(cont. on the next page)

Appendix B (cont.)

PDGFB	PLATELET-DERIVED GROWTH FACTOR SUBUNIT B (PTHR11633:SF2)	growth factor (PC00112)	growth factor receptor binding (GO:0070851)	ERK1 and ERK2 cascade (GO:0070371); cell population proliferation (GO:0008283); regulation of MAP kinase activity (GO:0043405); regulation of phosphatidylinositol 3-kinase signaling (GO:0014066); phosphatidylinositol 3-kinase signaling (GO:0014065); positive regulation of ERK1 and ERK2 cascade (GO:0070374); positive regulation of cell population proliferation (GO:0008284); protein autophosphorylation (GO:0046777); transmembrane receptor protein tyrosine kinase signaling pathway (GO:0007169); cell migration (GO:0016477)	extracellular space (GO:0005615)	Angiogenesis->Platelet-Derived Growth Factor; PDGF signaling pathway->Platelet-Derived Growth Factor
PLA2G4A	CYTOSOLIC PHOSPHOLIPASE A2 (PTHR10728:SF13)	phospholipase (PC00186)	phospholipase A2 activity (GO:0004623); calcium ion binding (GO:0005509); calcium-dependent phospholipid binding (GO:0005544)	glycerophospholipid catabolic process	nucleus (GO:0005634); cytosol (GO:0005829); Golgi apparatus (GO:0005794); endoplasmic reticulum (GO:0005783); vacuole (GO:0005773); plasma membrane (GO:0005886)	Endothelin signaling pathway->Cytosolic phospholipase A2; CCKR signaling map->PLA2; Inflammation mediated by chemokine and cytokine signaling pathway->Phospholipase A2; VEGF signaling pathway->Cytoplasmic Phospholipase A2; Angiogenesis->Phospholipase A2; Gonadotropin-releasing hormone receptor pathway->PLA2; Oxidative stress response

(cont. on the next page)

Appendix B (cont.)

PRDM1	PR DOMAIN ZINC FINGER PROTEIN1 (PTHR16515:SF26)	C2H2 zinc finger transcription factor (PC00248)				
PTGER4	PROSTAGLANDIN E2 RECEPTOR EP4 SUBTYPE (PTHR11866:SF6)	G-protein coupled receptor (PC00021)	carboxylic acid binding (GO:0031406); G protein-coupled receptor activity (GO:0004930)	inflammatory response (GO:0006954); response to external stimulus (GO:0009605); regulation of cAMP-mediated signaling (GO:0043949); regulation of adenylate cyclase activity (GO:0045761); negative regulation of inflammatory response (GO:0050728); cellular response to hormone stimulus (GO:0032870); adenylate cyclase-activating G protein-coupled receptor signaling pathway (GO:0007189); cellular response to oxygen-containing compound (GO:1901701); activation of adenylate cyclase activity (GO:0007190); cAMP-mediated signaling (GO:0019933); positive regulation of cytosolic calcium ion concentration (GO:0007204)	plasma membrane (GO:0005886)	Gonadotropin-releasing hormone receptor pathway->PG Rcs
PTPN2	TYROSINE-PROTEIN PHOSPHATASE NON-RECEPTOR TYPE 2 (PTHR46047:SF1)	protein phosphatase (PC00195)				
PTPN22	TYROSINE-PROTEIN PHOSPHATASE NON-RECEPTOR TYPE 22 (PTHR45983:SF1)		protein tyrosine phosphatase activity (GO:0004725)	peptidyl-tyrosine dephosphorylation (GO:0035335)		

(cont. on the next page)

Appendix B (cont.)

PTPRC	RECEPTOR-TYPE TYROSINE-PROTEIN PHOSPHATASE C (PTHR19134:SF539)	protein phosphatase (PC00195)		peptidyl-tyrosine phosphorylation (GO:0018108); negative regulation of protein kinase activity (GO:0006469); positive regulation of peptidyl-tyrosine phosphorylation (GO:0050731)		T cell activation->CD45; B cell activation->CD45; JAK/STAT signaling pathway->protein-tyrosine phosphatase
PUS10	TRNA PSEUDOURIDINE SYNTHASE PUS10 (PTHR21568:SF0)		intramolecular transferase activity (GO:0016866)	pseudouridine synthesis (GO:0001522); tRNA modification (GO:0006400)		
RASGRP1	RAS GUANYL-RELEASING PROTEIN 1 (PTHR23113:SF174)	guanyl-nucleotide exchange factor (PC00113)	GTP binding (GO:0005525); protein binding (GO:0005515); GTPase activity (GO:0003924); guanyl-nucleotide exchange factor activity (GO:0005085)	Ras protein signal transduction (GO:0007265); positive regulation of GTPase activity (GO:0043547)	plasma membrane (GO:0005886)	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway->Calcium and diacylglycerol-regulated guanine nucleotide exchange factor
RNASET2	RIBONUCLEASE T2 (PTHR11240:SF22)	endoribonuclease (PC00094)				
S100Z	PROTEIN S100-Z (PTHR11639:SF72)	calmodulin-related (PC00061)			cytoplasm (GO:0005737)	
SLAIN2	SLAIN MOTIF-CONTAINING PROTEIN 2 (PTHR22406:SF4)			positive regulation of supramolecular fiber organization (GO:1902905); positive regulation of cytoskeleton organization (GO:0051495); microtubule nucleation (GO:0007020); positive regulation of protein polymerization (GO:0032273); regulation of microtubule polymerization or depolymerization (GO:0031110)	microtubule plus-end (GO:0035371)	

(cont. on the next page)

Appendix B (cont.)

SLC16A4	MONOCARBOXYLATE TRANSPORTER 5 (PTHR11360:SF14)	transporter (PC00227)	monocarboxylic acid transmembrane transporter activity (GO:0008028)	monocarboxylic acid transport (GO:0015718);	integral component of plasma membrane (GO:0005887)	
SLC17A3	SODIUM-DEPENDENT PHOSPHATE TRANSPORT PROTEIN 4 (PTHR11662:SF134)	secondary carrier transporter (PC00258)	xenobiotic transmembrane transporter activity (GO:0042910); inorganic anion transmembrane transporter activity (GO:0015103); ion channel activity (GO:0005216); efflux transmembrane transporter activity (GO:0015562); voltage-gated ion channel activity (GO:0005244)	organic anion transport (GO:0015711)	integral component of membrane (GO:0016021)	
SLC22A23	SOLUTE CARRIER FAMILY 22 MEMBER 23 (PTHR24064:SF192)	secondary carrier transporter (PC00258)				
SLC22A4	SOLUTE CARRIER FAMILY 22 MEMBER 4 (PTHR24064:SF222)	secondary carrier transporter (PC00258)				
SLC22A5	SOLUTE CARRIER FAMILY 22 MEMBER 5 (PTHR24064:SF283)	secondary carrier transporter (PC00258)				

(cont. on the next page)

Appendix B (cont.)

SLC7A10	ASC-TYPE AMINO ACID TRANSPORTER 1 (PTHR11785:SF73)	transporter (PC00227)	L-amino acid transmembrane transporter activity (GO:0015179); neutral amino acid transmembrane transporter activity (GO:0015175)	alanine transport (GO:0032328); amino acid transmembrane transport (GO:0003333)		
SMAD3	MOTHERS AGAINST DECAPENTAPLEGIC HOMOLOG 3 (PTHR13703:SF53)	DNA-binding transcription factor (PC00218)	SMAD binding (GO:0046332); RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978); DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)	BMP signaling pathway (GO:0030509); cell differentiation (GO:0030154); transforming growth factor beta receptor signaling pathway (GO:0007179); transcription by RNA polymerase II (GO:0006366); anatomical structure morphogenesis (GO:0009653); regulation of transcription by RNA polymerase II (GO:0009653)	RNA polymerase II transcription regulator complex (GO:0090575); membrane (membrane)	TGF-beta signaling pathway->Rsmads; Gonadotropin-releasing hormone receptor pathway->SMAD2-4
SP140	NUCLEAR BODY PROTEIN SP140 (PTHR46386:SF8)		RNA polymerase II transcription regulatory region sequence-specific DNA binding (GO:0000977); DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)	transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)	nucleus (GO:0005634); membrane (GO:0016020)	

(cont. on the next page)

Appendix B (cont.)

STAT5B	SIGNAL TRANSDUCER AND ACTIVATOR OF TRANSCRIPTION 5B (PTHR11801:SF39)	DNA-binding transcription factor (PC00218)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978); DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)	cell population proliferation (GO:0008283); cytokine-mediated signaling pathway (GO:0019221); cellular response to peptide hormone stimulus (GO:0071375); transcription by RNA polymerase II (GO:0006366); transmembrane receptor protein tyrosine kinase signaling pathway (GO:0007169); defense response (GO:0006952); regulation of cell population proliferation (GO:0042127); receptor signaling pathway via JAK-STAT (GO:0007259); regulation of transcription by RNA polymerase II (GO:0006357)		JAK/STAT signaling pathway->Signal transducers and activators of transcription; PDGF signaling pathway->Signal transducers and activators of transcription; EGF receptor signaling pathway->Signal transducers and activators of transcription; Interleukin signaling pathway->Signal transducers and activators of transcription
STX8	SYNTAXIN-8 (PTHR19957:SF285)	SNARE protein (PC00034)	SNAP receptor activity (GO:0005484); SNARE binding (GO:0000149)	organelle localization (GO:0051640); vesicle fusion (GO:0006906); intracellular protein transport (GO:0006886)	integral component of membrane (GO:0016021); SNARE complex (GO:0031201); endomembrane system (GO:0012505); vacuole (GO:0005773)	
TAB2	TGF-BETA-ACTIVATED KINASE 1 AND MAP3K7-BINDING PROTEIN 2 (PTHR46253:SF2)		K63-linked polyubiquitin modification-dependent protein binding (GO:0070530)	positive regulation of protein kinase activity (GO:0045860); protein phosphorylation (GO:0006468)		p38 MAPK pathway->TAK1 binding protein 2

(cont. on the next page)

Appendix B (cont.)

TAP2	ANTIGEN PEPTIDE TRANSPORTER 2 (PTHR24221:SF237)	ATP-binding cassette (ABC) transporter (PC00003)	ATP hydrolysis activity (GO:0016887); active transmembrane transporter activity (GO:0022804)	transmembrane transport (GO:0055085)	integral component of membrane (GO:0016021)	
THADA	THYROID ADENOMA-ASSOCIATED PROTEIN (PTHR14387:SF7)	scaffold/adaptor protein (PC00226)		tRNA methylation (GO:0030488)	cytosol (GO:0005829)	
THEMIS	PROTEIN THEMIS (PTHR15215:SF1)			immune response (GO:0006955); T cell receptor signaling pathway (GO:0050852)		
TINAG	TUBULOINTERSTITIAL NEPHRITIS ANTIGEN (PTHR12411:SF274)	cysteine protease (PC00081)	cysteine-type endopeptidase activity (GO:0004197)	proteolysis involved in cellular protein catabolic process (GO:0051603)	extracellular space (GO:0005615); membrane (GO:0016020); lysosome (GO:0005764)	
TMBIM1	PROTEIN LIFEGUARD 3 (PTHR23291:SF35)	ion channel (PC00133)		apoptotic signaling pathway (GO:0097190); negative regulation of apoptotic signaling pathway (GO:2001234)	Golgi apparatus (GO:0005794); endoplasmic reticulum (GO:0005783); vacuole (GO:0005773); plasma membrane (GO:0005886)	
TMEM258	TRANSMEMBRANE PROTEIN 258 (PTHR13636:SF0)				membrane (GO:0016020); membrane-bounded organelle (GO:0043227)	
TNF	TUMOR NECROSIS FACTOR (PTHR11471:SF23)	intercellular signal molecule (PC00207)	tumor necrosis factor receptor superfamily binding (GO:0032813); signaling receptor activity (GO:0038023); cytokine activity (GO:0005125)	tumor necrosis factor-mediated signaling pathway (GO:0033209); transcription by RNA polymerase II (GO:0006366); extrinsic apoptotic signaling pathway via death domain receptors (GO:0008625); positive regulation of transcription by RNA polymerase II (GO:0045944); positive regulation of NF-kappaB transcription factor activity (GO:0051092)	extracellular space (GO:0005615); cell surface (GO:0009986)	Wnt signaling pathway->NFAT Target Genes; Apoptosis signaling pathway->Tumor necrosis factor

(cont. on the next page)

Appendix B (cont.)

TNFSF15	TUMOR NECROSIS FACTOR LIGAND SUPERFAMILY MEMBER 15 (PTHR11471:SF24)	intercellular signal molecule (PC00207)				
TNFSF18	TUMOR NECROSIS FACTOR LIGAND SUPERFAMILY MEMBER 18 (PTHR15267:SF1)	intercellular signal molecule (PC00207)				
TPPP2	TUBULIN POLYMERIZATION-PROMOTING PROTEIN FAMILY MEMBER 2 (PTHR12932:SF21)	non-motor microtubule binding protein (PC00166)	tubulin binding (GO:0015631)	microtubule bundle formation (GO:0001578); positive regulation of protein polymerization (GO:0032273); microtubule polymerization (GO:0046785)	microtubule (GO:0005874)	
TRA	T CELL RECEPTOR ALPHA CHAIN MC.7.G5-RELATED (PTHR19433:SF72)	immunoglobulin receptor superfamily (PC00124)		response to bacterium (GO:0009617)		
TRAV12-2	T CELL RECEPTOR ALPHA VARIABLE 12-2 (PTHR19343:SF3)	immunoglobulin receptor superfamily (PC00124)	peptide binding (GO:0042277)			
TUBD1	TUBULIN DELTA CHAIN (PTHR11588:SF4)	tubulin (PC00228)	GTP binding (GO:0005525); structural constituent of cytoskeleton (GO:0005200)	mitotic nuclear division (GO:0140014); microtubule cytoskeleton organization (GO:0000226)	cytoplasm (GO:0005737); microtubule (GO:0005874)	

(cont. on the next page)

Appendix B (cont.)

TYK2	NON-RECEPTOR TYROSINE-PROTEIN KINASE TYK2 (PTHR45807:SF6)	non-receptor tyrosine protein kinase (PC00168)	non-membrane spanning protein tyrosine kinase activity (GO:0004715); cytokine receptor binding (GO:0005126)	peptidyl-tyrosine phosphorylation (GO:0018108); cytokine-mediated signaling pathway (GO:0019221); hemopoiesis (GO:0030097); cell differentiation (GO:0030154); intracellular signal transduction (GO:0035556); receptor signaling pathway via JAK-STAT (GO:0007259)	cytosol (GO:0005829)	Inflammation mediated by chemokine and cytokine signaling pathway->Janus kinase
UBQLN4	UBIQUILIN-4 (PTHR10677:SF21)	scaffold/adaptor protein (PC00226)	polyubiquitin modification-dependent protein binding (GO:0031593)	ubiquitin-dependent protein catabolic process (GO:0006511)	cytosol (GO:0005829)	
UMOD	UROMODULIN (PTHR14002:SF40)	transmembrane signal receptor (PC00197)		neutrophil migration (GO:1990266)	extracellular space (GO:0005615); apical plasma membrane (GO:0016324)	
ZBTB38	ZINC FINGER AND BTB DOMAIN-CONTAINING PROTEIN 38 (PTHR24388:SF71)	C2H2 zinc finger transcription factor (PC00248)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978); DNA-binding transcription factor activity, RNA polymerase II-specific (GO:0000981)	transcription, DNA-templated (GO:0006351); regulation of transcription, DNA-templated (GO:0006355)		
ZGPAT	ZINC FINGER CCCH-TYPE WITH G PATCH DOMAIN-CONTAINING PROTEIN (PTHR46297:SF1)	RNA polymerase II cis-regulatory region sequence-specific DNA binding (GO:0000978)	transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)	nucleus (GO:0005634); membrane (GO:0016020)		

(cont. on the next page)

Appendix B (cont.)

ZMIZ1	ZINC FINGER MIZ DOMAIN-CONTAINING PROTEIN 1 (PTHR10782:SF7)	ubiquitin-protein ligase (PC00234)	nuclear receptor coactivator activity (GO:0030374); ubiquitin protein ligase activity (GO:0061659)	protein sumoylation (GO:0016925); transcription by RNA polymerase II (GO:0006366); regulation of transcription by RNA polymerase II (GO:0006357)		
ZNF365	PROTEIN ZNF365 (PTHR15739:SF2)			nucleic acid phosphodiester bond hydrolysis (GO:0090305); neuron projection development (GO:0031175); DNA biosynthetic process (GO:0071897); telomere maintenance (GO:0000723); regulation of double-strand break repair via homologous recombination (GO:0010569); regulation of neuron projection development (GO:0010975)	centriole (GO:0005814); centriolar satellite (GO:0034451)	

APPENDIX C. EXTENDED HAPLOTYPE HOMOZGOSITY PLOTS FOR 196 GENES FROM 14 POPULATIONS

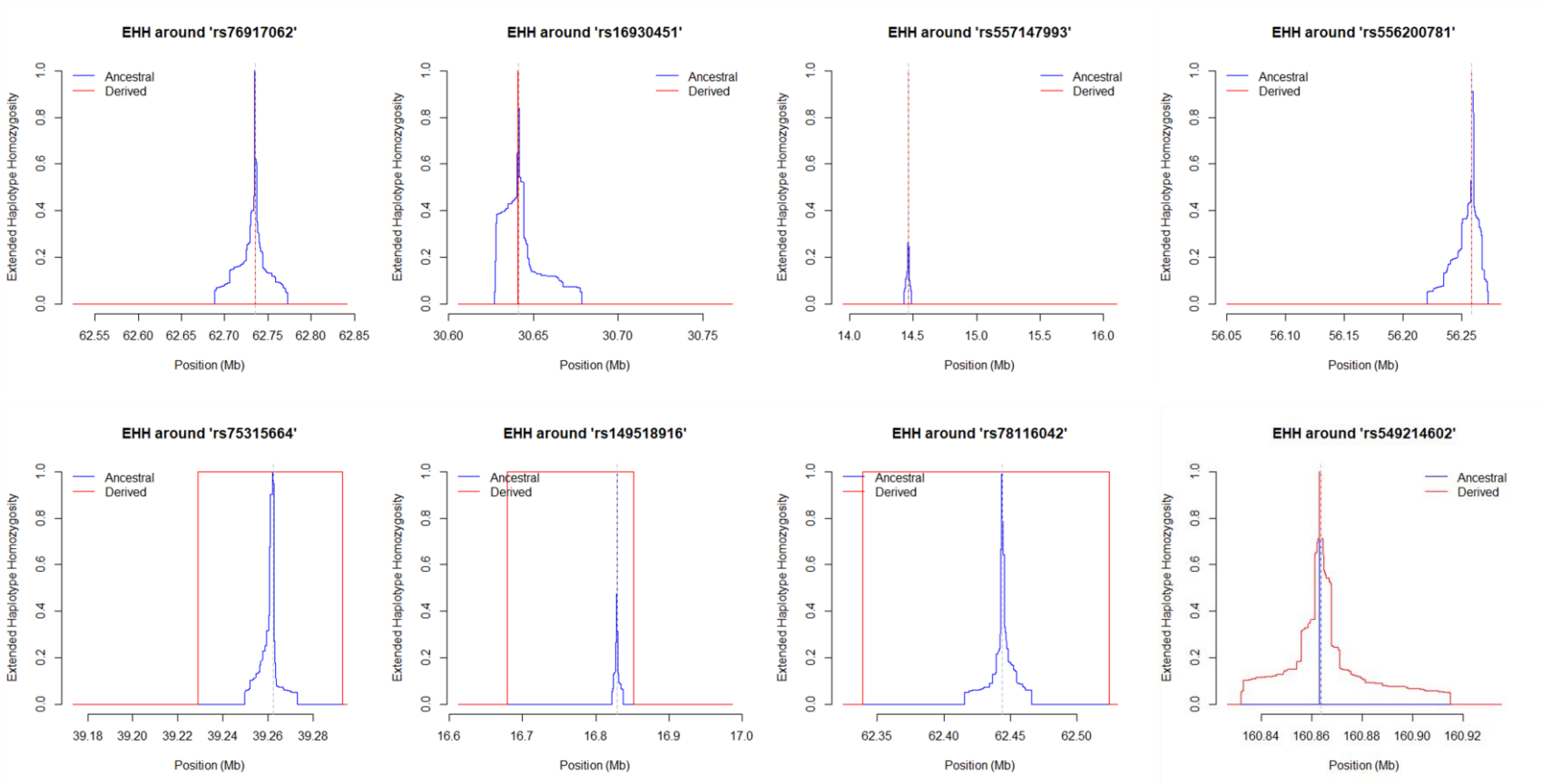

Figure C.1 Extended Haplotype Homozygosity plots for 196 genes of the CEU subpopulation of the European population.

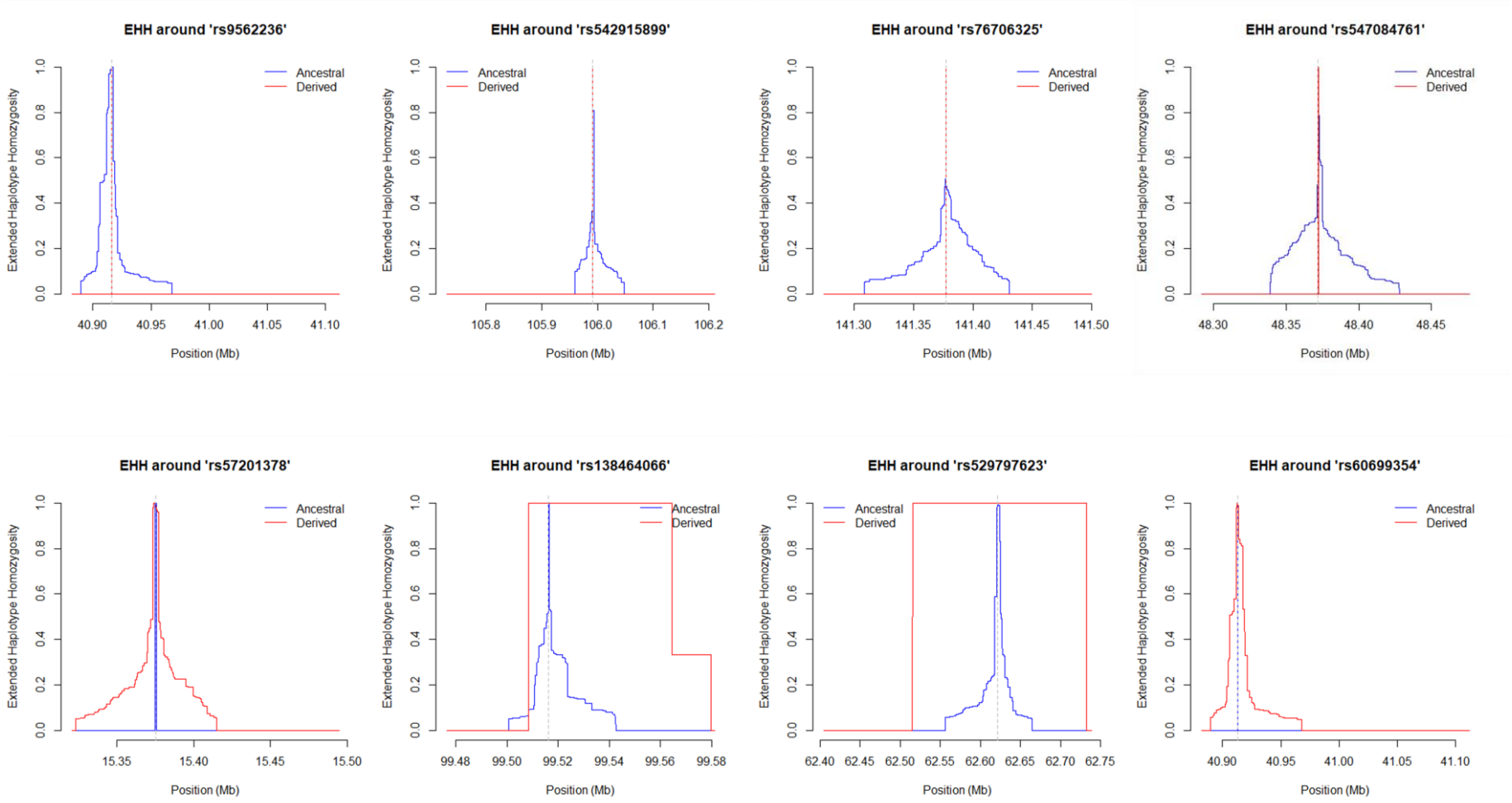

Figure C.2 Extended Haplotype Homozygosity plots for 196 genes of the FIN subpopulation of the European population.

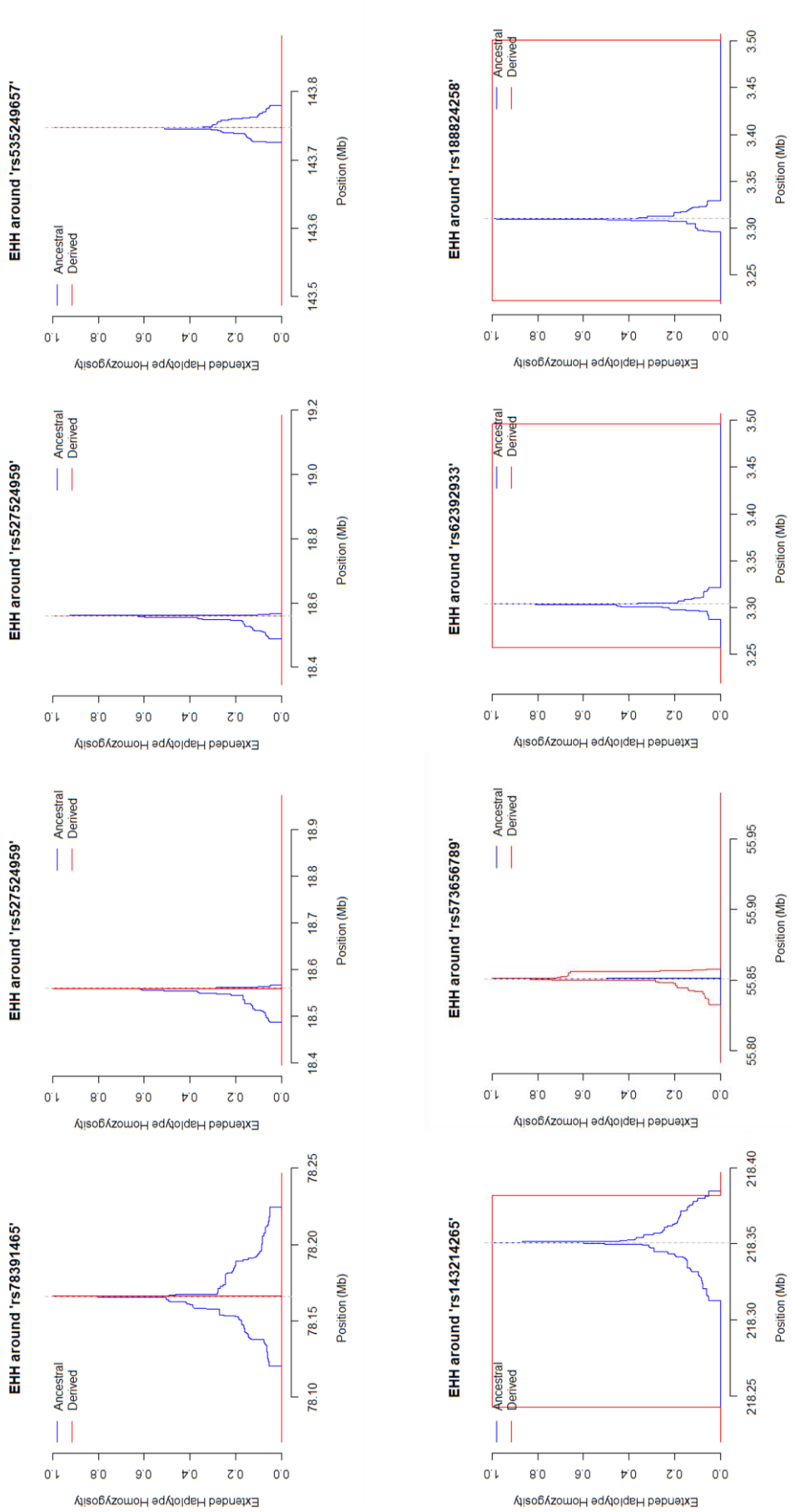

Figure C.3 Extended Haplotype Homozygosity plots for 196 genes of the GBR subpopulation of the European population.

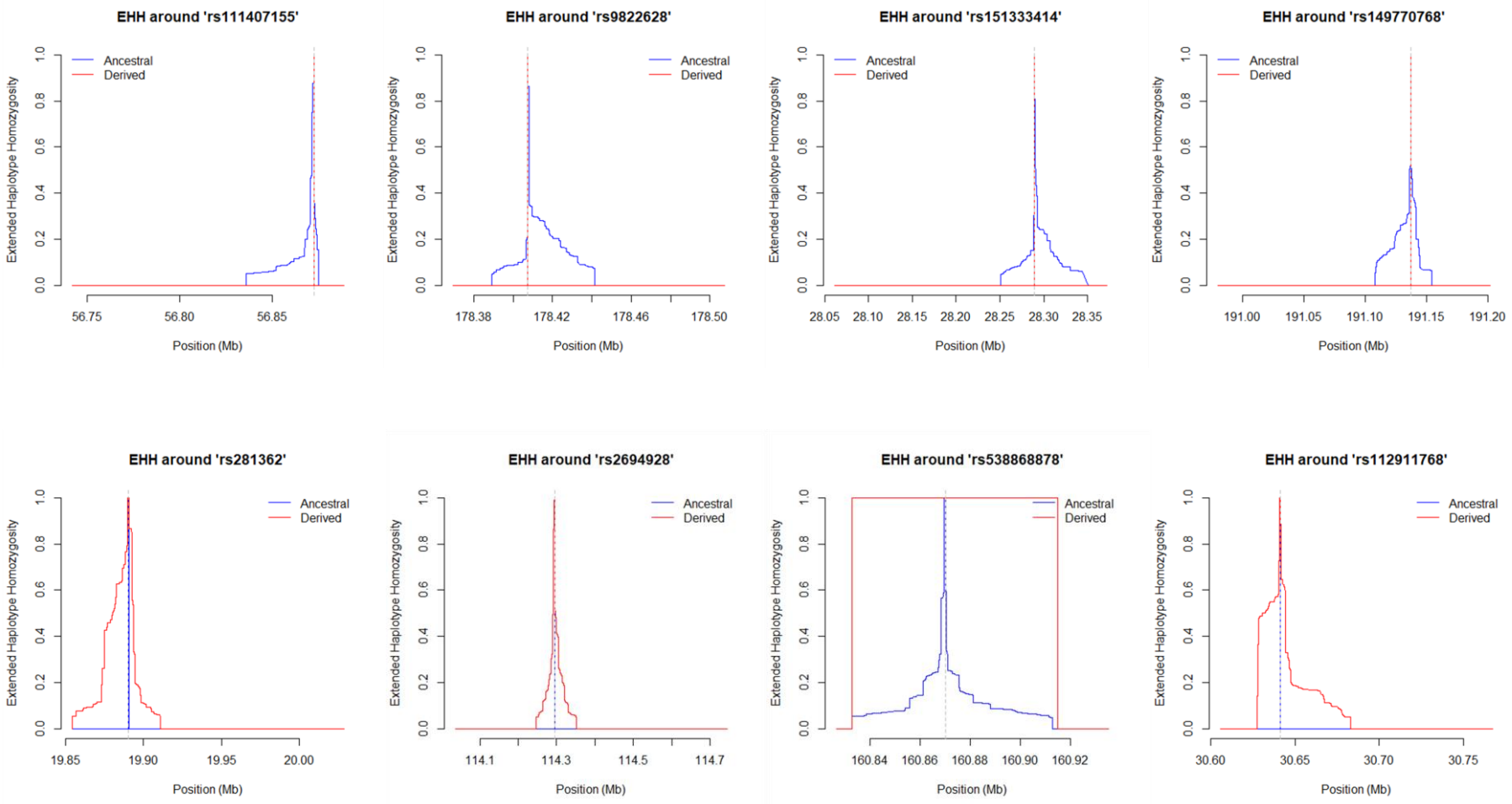

Figure C.4 Extended Haplotype Homozygosity plots for 196 genes of the IBS subpopulation of the European population.

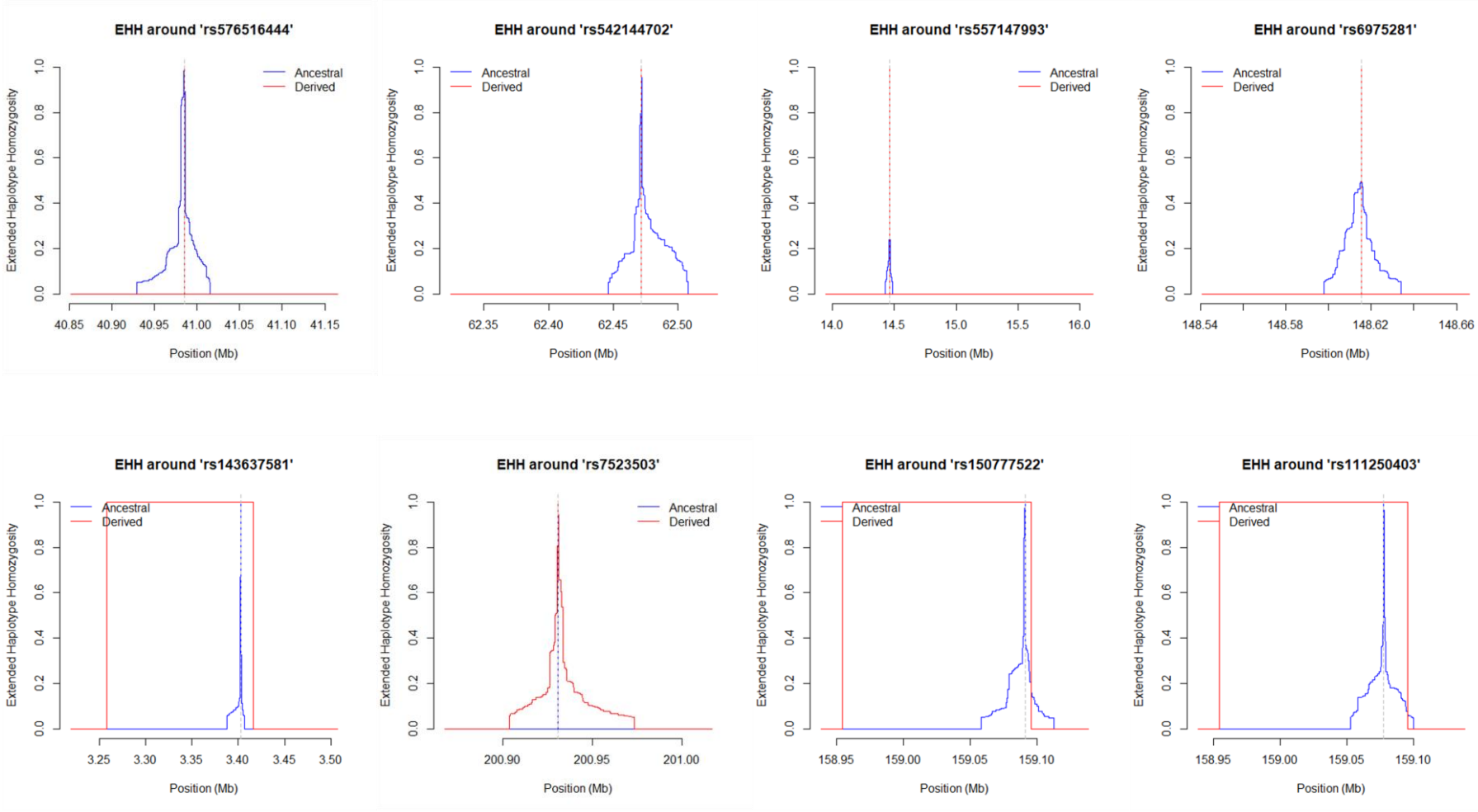

Figure C.5 Extended Haplotype Homozygosity plots for 196 genes of the TSI subpopulation of the European population.

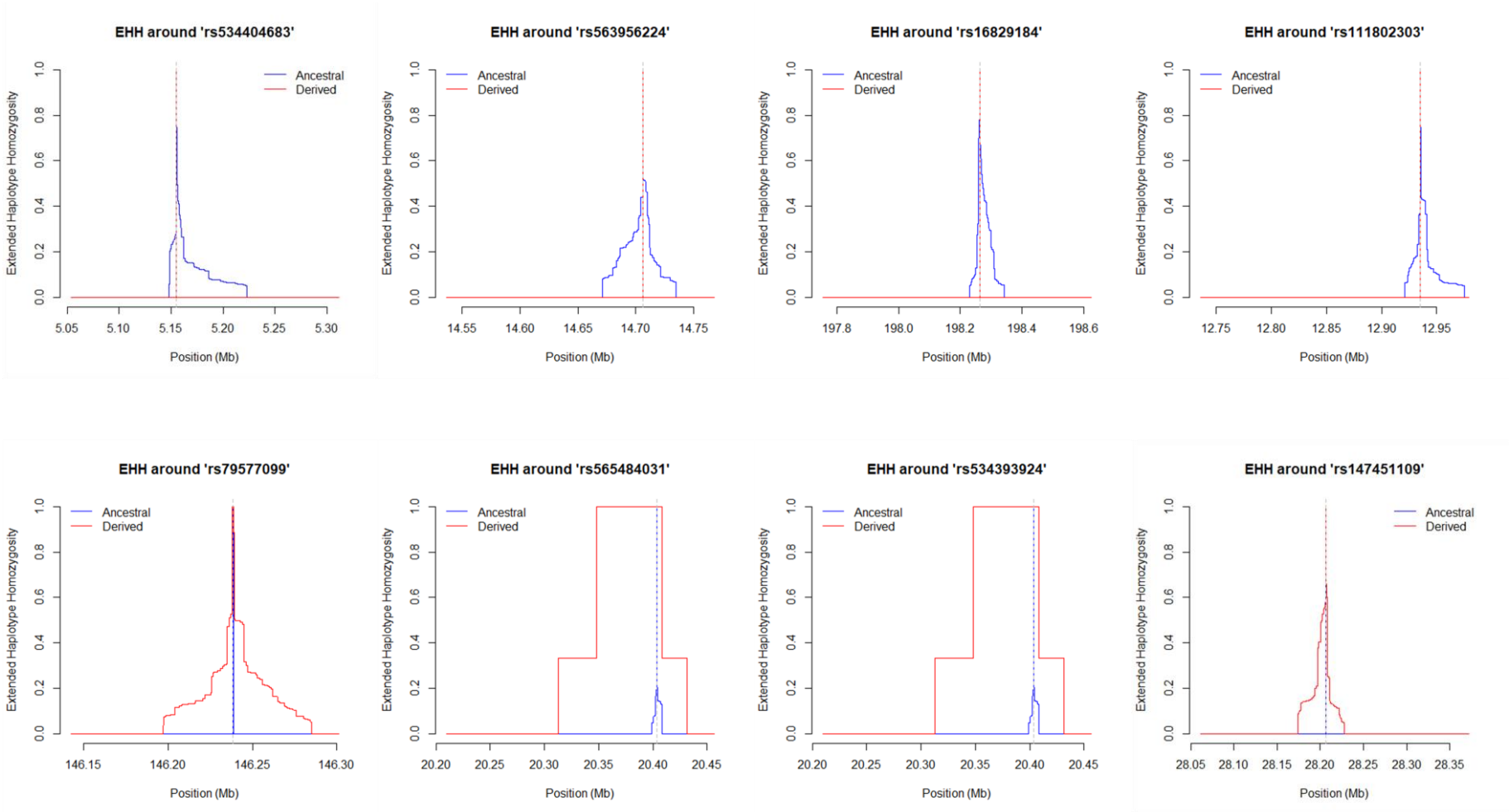

Figure C.6 Extended Haplotype Homozygosity plots for 196 genes of the CDX subpopulation of the East Asian population.

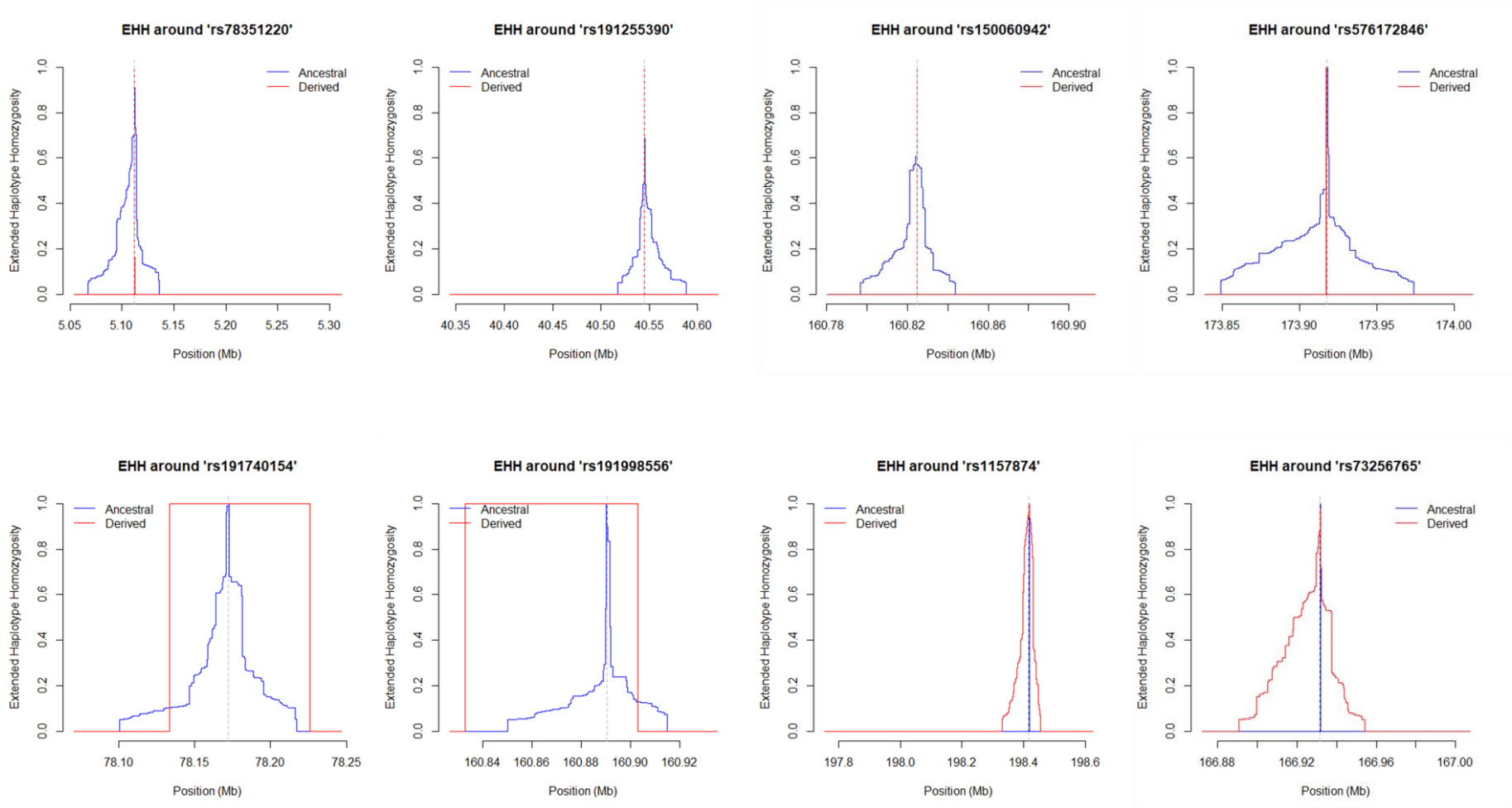

Figure C.7 Extended Haplotype Homozygosity plots for 196 genes of the CHB subpopulation of the East Asian population.

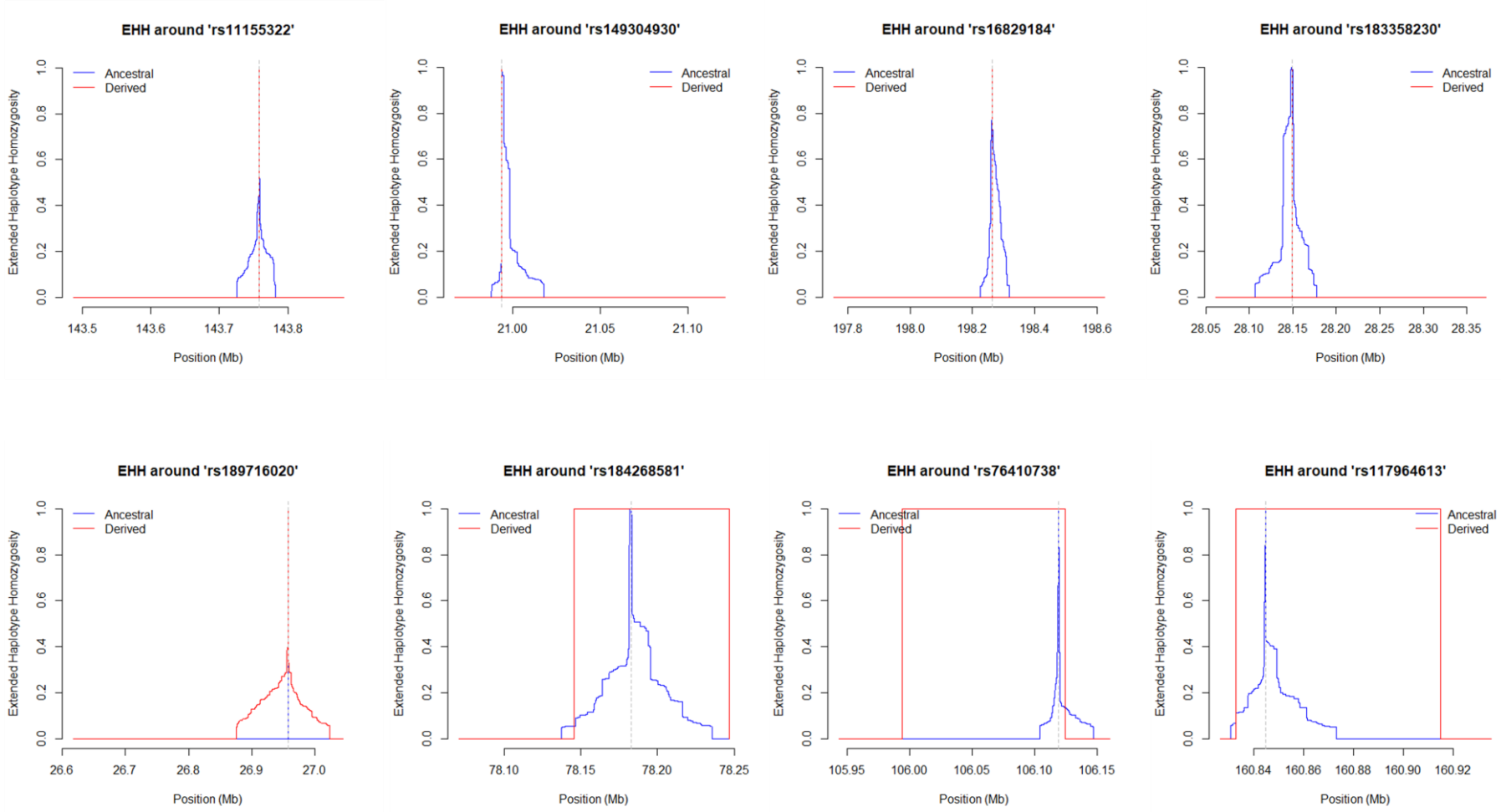

Figure C.8 Extended Haplotype Homozygosity plots for 196 genes of the CHS subpopulation of the East Asian population.

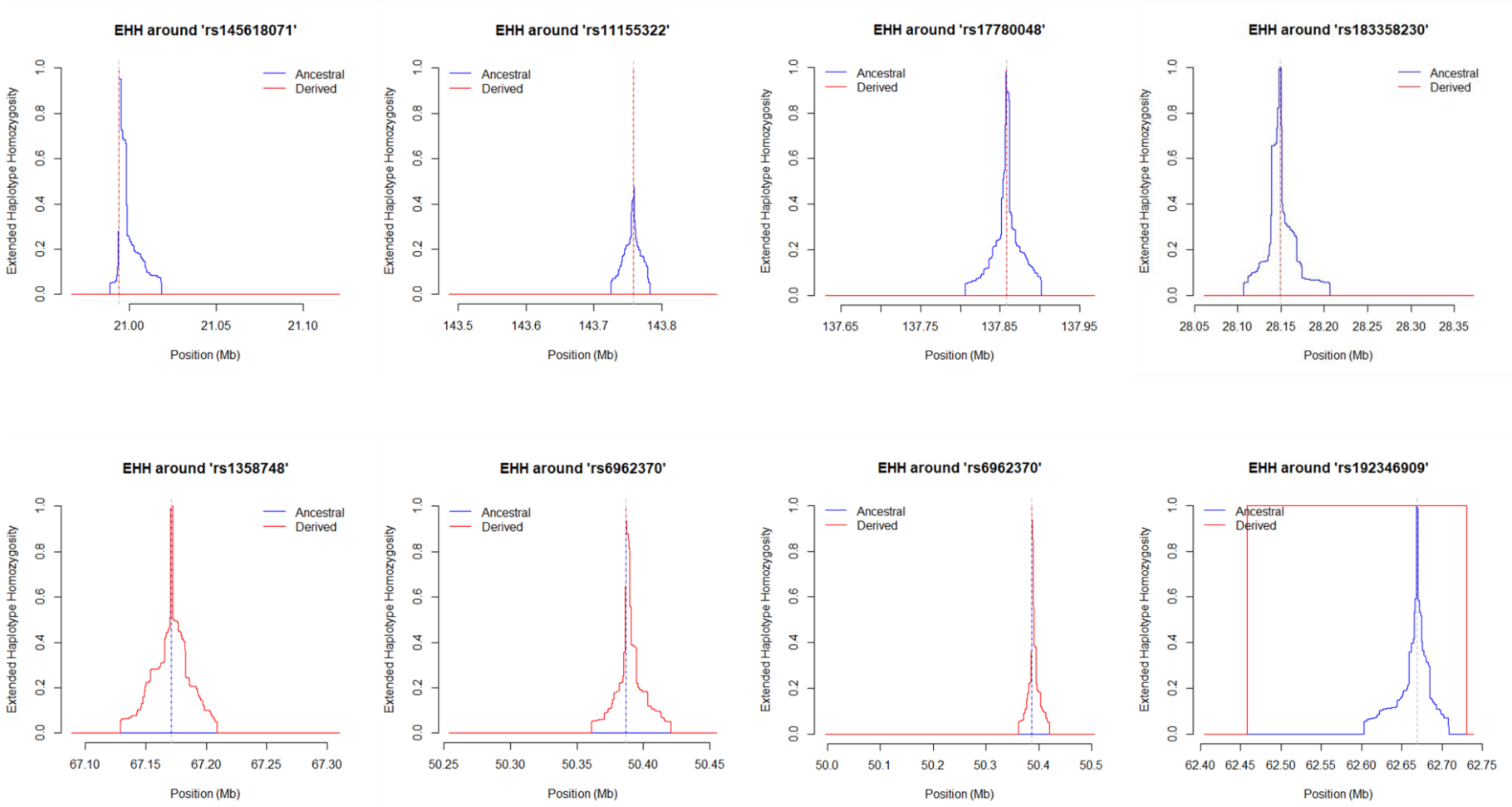

Figure C.9 Extended Haplotype Homozygosity plots for 196 genes of the JPT subpopulation of the East Asian population.

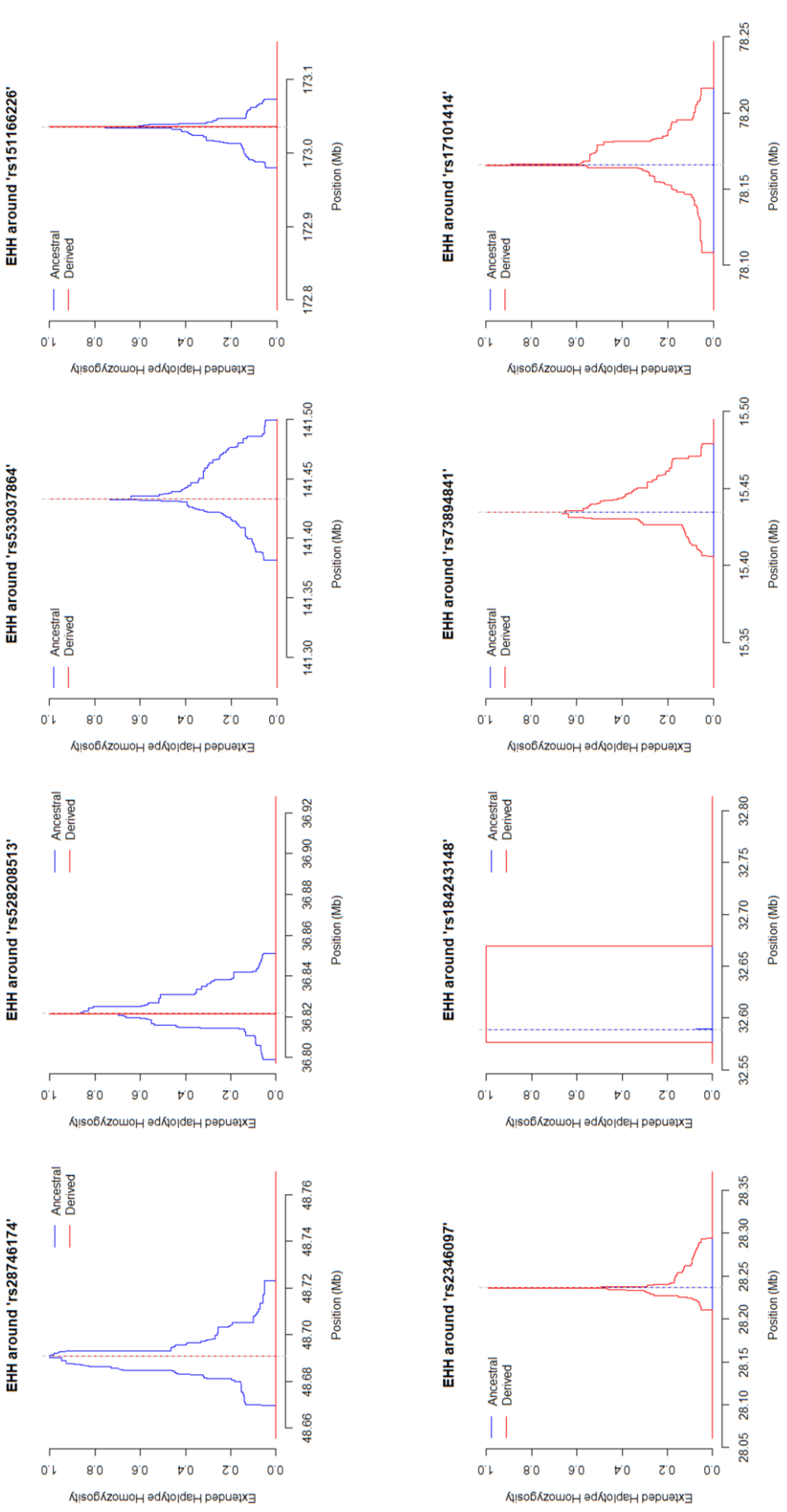

Figure C.10 Extended Haplotype Homozygosity plots for 196 genes of the KHV subpopulation of the East Asian population.

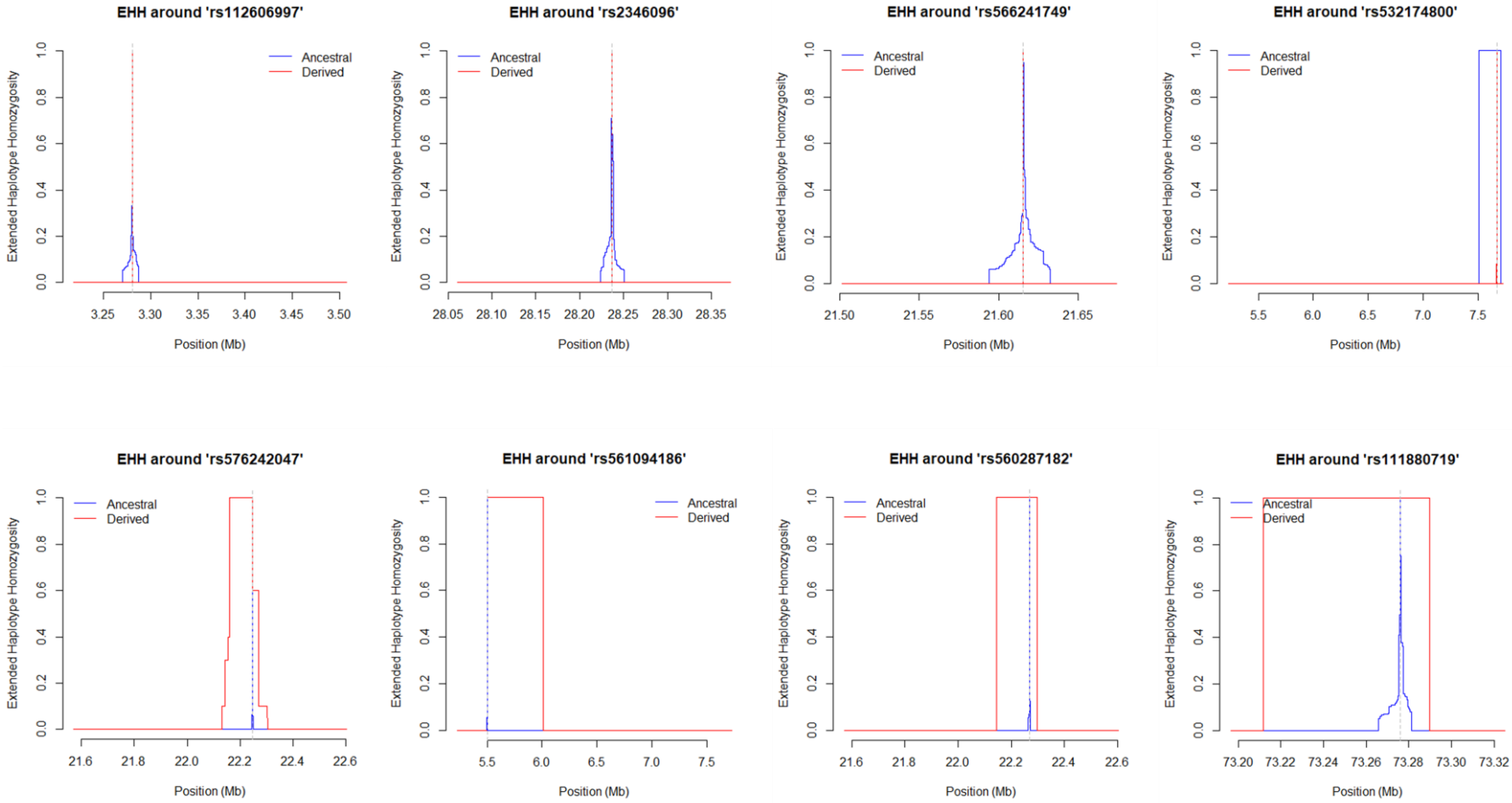

Figure C.11 Extended Haplotype Homozygosity plots for 196 genes of the GWD subpopulation of the African population.

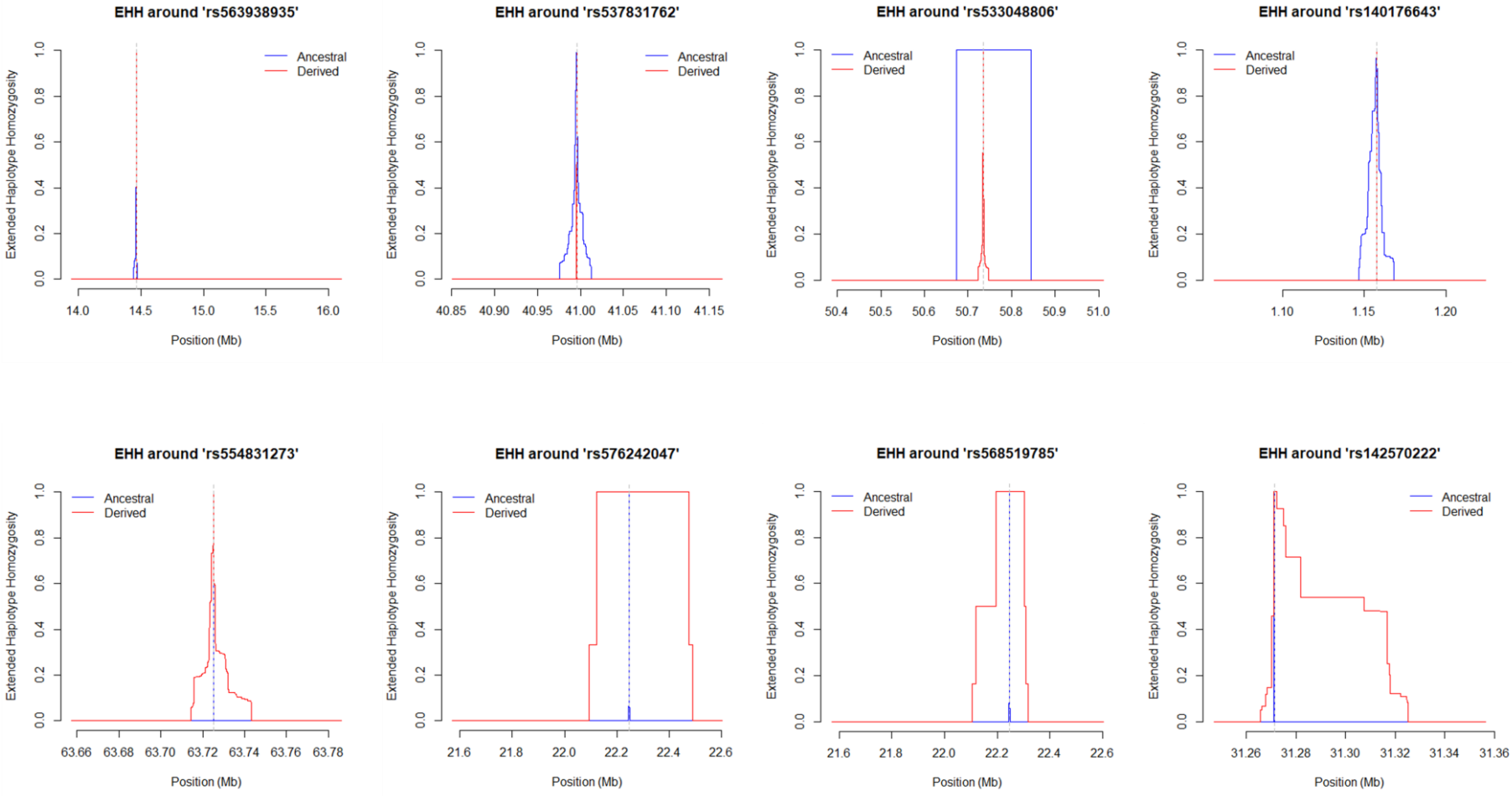

Figure C.12 Extended Haplotype Homozygosity plots for 196 genes of the YRI subpopulation of the African population.

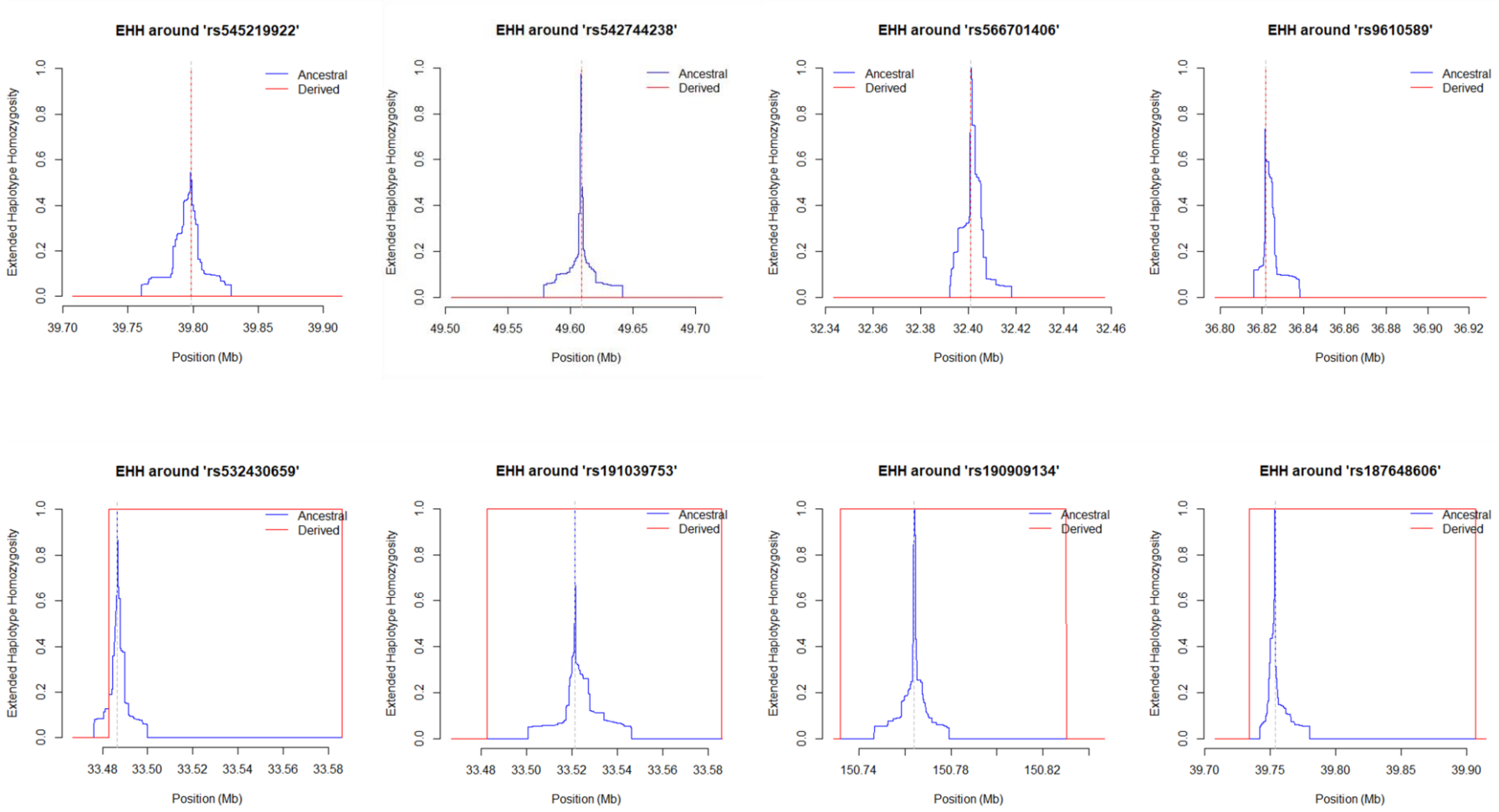

Figure C.13 Extended Haplotype Homozygosity plots for 196 genes of the ESN subpopulation of the African population.

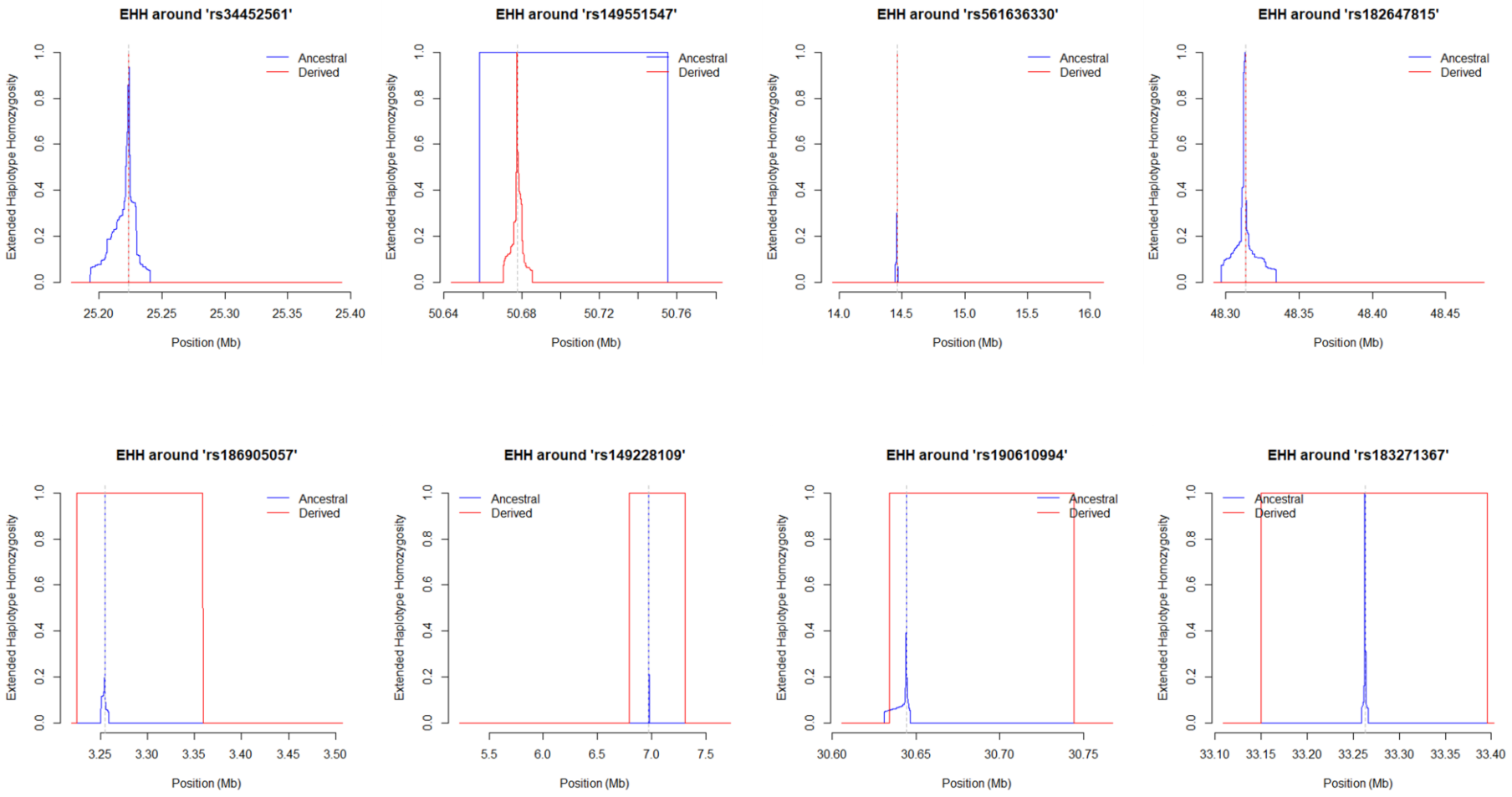

Figure C.14 Extended Haplotype Homozygosity plots for 196 genes of the LWK subpopulation of the African population.