

The Validity and Statistical Power of the Case-Only Study Design for Interaction Analysis: Gene-Gene Interaction and the Role of Genotype Imputation in Gene-Environment Interaction

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Milda Aleknonytė-Resch

Kiel, 2021

Erste Gutachterin: Prof. Dr. Astrid Dempfle
Zweiter Gutachter: Prof. Dr. Hinrich Schulenburg

Tag der mündlichen Prüfung: 25.02.2021

Summary

In search of the origin of complex human diseases such as inflammatory bowel disease (IBD) and Parkinson disease (PD), not only are genetic and environmental factors thought to play a role, but gene-gene (G×G) and gene-environment (G×E) interactions may also contribute to the disease etiology. However, examining these interactions is challenging, as high statistical power is needed in order to detect them, especially when the effects are small and single nucleotide polymorphisms (SNPs) with low minor allele frequencies (MAFs) are examined. In epidemiological studies, the traditional case-control (CC) design is often employed, however, it often does not achieve the necessary statistical power for interaction analysis. The case-only (CO) study design proves to be of great use in these circumstances, as it not only obviates the need for controls, but given the same number of cases, it is statistically more powerful than the CC study design. However, two key assumptions must be fulfilled in order for the CO study design to be valid: (i) the disease of interest must be sufficiently rare, and (ii) the two risk factors (gene and environment in case of G×E interaction, both genetic in G×G interaction) must be independent in the general population. Nevertheless, the practical implementation of the CO study design in the context of G×G interaction analysis remained unexplored.

Another aspect that increases the statistical power to detect interaction effects is the number of observations available for the analysis. Thus, combined data from the largest consortia comprising of numerous centers and thousands of cases gives the highest possible chance of detecting interactions to date. Depending on the center, a different genotyping chip is often used resulting in different genotyped SNPs. Genotype imputation uses a reference database to impute missing data and thus allows to gain information on numerous SNPs and make the analysis of data from different centers possible. It is a standard procedure in genome-wide-association-studies which analyse genetic main effects (MEs). The reference base used for genotype imputation is population based and assumed to consist of healthy individuals, therefore, their linkage disequilibrium (LD) structure may differ from diseased cases, particularly in areas with MEs. Thus, whether genotype imputation has an impact on the validity and statistical power of statistical tests for G×E interactions in CO studies would be a useful asset in the analysis, yet was unknown.

This thesis examined two aspects of interaction analysis in the CO study design. First, whether imputing data from a reference base consisting of healthy individuals into diseased cases has consequences for the downstream G×E interaction analysis. The results showed, that imputation does not work well in areas with MEs and low minor allele frequencies of SNPs. The lower the LD to neighbouring SNPs was, the more the MAF resembled the reference base controls than the cases from the used dataset. This imputation bias further led to a loss of statistical power in the G×E interaction analysis.

The second aspect of this thesis is the practical implementation of G×G interaction analysis in which SNPs were considered as proxies for genes. The (ii) assumption of independence of both factors is problematic in G×G interactions due to LD. Moreover, computational issues arise due to the large number of possible genome-wide interaction pairs that, given more than one

center, need to be calculated separately for each. Thus, a method was proposed that practically implements G×G interaction analysis. The method includes, among others aspects, analysing SNPs on different chromosomes or chromosome arms to fulfil the (ii) assumption and focusing on SNPs with known MEs in order to reduce the computational burden. The largest available datasets for IBD and PD to date were used for the analysis of G×G interactions for these complex diseases. While the G×G interaction analysis for IBD found G×G interactions to be scarce, it yielded 10 unique significant G×G interactions for PD after multiple test correction.

The findings of this thesis will add to an improved understanding of G×E and G×G interaction analysis in the CO study design. It points out areas of caution when examining G×E interaction using imputed data. Furthermore, this work shows how G×G interaction can be implemented in a statistically sound and computationally efficient manner. This could lead to further G×G interaction analyses, opening doors to more in-depth knowledge on the etiology of complex human diseases.

Zusammenfassung

Auf der Suche nach dem Ursprung komplexer menschlicher Krankheiten wie z.B. entzündlichen Darmerkrankung (engl. Inflammatory bowel disease, IBD) und Parkinson (engl. Parkinson disease, PD) spielen nicht nur genetische und umweltbedingte Risikofaktoren eine Rolle, sondern auch die daraus entstehenden Gen-Gen- (G×G) und die Gen-Umwelt- (engl. gene-environment, G×E) -Interaktionen können ebenfalls zur Ätiologie der Krankheiten beitragen. Die Untersuchung dieser Interaktionen ist jedoch schwierig, da eine hohe statistische Aussagekraft erforderlich ist, um sie nachzuweisen, insbesondere wenn die Effekte gering sind und auf ein einziges Nukleotid reduzierten Bereiche (engl. single nucleotide polymorphisms, SNPs) mit niedrigen Häufigkeiten des selteneren Allels (engl. minor allele frequency, MAFs) untersucht werden. In epidemiologischen Studien wird häufig das traditionelle Case-Control-Design (CC) verwendet, dies erreicht jedoch häufig nicht die erforderliche statistische Aussagekraft für die Interaktionsanalyse. Das Case-Only-Studiendesign (CO) erweist sich unter diesen Umständen als sehr nützlich, da es nicht nur die Notwendigkeit von Kontrollen überflüssig macht, sondern bei gleicher Anzahl von Fällen statistisch leistungsfähiger ist als das CC-Studiendesign. Zwei Schlüsselannahmen müssen jedoch erfüllt sein, damit das CO-Studiendesign gültig ist: (i) Die zu untersuchende Krankheit muss ausreichend selten sein, und (ii) die beiden Risikofaktoren (Gen- und Umweltfaktor bei G×E-Interaktionen, beide genetische Faktoren in der G×G-Interaktion) müssen in der Allgemeinbevölkerung unabhängig sein. Die praktische Umsetzung des CO-Studiendesigns im Rahmen der G×G-Interaktionsanalyse blieb bis jetzt jedoch unerforscht.

Ein weiterer Aspekt, der die statistische Aussagekraft zur Erkennung von Interaktionseffekten erhöht, ist die Anzahl der für die Analyse verfügbaren Beobachtungen. Somit bieten Datensätze aus den größten Konsortien, die aus zahlreichen Zentren und Tausenden von Fällen bestehen, bis heute die höchstmögliche Chance, Interaktionen zu erkennen. Je nach Zentrum wird häufig ein anderer Genotypisierungschip verwendet, was zu unterschiedlichen genotypisierten SNPs führt. Die Genotyp-Imputation verwendet eine Referenzdatenbank, um fehlende Daten zu unterstellen, und ermöglicht so die Information über zahlreiche SNPs und die Analyse von Daten aus verschiedenen Zentren. Es ist ein Standardverfahren in genomweiten Assoziationsstudien, die genetische Haupteffekte (engl. main effects, MEs) analysieren. Die Referenzbasis für die Genotyp-Imputation basiert auf der Bevölkerung und besteht vermutlich aus gesunden Personen. Daher kann sich die Struktur des Kopplungsungleichgewichts (engl. linkage disequilibrium, LD) in der Referenzbasis von den erkrankten Fällen unterscheiden, insbesondere in Regionen mit MEs. Ob die Genotyp-Imputation einen Einfluss auf die Validität und statistische Aussagekraft statistischer Tests für G×E-Interaktionen in CO-Studien hat, war daher bis jetzt unbekannt.

Die Vorliegende Dissertation untersuchte zwei Aspekte der Interaktionsanalyse im CO-Studiendesign. Erstens, ob die Zuordnung von Daten aus einer Referenzbasis, die aus gesunden Personen besteht, zu erkrankten Fällen Konsequenzen für die nachfolgenden G×E-Interaktionsanalyse hat. Die Ergebnisse zeigten, dass die Imputation in Regionen mit MEs und niedrigen MAFs von SNPs nicht gut funktioniert. Je niedriger das LD gegenüber benachbarten

SNPs war, desto mehr ähnelte der MAF den Referenzbasis-Kontrollen als den Fällen aus dem verwendeten Datensatz. Diese Verzerrung, aufgrund der Imputation führte ferner zu einem Verlust der statistischen Aussagekraft in der G×E-Interaktionsanalyse.

Der zweite Aspekt, der in dieser Dissertation behandelt worden ist, ist die praktische Implementierung der G×G-Interaktionsanalyse, bei der SNPs als Proxys für Gene betrachtet wurden. Die (ii) Annahme des CO-Studiendesigns der Unabhängigkeit beider Faktoren voraussetzt ist bei G×G-Interaktionen aufgrund von LD problematisch. Darüber hinaus ergeben sich Rechenprobleme aufgrund der großen Anzahl möglicher genomweiter Interaktionspaare, die bei mehr als einem Zentrum für jedes separat berechnet werden müssen. Daher wurde in dieser Dissertation ein Verfahren vorgeschlagen, dass die G×G-Interaktionsanalyse umsetzbar implementiert. Das Verfahren umfasst unter anderem die Analyse von SNPs auf verschiedenen Chromosomen oder Chromosomenarmen, um die (ii) Annahme zu erfüllen, und die Konzentration auf SNPs mit bekannten MEs, um den Rechenaufwand zu verringern. Die bislang größten verfügbaren Datensätze für IBD und PD wurden zur Analyse der G×G-Interaktionen für diese komplexen Krankheiten verwendet. Während die G×G-Interaktionsanalyse für IBD ergab, dass G×G-Interaktionen für diese Krankheit wahrscheinlich selten sind, bei PD ergab sie nach multipler Testkorrektur 10 einzigartige signifikante G×G-Interaktionen.

Die Ergebnisse dieser Arbeit werden zu einem besseren Verständnis der G×E- und G×G-Interaktionsanalyse im CO-Studiendesign beitragen. Bei der Untersuchung der G×E-Interaktionen mit imputierten Daten ist bei seltenen SNPs Vorsicht geboten. Darüber hinaus zeigt diese Arbeit, wie die G×G-Interaktion statistisch fundiert und rechnerisch effizient implementiert werden kann. Dies könnte zu weiteren G×G-Interaktionsanalysen führen und Türen zu tieferem Wissen über die Ätiologie komplexer menschlicher Krankheiten öffnen.

Table of Contents

Summary	i
Zusammenfassung.....	iii
List of Publications.....	vii
List of Tables and Figures	viii
Abbreviations	xi
Preface.....	xii
1 Introduction.....	1
1.1 Motivation	1
1.2 Complex Diseases of Interest: Inflammatory Bowel Disease and Parkinson Disease.	2
1.3 Case Only Study Design	4
1.4 Gene-Environment Interactions	7
1.5 Gene-Gene Interactions	9
1.6 Genotype Imputation	11
2 Results.....	14
2.1 The effect of genotype imputation on the validity and power of statistical tests for gene-environment interactions in case-only studies	14
2.1.1 Summary	14
2.1.2 Publication.....	16
2.2 Case-Only Analysis of Gene-Gene Interactions in Inflammatory Bowel Disease.....	43
2.2.1 Summary	43
2.2.2 Publication.....	46
2.3 Case-Only Analysis of Gene-Gene Interactions in Parkinson Disease.....	98
2.3.1 Summary	98
2.3.2 Publication.....	100
3 Discussion	131
3.1 Imputation	131
3.2 Rare SNPs.....	132
3.3 An Approach for Investigating Gene-Gene Interactions	133
3.4 Strengths and Limitations.....	135
3.5 Conclusion and Outlook.....	136

Bibliography.....	139
Acknowledgements.....	148
Declaration	149
Curriculum Vitae.....	150

List of Publications

Published work by the author incorporated into this thesis

Aleknonytė-Resch, M., Freitag-Wolf, S., International Inflammatory Bowel Disease Genetics Consortium, Schreiber, S., Krawczak, M., & Dempfle, A. (2020). Case-only analysis of gene–gene interactions in inflammatory bowel disease. *Scandinavian journal of gastroenterology*, 1-10.

Unpublished work ready for submission to a journal by the author incorporated into this thesis

Aleknonytė-Resch, M., Szymczak, S. Freitag-Wolf, S., Krawczak, M., & Dempfle, A. (to be submitted). The effect of genotype imputation on the validity and power of statistical tests for gene-environment interactions in case-only studies.

Aleknonytė-Resch, M., et al. (to be submitted). Case-only analysis of gene–gene interactions in Parkinson’s disease.

List of Tables and Figures

1. Introduction

Figure 1	Power of the case only and case control designs of G×G analysis	6
----------	---	---

2.1. The effect of genotype imputation on the validity and power of statistical tests for gene-environment interactions in case-only studies

Figure 1	MAF of target SNPs	22
Figure 2	Kappa value of target SNPs	23
Figure 3	Estimated imputation accuracy score against kappa	24
Figure 4	Boxplot of statistical power of the G×E interaction by LD threshold	25
Figure 5	G×E interaction beta coefficient estimation bias	26

Appendix 1

Table	SNPs used in analysis	29
-------	-----------------------------	----

Appendix 2

Table	Exposure probabilities	32
-------	------------------------------	----

Appendix 3

Figure 1	Medium-low MAF group with ME	37
Figure 2	Medium-high MAF group with ME	37
Figure 3	High MAF group with ME	38
Figure 4	Medium-low MAF group without ME	38
Figure 5	Medium-high MAF group without ME	39
Figure 6	High MAF group without ME	39

Appendix 4

Figure 1	Boxplot of statistical power of G×E interaction by simulated SNP	40
Figure 2	Mean statistical power of G×E interaction from imputed data	40

Appendix 5

Figure 1	Boxplot of type I error rate of G×E interaction by simulated SNP	41
Figure 2	Boxplot of type I error rate of G×E interaction by LD threshold	41
Figure 3	Mean type I error rate of G×E interaction	42
Figure 4	G×E interaction beta estimation bias	42

2.2. Case-Only Analysis of Gene-Gene Interactions in Inflammatory Bowel Disease

Summary

Figure 1	Independence assumption validation	44
----------	--	----

Publication

Figure 1	Power of the CO and CC designs of G×G analysis	52
Table 1	Pairs of IBD-associated SNPs meeting different criteria for genotypic non-independence	50
Table 2	Top G×G interactions between SNPs associated with CD	53
Table 3	Top G×G interactions between SNPs associated with UC	53
Table 4	G×G interaction on CD risk of SNPs rs26528 and rs9297145	53

Supplementary Figures

Figure 1a	Distribution of interaction odds ratios observed among candidate SNP pairs (grey bars) in CD	57
Figure 1b	Distribution of interaction odds ratios observed among candidate SNP pairs (grey bars) in UC	58
Figure 2	Forest plot of CO interaction odds ratio for SNPs rs26528 and rs9297145 in CD	59
Figure 3	Forest plot of CC interaction odds ratio for SNPs rs26528 and rs9297145 in CD	60

Supplementary Tables

Table 1a	Top G×G interactions between SNPs associated with CD (CO design)	61
Table 1b	Top G×G interactions between SNPs associated with UC (CO design)	65
Table 1c	Top G×G interactions between SNPs associated with IBD (CO design)	69
Table 2	G×G interactions between SNPs 30 kb upstream or downstream from <i>IL27</i> and <i>KPNA7</i> gene regions	73
Table 3a	G×G interactions between SNPs associated with CD on the same chromosome arm in the CC design	80
Table 3b	G×G interactions between SNPs associated with UC on the same chromosome arm in the CC design	89

2.3. Case-Only Analysis of Gene-Gene Interactions in Parkinson Disease

Figure 1	Power of the CO and CC designs of G×G analysis for PD	106
Figure 2	Locus zoom plots of chromosome 12	107
Figure 3	Interaction between rs76904798 in <i>AC079630.4</i> on chromosome 12 and rs1007709 near <i>STY10</i> on chromosome 12	108
Figure 4	Interaction between rs76904798 in <i>AC079630.4</i> on chromosome 12 and rs187879258 in <i>BICD1</i> on chromosome 12	108
Figure 5	Interaction between rs76904798 in <i>AC079630.4</i> on chromosome 12 and rs117835488 in <i>FGD4</i> on chromosome 12	109
Figure 6	Interaction between rs76904798 in <i>AC079630.4</i> on chromosome 12 and rs139007869 in <i>PKP2</i> on chromosome 12	109
Figure 7	Interaction between rs76904798 in <i>AC079630.4</i> on chromosome 12 and rs150084348 on chromosome 12	110

Figure 8	Interaction between rs76904798 in <i>AC079630.4</i> on chromosome 12 and rs141945110 on chromosome 12	110
Figure 9	Interaction between rs34637584 in <i>LRRK2</i> on chromosome 12 and rs151094822 in <i>FGD4</i> on chromosome 12	111
Figure 10	Interaction between rs34637584 in <i>LRRK2</i> on chromosome 12 and rs80291054 in <i>PKP2</i> on chromosome 12	111
Figure 11	Interaction between rs112485576 on chromosome 6 and rs7856915 in <i>KANK1</i> on chromosome 9	112
Figure 12	Interaction between rs26431 in <i>PAM</i> on chromosome 5 and rs139186308 in <i>TTC6</i> on chromosome 14	112
Supplement 1		
Table	Number of cases per center	115
Supplement 2		
Table	Significant SNP×SNP interactions	116
Supplement 3		
Table	Sensitivity analysis	129

Abbreviations

CC	Case-Control
CD	Crohn's Disease
CI	Confidence Interval
CO	Case-Only
D	Disease Status
E	Environment
FEV	Forced Expiratory Volume
FVC	Forced Vital Capacity
G	Gene
G×E	Gene-Environment
G×G	Gene-Gene
GWAS	Genome-Wide-Association Studies
HRC	Haplotype Reference Consortium
IBD	Inflammatory Bowel Disease
IIBDGC	International Inflammatory Bowel Disease Genetics Consortium
IPDGC	International Parkinson Disease Genomics Consortium
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
ME	Main Effect
MHC	Major Histocompatibility Complex
OR	Odds Ratio
PD	Parkinson's Disease
PKU	Phenylketonuria
SNP	Single Nucleotide Polymorphisms
UC	Ulcerative Colitis

Preface

This thesis focuses on the statistical power and practical implementation of gene-gene (G×G) and gene-environment (G×E) interactions in case-only (CO) studies. Specifically, these two issues are addressed in more detail:

1. The effect of genotype imputation on the validity and power of statistical tests for G×E interactions in CO studies.
2. The practical implementation of G×G interaction analysis in CO studies.

In Chapter 1, the key concepts and issues related to this thesis are introduced. The first chapter begins with the motivation for this thesis (subchapter 1.1). It is followed by introducing complex diseases and going into detail about the two complex diseases of interest in this thesis, namely inflammatory bowel disease (IBD) and Parkinson's disease (PD) (subchapter 1.2). Furthermore, the concept and modelling of the CO study design is examined and the advantages and disadvantages of this design are mentioned (subchapter 1.3). Then moving on to introduce G×E (subchapter 1.4) and G×G (subchapter 1.5) interactions along with the challenges and issues associated with the analysis of these interactions. Finally, genotype imputation is introduced in subchapter 1.6 with its benefits and drawbacks that are relevant for this thesis.

The core of this thesis consists of three publications presented in Chapter 2. In each subchapter, a brief summary of each publication is followed by the publication itself. For already published work, the bibliography used in that publication is found directly after the publication itself. For not yet published work, the bibliography is merged with the overall citations in this thesis and can be commonly found after Chapter 3. A short summary and bibliographic details of the three publications, which are hereafter referred to as “publication (i)”, “publication (ii)” and “publication (iii)” are presented below:

- (i) Aleknytyė-Resch, M., Szymczak, S. Freitag-Wolf, S., Krawczak, M., & Dempfle, A. (to be submitted).

The effect of genotype imputation on the validity and power of statistical tests for gene-environment interactions in case-only studies.

Abstract

The case-only (CO) design is a powerful approach to study gene-environment (G×E) interactions using only affected subjects. Genotype imputation uses a reference sample such as the Haplotype Reference Consortium (HRC) to predict genotypes at untyped loci. However, using healthy controls as a reference in a CO study may introduce systematic error, especially in regions of genetic main effects. Using data from 719 Crohn's Disease (CD) cases from Kiel, Germany, we investigated the imputation accuracy for target SNPs with varying minor allele frequencies (MAFs) with and without genetic main effects (MEs). Target SNPs were imputed using neighbouring proxy SNPs with different levels of linkage disequilibrium (LD) and the HRC as a reference base. True genotypes of target SNPs were available for comparison.

Furthermore, we simulated different levels of G×E interaction by assigning environmental exposure conditional on SNP genotype in order to evaluate the loss in statistical power. The comparison of true and imputed MAFs of target SNPs showed that the highest differences between true and imputed MAFs were of SNPs found in gene regions such as IL27 and NOD2, which are known to play a role in CD. Some target SNPs with low MAF (≤ 0.05) and MEs exhibited a high imputation accuracy score, yet the agreement between true and imputed genotypes was low. SNPs of interest with lower MAF achieve less statistical power and a gradual decrease in statistical power can be observed as the level of LD decreases. In conclusion, our study describes constellations in which imputed data should be used with caution when testing for G×E interactions in CO studies and exemplifies how G×E interactions can remain undetected due to statistical power loss resulting from imputation of cases when using a reference base consisting of controls.

- (ii) Aleknytytė-Resch, M., Freitag-Wolf, S., International Inflammatory Bowel Disease Genetics Consortium, Schreiber, S., Krawczak, M., & Dempfle, A. (2020)

Case-only analysis of gene–gene interactions in inflammatory bowel disease.

Scandinavian journal of gastroenterology, 1-10.

Abstract

Background

Gene–gene interactions ($G \times G$) potentially play a role in the etiology of complex human diseases, including inflammatory bowel disease (IBD), and may partially explain their ‘missing heritability’.

Methods

Using the largest genotype dataset available for IBD (16,636 Crohn’s disease (CD) and 12,888 ulcerative colitis (UC) cases) we analysed $G \times G$ with the powerful case-only (CO) design. We studied 169 single nucleotide polymorphisms (SNPs) for CD (156 for UC), previously shown to be associated with the respective diseases. To ensure the validity of the CO design, we confined our analysis to pairs of unlinked SNPs. We used principal component analysis at the center level to adjust for possible causes of genotypic association other than $G \times G$, such as population stratification and genotyping batch effects. Results from center-wise logistic regression analyses were combined by a random effects meta-analysis.

Results

A number of nominally significant ($p < .05$) $G \times G$ interactions were observed, but none of these withstood the Bonferroni multiple testing correction. However, one SNP pair, comprising rs26528 in the *IL27* gene and rs9297145 in the *KPNA7* gene region was characterized by an interaction odds ratio of 1.18 (95% CI: 1.10–1.27) for CD and a p-value of 7.75×10^{-6} . Owing to the concurrent role of the *IL27* and *KPNA7* genes in NF- κ B signaling, a master regulator of pro- and anti-inflammatory processes in IBD, the observed interaction also has biological plausibility.

Conclusions

We were able to exemplify the utility of the CO design for analyzing $G \times G$, but had to recognize that such interactions are probably scarce for IBD.

(iii) **Aleknonytė-Resch, M.**, et al. (to be submitted).

Case-only analysis of gene–gene interactions in Parkinson disease.

Abstract

Introduction

Gene-gene interactions ($G \times G$) potentially play a role in the etiology of complex human diseases, including Parkinson disease (PD), and may contribute to the explanation of their “missing heritability”. It is important to distinguish between biological and statistical $G \times G$. The former denotes that the gene products in question share a common role in the etiology of the disease while the latter is interpreted as effect modification.

Methods

Using one of the largest available genotype datasets for PD (36 362 cases) and considering early-onset, not-early-onset PD as well as all data combined, we analysed $G \times G$ with the powerful case-only (CO) design and used single nucleotide polymorphisms (SNPs) as proxies for genes. We confined one of the SNPs in the interaction pair to one of the 90 SNPs previously shown to be associated with PD. To ensure the validity of the CO design, we only examined pairs of unlinked SNPs with a hard-call rate of at least 0.8. We used principal component analysis at the center level to adjust for possible causes of genotypic association other than $G \times G$, such as population stratification and genotyping batch effects. Results from center-wise logistic regression analyses were combined by a random effects meta-analysis.

Results

The genome-wide significance level was set to 5.56×10^{-10} after multiple test correction. Our study found 337 significant SNP×SNP interactions, ten of which could be identified as unique $G \times G$. The pair with the highest density of significant SNP pairs included rs76904798 with a main effect, in the AC079630.4 antisense gene region, overlapping with LRRK2 and rs1007709 in the promoter region of SYT10 with an interaction OR of 1.80 (95% CI: 1.65-1.95) and a p-value of 2.67×10^{-48} .

Conclusions

The rs76904798 nearby LRRK2 and rs1007709 near SYT10 pair has biological plausibility due to LRRK2 direct link to PD and involvement in neural plasticity as well as SYT10 contribution to the exocytosis of secretory vesicles in neurons. This pair, along with the nine other unique significant $G \times G$ lay ground for further specific research of these pairs in their combined role in PD.

Finally, in Chapter 3, the overall primarily common findings of this thesis are discussed. The main common issues, namely how imputation can affect further analysis (subchapter 3.1) and the importance as well as difficulties of examining rare SNPs (subchapter 3.2) are elaborated. An approach for investigating $G \times G$ interactions in practice that I described and implemented

in publications (ii) and (iii) is discussed in subchapter 3.3. The strengths and limitations of this thesis are elaborated in subchapter 3.4 followed by final conclusions and outlook for further research in the areas of G×G and G×E interactions in the CO study design (subchapter 3.5). The three main chapters are followed by a bibliography of references used in this thesis, acknowledgements, an official declaration that this thesis is the outcome of my own work and effort as well as my current curriculum vitae at the time of submission of this thesis.

Chapter 1

1 Introduction

1.1 Motivation

The main motivation for this thesis stems from the interest in the case only (CO) study design and focus on the missing heritability of complex human diseases. The CO study design has many advantages in comparison the standard case-control (CC) approach, yet is not as widely used. This may be partially due to the lack of analysis and methodological guidelines in certain tricky scenarios when applying the CO study design. Therefore, I took a closer look at the effect of genotype imputation on the validity and power of statistical tests for gene-environment (G×E) interactions in CO studies and how the CO study design could be used to detect gene-gene (G×G) interactions.

The other interest focus is complex human diseases. In general, human diseases are categorized as being “simple” or “complex”. Simple diseases, also known as monogenic or Mendelian disorders, follow the concept of dominant and recessive traits, where the disease is controlled by single genes (Antonarakis and Beckmann 2006). Such Mendelian disorders include, among others, sickle-cell anaemia, haemophilia and cystic fibrosis. On the other hand, complex diseases do not follow simple Mendelian inheritance patterns. The genetic architecture of complex diseases is not yet fully understood. Complex diseases include cancer, cardiovascular diseases such as myocardial infarction, chronic inflammatory disorders such as inflammatory bowel disease and neurodegenerative disorders such as Alzheimer’s disease, or Parkinson disease to name a few. Up to date, complex diseases are believed to result from G×G and G×E interactions, genetic heterogeneity and potentially even more reasons, which are yet unknown (Manolio et al. 2009).

Genome-wide-association studies (GWAS), measure and analyse variations in DNA sequence across the whole human genome and compare the variations between healthy controls and diseased cases. The main goal of GWAS is to identify genetic risk factors for diseases (Bush and Moore 2012). GWAS comprise of over a million single nucleotide polymorphisms (SNPs), gathered from thousands of individuals have already provided useful insights and have helped to identify causal links between genes or gene regions and diseases (Hardy and Singleton 2009; Buniello et al. 2019). However, many complex traits still have a large portion of unexplained heritability. For example, through GWAS and twin studies, only 60-70% of heritability is explained for ulcerative colitis (UC) (G.-B. Chen et al. 2014). One explanation for missing heritability includes the idea that there is a much larger number of variants with smaller effect sizes that are yet to be found (Manolio et al. 2009). Another explanation states that there could be rare variants (with, possibly, larger effects), which fall through quality

control and are poorly detected by current genotyping arrays that focus on variants present in at least 5% of the population (Maroille and Tarailo-Graovac 2019). Both issues are not only relevant for main effect analysis, but also play a role when examining G×G and G×E interactions, because a higher number of cases is needed to have enough statistical power to find small effects and rare variants could also have large interaction effects. This underlines the motivation to apply the CO study design, as it has a higher statistical power than the CC study design, given the same number of cases (W. J. Gauderman 2002; W. James Gauderman 2002) (see Chapter 1.3).

The larger the dataset, the higher the statistical power to find smaller effects. Thus, I was eager to work with consortiums, which could provide access to the largest available datasets gathered on specific complex diseases. Since my diseases of interest were inflammatory bowel disease (IBD) and Parkinson disease (PD), datasets from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) and the International Parkinson Disease Genomics Consortium (IPDGC) were ideal to work with as they both had genetic data on over 17 000 patients of the respective diseases. This gave me the highest possible statistical power to date to be able to find G×G interactions and contribute to the understanding of these complex diseases.

1.2 Complex Diseases of Interest: Inflammatory Bowel Disease and Parkinson Disease

IBD is a term used to describe diseases that involve the chronic inflammation of the digestive tract. Such diseases are Crohn's disease (CD) and ulcerative colitis (UC). Clinically speaking, CD could involve any part of the gastrointestinal tract while UC is limited to rectal and colonic mucosal layers. Whereas the bowel inflammation in CD is transmural, discontinuous, and may involve granulomas as well as intestinal or perianal fistulas, in UC the inflammation is continuous and neither fistulas nor granulomas seem to occur (Panaccione 2013). In approximately 10% of IBD cases, no definitive classification as either CD or UC can be made (Liu et al. 2015). The age of onset is roughly the same for both diseases and usually occurs in the second to fourth decade of life (Molodecky et al. 2012). Complications due to these diseases develop in half of the patients, often resulting in surgery (Thia et al. 2010).

CD and UC respectively affect up to 0.3% and 0.5% of the European population (Ng et al. 2017) with rising prevalence, indicating IBD to be an emerging global disease. The highest prevalence rates of CD and UC are found in Europe (322 per 100 000 for CD in Italy, 505 per 100 000 for UC in Norway) and Canada (319 per 100 000 for CD and 248 per 100 000 for UC) (Alatab et al. 2020). Studies show, that there are areas of high and low incidence and prevalence, with greater numbers in developed rather than developing countries and urban areas rather than rural areas (Torres et al. 2017). Interestingly, incident and prevalence rates increase in correlation with the urbanization of developing areas (Ng et al. 2013). Moreover, there is evidence of immigrant studies that suggest the increase of incidence of IBD in the first and second generations of Asian migrants that have moved to western countries. These incident rates sometimes exceeded those of the local population (Pinsk et al. 2007; Tsironi et al. 2004).

These findings support the theory that single factors alone are not causal for the diseases and that there must be some sort of interaction of multiple factors.

Genetically, a considerable overlap has been found for CD and UC alike in terms of their genetic risk loci. Liu et al mentioned a total of 169 SNPs (with a MAF ≥ 0.05) disease associated loci, 38 of which were newly identified and 27 of those were associated with both diseases (Liu et al. 2015). A number of studies with smaller case numbers (less than 3000 individuals each, except for the study by Zhang et al. (J. Zhang et al. 2019)) have shown interactions between genes or gene regions in CD and UC. These include, but are not limited to findings of interactions between SNPs in the *IL12B* and *STAT4* (Glas, Seiderer, et al. 2012), *FCRL3* and *MHC* (Martínez et al. 2007) genes for general IBD, between *JAK2* and *STAT3* (Polgar et al. 2012), *ATG16L1* and *PTPN2* (Glas, Wagner, et al. 2012), *CARD8* and *NAPL3* (Roberts et al. 2010) for CD, and between *ATG16L1* and *PTPN2* (Glas, Wagner, et al. 2012) and *JAK2* and *STAT3* (Polgar et al. 2012) for UC. Zhang et al. (J. Zhang et al. 2019) undertook an extensive search for G×G interactions using the data from the IIBDGC and, using a screening method and CC approach, found nine weak interactions in the *MHC* region on chromosome 6. Other studies have found evidence of G×E interactions in IBD, for example Yadav et al. identified 64 SNP – smoking interactions and validated the findings in a mouse model (Yadav et al. 2017). Other environmental exposures that have been associated with risk for IBD include medications and infections (Rogler and Vavricka 2015). The diet is also thought to be a factor, primarily the “Western” diet, high in fat and protein, low in fruits and vegetables could have an effect on IBD (Amre et al. 2007). Thus, there is confirmation of the belief that IBD is a result of the interplay of genetic and environmental factors including gut microbiota.

PD is also considered a complex disease and is the most common movement disorder and the second most common neurodegenerative disease after Alzheimer’s disease (Kalia and Lang 2015). First noticed by James Parkinson in 1817 and named “shaking palsy” (Parkinson 2002), PD leads to loss of motor skills and starts with shaking, stiffness and progresses to difficulty walking, balance and coordination issues. Mental, behavioural, talking difficulties, insomnia, depression, memory difficulties and fatigue are possible symptoms of PD (Reich and Savitt 2019). PD is a highly age-related disease, while there are cases of early-onset PD, it is rare before the fifth decade (de Lau and Breteler 2006) and usually the onset is at the age of 65-70. Naturally, the prevalence rates increase with age. In the developed countries, the prevalence rate of PD is estimated to be 0.3%, roughly 1-2% in the population older than 60 (Nussbaum and Ellis 2003; Capriotti and Terzakis 2016). Being a neurodegenerative disease with no present cure, progressing PD patients require an increasing amount of assistance, eventually leading to around-the-clock care (S. L. Wong, Gilmour, and Ramage-Morin 2014).

While the exact cause of PD is still unknown, there are two forms – monogenic and sporadic – and there is evidence that genetic variability plays a role in both. Around 1-5% of all PD cases are considered to be monogenic and several genes appear to be causal (Singleton and Hardy 2016). However, some of the genes that cause monogenic PD, such as *SNCA*, *LRRK2* and *VPS13C*, among others, seem to play a role in the sporadic disease as well (Trinh et al. 2018). A recent study by Nalls et al. identified 90 genome-wide significant variants across 78 genomic regions that explained 16-36% of the heritable risk of PD (depending on the prevalence) (Nalls et al. 2019). Moreover, Blauwendraat et al. have found two genome-wide significant

associations between PD and age at onset: one in the *SNCA* gene and another in *TMEM175* (Blauwendraat et al. 2019). Interestingly, the male-female ratio in PD increases with age and men are roughly 1.5 times more susceptible to PD (Moisan et al. 2016). Environmental factors such as exposure to some metals, pesticides, herbicides, and fungicides have all been linked to PD (Bjorklund et al. 2018; Tanner et al. 2011; Liou et al. 1997). G×G and G×E in the CO design are yet to be thoroughly systematically analysed. Even though the population is aging in the developed countries, there may be an increase in the prevalence of PD that cannot be explained by demographic changes of the population alone.

1.3 Case Only Study Design

Before discussing the CO study design, it is useful to elaborate CC studies. CC studies are a type of retrospective observational studies that are popular in medical and epidemiological research. They are designed to help determine the association between an exposure and an outcome. As the name already describes, such studies require cases known to have the outcome of interest and controls that do not have the desired outcome. Simply put, the association is then determined by comparing the frequencies of the exposure between the cases and controls (van Stralen et al. 2010). Data gathered in CC studies can easily be used to analyse multiplicative interactions between variables. An example for exploiting G×G interaction in the standard CC design would be by means of a logistic regression in a dominant genetic model:

$$\text{logit}\{P(D = 1)\} = \theta_0 + \theta_1 G_1 + \theta_2 G_2 + \theta_3 G_1 G_2 \quad (1)$$

In the dominant genetic model, genotypes (G) were encoded assuming a dominant effect of the minor allele, i.e. $G=1$ for homozygous or heterozygous carriers of the minor allele, $G=0$ for homozygous carriers of the major allele. In case of an additive genetic model, the predictor variable would be encoded with 0, 1 or 2 depending on the number of minor alleles and the response variable would follow the dominant genetic model. In the example equation (1), D , the response variable, is the disease status whereas genotypes G_1 and G_2 are treated as predictor variables, alongside an interaction $G_1 G_2$. θ_1 and θ_2 represent the main effects of G_1 and G_2 , respectively, while θ_3 represents the interaction effect in this model.

CC studies have numerous advantages in comparison to other types of observational studies including their simplicity and ability to generate a great amount of information from relatively few subjects in a short amount of time, especially when uncommon diseases are analysed (Mann 2003). The two prime possible drawbacks of CC studies are recall bias and control group selection (Schulz and Grimes 2002). Due to the fact that CC studies are retrospective, the gathering of predictor variables largely depends on the perception of the probands, investigators or both. This may make gathering information about previous environmental factors biased if differences arise between the case and control groups, thus introducing recall bias to the data. Especially in genetic association studies, control group selection can be a sensitive topic. Not only do the controls have to be matched to phenotypic and demographical aspects of the case group, but genetics must also be considered in order not to cause

population stratification issues. Solutions such as using family-based instead of unrelated control groups have been proposed. However, they bring their own drawbacks such as the difficulty to accumulate a sufficient sample size (Evangelou et al. 2006).

The CO design could be considered as a “spin-off” of the traditional CC design, when the goal of the study is to analyse interaction effects of possible risk factors. In this study design, only the group of cases known to have the outcome of interest is needed. Continuing the example of G×G interaction, in the CO design, the genotype of the first SNP (G_1) is treated as a predictor variable in the logistic regression, with respective regression coefficient δ_1 representing the interaction effect, whereas the other genotype (G_2) is treated as the response variable, i.e.

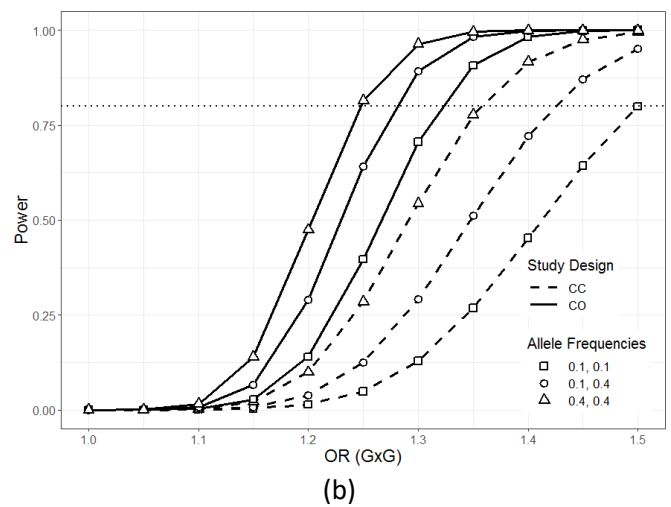
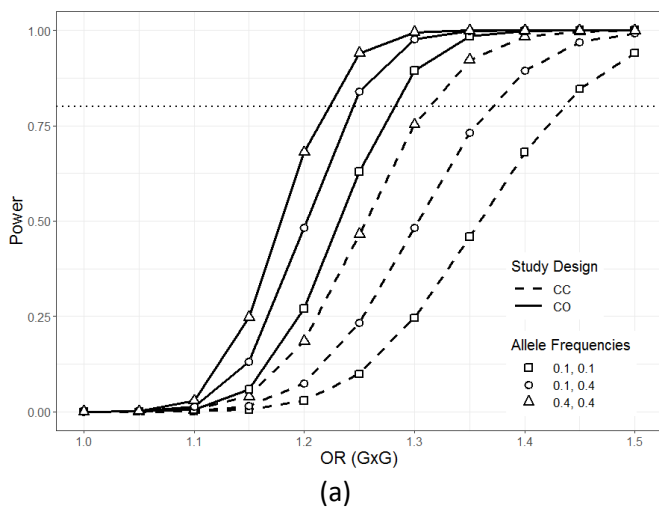
$$\text{logit}\{P(G_2 = 1)\} = \delta_0 + \delta_1 G_1 \quad (2)$$

According to Piegorsch et al. (Piegorsch, Weinberg, and Taylor 1994), no typical confounders such as age or sex can be included in the CO model, because their main effects on D cannot sensibly be modelled when using a CO design. One would be purely modelling the interaction of the confounder and G_2 .

For the CO design to be valid, two assumptions must be fulfilled (Piegorsch, Weinberg, and Taylor 1994):

1. disease of interest is sufficiently rare, i.e. has prevalence of less than 5%,
2. the two risk factors under study are uncorrelated in the general population.

The CO study design brings many advantages in comparison to the CC study design when analysing interaction effects. Firstly, as the name states, it obviates the need for controls. Without controls some issues that are disadvantageous in the CC study design regarding the selection of adequate controls do not occur. No need for controls also has financial aspects for a study, as it is less time and money consuming to consider only cases and no resources are needed for the search of suitable controls. Methodologically, the CO study design is attractive due to the gain in statistical power (given the same number of cases), in comparison the CC study design (W. J. Gauderman 2002).



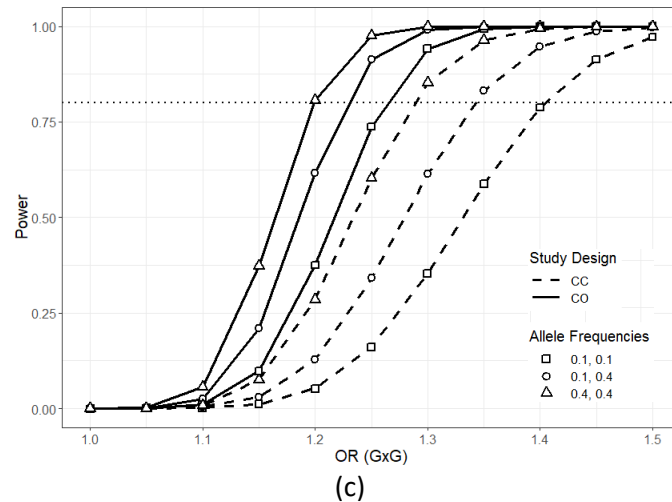


Figure 1: Power of the CO (solid line) and CC (dashed line) designs of GxG analysis (calculated using the Quanto software with parameter settings apt for (a) CD, (b) UC and (c) PD; Bonferroni-corrected significance levels: 3.66×10^{-6} , 4.32×10^{-6} , 1.25×10^{-5} for CD, UC and PD respectively). Pairs of SNP allele frequencies are marked by symbols: 0.1, 0.1 (square); 0.1, 0.4 (circle); 0.4, 0.4 (triangle). Dotted horizontal line marks 80% statistical power.

When analysing GxG interactions, factors that must be considered when assessing statistical power include the allele frequencies of the respective SNPs in question, disease prevalence rate and of course, sample size. In addition, when examining GxE interactions, the prevalence of the environmental factor must also be considered. Tailored to represent the sample sizes from the datasets of the IIBDGC (16 636 CD, 12 888 UC) and IPDGC (17 415) I had at my disposal and the respective disease prevalence of CD, UC and PD, Figures 1a-1c show the statistical power for CC and CO study designs to detect GxG interactions, given a dominant mode of inheritance. It is obvious that, regardless of the allele frequency pairing, the CO design achieves dramatically higher statistical power in comparison to the CC design. Let us explore one hypothetical case of GxG interaction in more detail: allele frequencies of 0.1 and 0.4 for two respective SNPs are given, prevalence of the disease is set at 0.3% for CD and main gene effects of 1.08 are present. In such a scenario, 80% statistical power is achieved with a GxG interaction OR of already 1.23 in the CO design, whereas in the CC design, 80% statistical power is achieved with an observed interaction OR of 1.35, as seen in Figure 1a. Therefore, if the research question focuses on interactions, the CO study design has a major advantage of higher statistical power, in comparison to the standard CC study design.

While the CO study design proves to have strong advantages in comparison to the CC study design, it also has drawbacks. Similarly to the CC study design, the CO study design is still subject to all issues associating the selection and documentation of cases, for example, recall bias tainting the data on exposures experienced (Khoury and Flanders 1996). Moreover, due to the nature of the CO study design, only interaction effects can be investigated, main effects can only be analysed in the CC study design. Therefore, one can say that the CC study design can give more diverse information about the risk factors in comparison to the CO study design. By far the largest disadvantage of the CO study design is the need to fulfil the assumption of independence of the two risk factors under investigation. To make sure that the independence assumption is fulfilled, a two-step procedure may be used. It involves first testing for

correlation between the SNPs in the general population and then, based on the results stratifying the analysis by using either the CO or the CC study design where appropriate (Mukherjee et al. 2008). However, this approach may introduce the burden of selecting an appropriate population reference sample, and, in scenarios where controls need to be used, abolishes one major advantage of the CO study design – no need for controls.

Linkage disequilibrium (LD) proves to be an issue when investigating G×G interactions, because it also causes the departure from the independence assumption (see Chapter 2.2 for more details). Moreover, in G×E interaction studies it is often presumed that the genes and environment risk factors are independent. This, however, is not always the case and can lead to false positive results. Albert et al. have shown that the CO study design is sensitive to even small amounts of G×E association in the general population and results may easily be distorted due to the departure from the independence assumption (Albert 2001). Given the disadvantages of the CO study design and the imperious advantage of higher statistical power in comparison to the CC study design, some methodologies have been developed combining the two approaches. One idea is to use a two-step method by first screening the dataset of cases and controls for marginal effects using the CO equation (2) and then using the standard CC study design to test for interactions (Murcray, Lewinger, and Gauderman 2008). Other combinations of the two methods that try to keep the advantage of higher statistical power from the CO design and the stability of the CC design include model averaging (D. Li and Conti 2008), an empirical Bayes estimator (Mukherjee and Chatterjee 2008) and a weighted empirical Bayes estimator (Mukherjee et al. 2012). However, all these combinations of the study designs still need controls, because they include the CC study design.

1.4 Gene-Environment Interactions

A simple example of G×E interaction in human diseases is phenylketonuria (PKU). PKU is caused by recessive mutations in the gene *PAH* that codes for an enzyme needed to break down phenylalanine, which is found mostly in foods that contain protein (Manta-Vogli and Schulpis 2018). If individuals with this genetic predisposition consume phenylalanine through their diet, they become susceptible to PKU, which can cause neurological disorders including seizures, skin rashes, psychiatric disorders, among others. Thus, by controlling the environment, i.e. the diet, of the individuals with the genetic predisposition, i.e. genetic risk factor, PKU will not develop. Another example for G×E interaction is melanoma. It has been shown that ultraviolet light from, for example, sunlight (environmental factor) increases the chances of melanoma in fair-skinned individuals in comparison to dark-skinned (genetic component) (Rees 2004). However, defining G×E interactions in complex diseases such as IBD or PD could be far more challenging due to the numerous factors that play a part in the etiology of these diseases.

The precise definition of G×E interactions slightly depends on the field of discipline and context. A vague definition of G×E interaction describes it as purely some sort of interplay between genetic and environmental factors (Dempfle et al. 2008). Sometimes it may even be used to describe several factors that both contribute to disease risk, without meaning that the two factors are completely independent. Interaction can be defined as biological (causal) or

statistical (H. J. Cordell 2002). Biological interaction is described as the joint effect of two factors that interact physically or chemically with one another, or if they affect the same disease-relevant biological pathway (Yang and Khoury 1997). Functional studies are carried out to examine these etiological mechanisms due to biological interaction. Statistical interaction, on the other hand, is defined as “departure from additivity of effects on a specific outcome scale” (Rothman, Greenland, and Lash 2008). Effect modification is also a term used synonymous to statistical interaction, meaning that the risk difference associated with a given factor G depends on risk factor E . Ideally, findings of statistical interaction give ground for further research of plausible biological interaction. However, biological and statistical interaction do not necessarily have to overlap and statistical interaction does not imply a useful biological mechanism (Cowman and Koyutürk 2017).

G×E interactions are not only believed to play an important role in complex diseases, but have significant applications in personalized medicine when it comes to treatment and prevention (Dempfle et al. 2008). Pharmacogenetics study how people respond to drug therapy (environmental factor) based on their genetics. The main idea is that patients with different genetic make-up would benefit from (or be harmed by) different medications and/or dosages and these effects would be predictable (Gardiner and Begg 2005). Regarding disease prevention, specific recommendations or intervention plans can be developed if the environmental risk effects strongly depend on an identified genetic predisposition. While there are obviously some interventions that can yield great effects for some diseases, such as dietary changes for individuals with PKU, there is also criticism on this topic that should be considered. On the one hand, studies can easily produce biased estimations of effects and overestimate risk factors (Ioannidis et al. 2001). On the other hand, there is the argument that strong intervention for high-risk individuals may be smaller than weak intervention on the whole population, resulting in a higher overall benefit of smaller, but more wide-spread interventions (Rose 2001). Despite these valid arguments, G×E play an important role in complex diseases and increasingly more research is done in this field.

Family-based as well as population-based study designs can be used to investigate G×E interactions. Clayton and McKeigue (Clayton and McKeigue 2001) provide an overview of possible study designs depicting the main advantages and drawbacks of each. While the CO study design has its disadvantages as discussed in Chapter 1.3, it also has the overwhelming advantage of high statistical power given the same number of cases as the CC study design (Yang, Khoury, and Flanders 1997). Therefore, in my dissertation, I will be focusing on the CO study design and using a logistic regression model for studying G×E interactions. Similar to the example in Chapter 1.3, the logistic regression model for G×E in the CO design is:

$$\text{logit}\{P(E = 1)\} = \beta_0 + \beta_1 G \quad (3)$$

Here, environmental factor E is binomial, categorizing the environmental factor as either present or absent. The genotype of the SNP (G) is treated as a predictor variable with β_1 corresponding to the interaction effect. As mentioned before, in this CO logistic regression, no typical confounders should be included in the CO model (Piegorisch, Weinberg, and Taylor 1994). The two assumptions (assumption of independence of the two risk factors and the assumption that the disease under study is sufficiently rare) apply as well.

While the CO study design has many advantages as discussed in Chapter 1.3, the limitations of the CO study design for identifying G×E interactions must be kept in mind. The assumption that the gene and environmental factor under investigation are independent in the population can be problematic. Simulations have shown that even small departures from the independence assumption can alter the results (Albert 2001). Smoking is quite often perceived as an independent environmental factor. However, studies have shown that there are genetic predispositions that cause nicotine addiction, which could cause some individuals to smoke for longer periods of their lives than others (Weiss et al. 2008). This in turn, could introduce bias when investigating G×E interactions. Population stratification is an issue that must be addressed. Population stratification can either be genetic, environmental or jointly genetic and environmental (Yadav, Freitag-Wolf, Lieb, Dempfle, et al. 2015). Moreover, the misclassification of genotypes may also lead to spurious associations (Cheng and Lin 2009). Misclassification of the environmental exposure could introduce bias. The quality of the information about the environmental factor may include recall bias or an unclear definition of the environmental exposure and thus introduce bias into the study. These measurement errors lead to dramatic reductions in the power of the G×E interaction test and increase the bias towards the null hypothesis (Garcia-Closas, Rothman, and Lubin 1999). Although some issues may be improved or resolved through a well-planned study, larger sample size or different statistical methods (Aschard et al. 2018; M. Y. Wong et al. 2004), the threat of bias should be considered. Overall, the G×E interaction model can only be as reliable and respectable as the input data for it.

1.5 Gene-Gene Interactions

G×G interactions, also known as epistasis, are defined as the interplay between different genes (H. J. Cordell 2002). There has been an increasing interest in G×G interactions as it may explain part of the missing heritability of complex diseases and provide insights into the biological processes that cause, or make one susceptible to complex diseases. To date, a lot of evidence for G×G interactions comes from model organism studies, for example yeast, nematodes and flies (Mackay 2014). Interest of G×G interactions in complex human diseases is also emerging. A study on knee osteoarthritis has found G×G interactions (Fernández-Torres et al. 2020) and, as mentioned in Chapter 1.2, smaller studies in CD and UC have shown evidence of interplay of two SNPs.

Similar to G×E interactions, there is a little confusion as to what exactly is meant by G×G interactions. It can be understood as functional/biological, compositional or statistical epistasis (Phillips 2008). Biological interaction focuses on the molecular interactions that proteins as well as other genetic elements have with one another and how biological pathways are affected (Boone, Bussey, and Andrews 2007). Compositional epistasis refers to the blocking of one allelic effect by another in a different location on the genome (Phillips 2008). In my dissertation, I refer to G×G interactions in the statistical sense, identically to G×E statistical interaction, meaning effect modification and the departure from a linear model. Identical to G×E interaction definitions, it must be noted that statistical interaction does not imply biological interaction.

Although there are a few approaches for studying G×G interactions, such as family-based or CC studies, I will be focusing on the CO study design using the logistic regression to test for G×G interactions. Descriptions of other possible approaches can be found in the review by Cordell (Heather J Cordell 2009) and Clark et al. for mathematical approaches in the CO study design for G×G interactions (Clarke, Pettersson, and Morris 2009). While the term G×G interactions is used, the actual calculations are executed using SNPs. The logistic model for G×G interactions is similar to that for G×E interactions, only the environmental exposure E is interchanged with the second genotype as depicted in Function 2 and described in Chapter 1.3. According to Piegorsch et al., no confounding variables can be incorporated in this model (Piegorsch, Weinberg, and Taylor 1994). The same assumptions as for G×E apply, namely the assumption that both factors, in this case genes, are independent in the general population and that the disease of study is sufficiently rare.

The fulfilment of the independence assumption is slightly problematic in the G×E interaction analysis, however, it is even more so in G×G interactions. The CO study design is only suitable for SNPs that are not in linked in any way. Thus, linkage disequilibrium (LD), population stratification and cryptic relatedness pose difficulties. The CO study design is not universally applicable to all possible SNP pairs, as those within close distance to each other or show correlation due to other reasons (Heather J Cordell 2009). For example, certain genotype combinations are related to viability and would thus be inappropriate to test. False positive results could also be created by not taking population stratification into account. Moreover, ignoring technical artefacts such as batch effects could potentially bias the results. Mukherjee et al have proposed a two-step procedure considering the independence assumption when analysing interactions in the CO study design (Mukherjee et al. 2008). It involves first testing for correlation between the SNPs in the general population and then, based on the results stratifying the analysis by using either the CO or the CC study design where appropriate. However, this approach may introduce the burden of selecting an appropriate population reference sample, and, in case controls need to be used, abolishes one mayor advantage of the CO study design – no need for controls. Thus, there is potential for a methodology where the advantages of the CO study design are retained and the associations are addressed in order to retain the validity of the CO study design.

An issue that is not applicable to G×E interactions, but poses a problem with G×G interactions is the computational burden. Practical issues may arise due to the fact that the number of possible pairwise SNP combinations equals to the quadratic function of the number n SNPs under investigation:

$$\frac{n \times (n - 1)}{2} \quad (4)$$

This makes genome-wide analysis of all possible SNP pair combinations computationally challenging. For example, the Illumina HumanOmni2.5-8 chip, which is quite often used in GWAS covers 2.5 million SNPs, which would generate more than three trillion possible SNP pairs for investigation.

Given that many statistical tests, the multiple testing problem becomes an issue as it makes sufficient type one error control crucial. Pecanka et al. has proposed a two-step procedure by first performing an independence test on all SNP pairs and then analysing only those which passed the test (Pecanka et al. 2017). The first step is less computationally demanding; therefore, it reduces calculation time. Moreover, fewer tests in the second step reduce the multiple testing burden as the statistical significance threshold is only adjusted by the number of tests conducted in the second step. There are some computational solutions, for example by Wan et al. (Wan et al. 2010), that could decrease the computational burden of investigating G×G interactions. Cowman and Koyutürk propose a method of reducing the number of G×G interaction tests and thus addressing the multiple testing problem through hierarchical representation of genomic redundancies (Cowman and Koyutürk 2017). Another approach, that I used in my studies (Chapters 2.2 and 2.3) involves focusing the G×G interaction analysis by preselecting SNPs with known main effects, because they are more likely to exhibit interaction effects. Thus, the number of possible G×G interactions not only poses a computational problem, but a statistical multiple testing issue as well.

1.6 Genotype Imputation

Genotype imputation is the process of predicting (a.k.a. imputing) genotypes that have not been directly assayed in a given sample of observations by using a reference panel of haplotypes (Marchini and Howie 2010). It is used for a number of reasons: data harmonization, improvement of statistical power and to increase the overall number and density of genotyped variants for association testing (Naj 2019). If whole genome sequencing would be used in all studies, it would be simple to merge and compare the data between different studies. The maximum amount of information would be available on over 300 million variants identified to date (Sherry 2001), including rare ones. This, however, would be an expensive and time-consuming process. Therefore, different GWAS chips from different companies (e.g. Illumina or Affymetrix) are usually used in studies, which do not fully overlap in their sequenced variants. When meta analysing the results, data from different studies can be harmonized and assessed by imputing the non-overlapping variants. Data harmonization in turn can increase the statistical power of a study by increasing the number of observations available for association testing. For moderately strong associations (i.e. with and OR between 1 and 2), samples sizes of over 10 000 are required to obtain sufficient statistical power (Naj 2019). This is a rather difficult task for small research centres and, therefore, cooperation through consortia and meta analysing the results is an attractive option. Finally, increasing the density of genotyped variants can be beneficial due to the fact that the associated SNPs in GWAS are not necessarily the causal variants, but variants in high LD with the causal variant. Thus, for fine-mapping, a higher density of SNPs is advantageous. Genotype imputation can be summarised as a useful tool to combine and improve association analyses.

There are numerous algorithms for genotype imputation that have been improved over the years. Some of the programs, which use different algorithms include MaCH (Y. Li et al. 2010), minimac4 (Das et al. 2016), IMPUTE2 (Howie, Marchini, and Stephens 2011), BEAGLE (Browning, Zhou, and Browning 2018), PLINK (Purcell et al. 2007) and fastPHASE (Scheet and

Stephens 2006). Each of these algorithms have their strengths and weaknesses that must be considered based on the needs and goals for imputing. For example, PLINK and BEAGLE are more computationally efficient than the others, but MaCH, minimac4 and IMPUTE2 are more accurately imputing rare and low frequency variants (Naj 2019). A thorough overview of the advantages and disadvantages of each method is summarised by Das et al. (Das, Abecasis, and Browning 2018). It is therefore important to keep in mind that there is no one universally ideal imputation algorithm and to choose a plausible option tailored to the needs and possibilities of the study.

Genotype imputation methods can be categorized as either single-step or two-step. The one-step approach is used for localized imputation and consist of imputing the actual genotypes. Beagle, PLINK and fastPHASE can be used for this method. In the two-step approach the data is first pre-phased, meaning that haplotypes are estimated and is then imputed in the second step. This method is favoured by MaCH, minimac4 and IMPUTE2. The first step in the two-step method is more computationally consuming, but is much faster in the actual imputation, which usually makes the whole process faster (Naj 2019).

In practice, genotype imputation can either be executed on in-house computers or by using an imputation server. Examples of fast and efficient imputation servers include the Michigan Imputation Server (Das et al. 2016) and Sanger Imputation Server (the Haplotype Reference Consortium 2016). Using in-house computers for genotype imputation has the advantages of having full control of the algorithm used, yet may be time and computer capacity consuming. Uploading data to an imputation server has the advantages of being fast and efficient, yet may involve more time preparing the data in the format accepted by the imputation server, have idle waiting time due to limited job lots, is not always flexible in the reference bases available and may data privacy issues must also be considered.

Due to the fact that genotype imputation uses a reference base for the algorithm to impute missing genotypes, it is logical, that the quality of imputation depends on the quality and appropriateness of the reference base. Popular reference bases include 1000 Genomes (The 1000 Genomes Project Consortium 2015) with 2 504 individuals and 49 143 605 SNPs and the Haplotype Reference Consortium (HRC) (the Haplotype Reference Consortium 2016) with 32 470 observations and 39 635 008 genotyped sites. Both reference bases cover a variety of different populations. While it would be intuitive to assume that the larger the reliable reference base, the higher the imputation accuracy will be, this is not the case because of population stratification. For example, in case of an African American population, the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) (CAAPA et al. 2016) would be a more suitable selection of the reference base, even though it consists of only 883 observations, it would reflect the population better. Therefore, it must be kept in mind that the genotype imputation quality can only be as high as the quality and suitability of the reference base.

After the genotype imputation has been conducted, it is necessary to assess the quality of the genotype imputation in order to determine the success of the imputation and identify poorly imputed variants, which should be excluded from any further association analyses. There are a few different genotype imputation quality metrics that have been developed and different genotype imputation programs vary in the quality metrics that they apply. For example,

minimac4 generates an R^2 quality score, which denotes the ratio of the empirically observed variance of the imputed allele dosage to the expected binomial variance at Hardy-Weinberg equilibrium. On the other hand, IMPUTE2 incorporates a score that measures the relative statistical information about the variant allele frequency derived from the imputed data (Marchini and Howie 2010). Unfortunately, there is also no consensus on the appropriate threshold of the quality scores for the selection of SNPs for further analysis. The minimac4 R^2 is widely used and there are opinions that suggest values higher than 0.3 are appropriate (Scott et al. 2007), other consider 0.5 (Meschia et al. 2011) or 0.8 (Anderson et al. 2008) as adequate cut-off thresholds. Another popular approach is to consider a varying R^2 threshold depending on the allele frequency (Das, Abecasis, and Browning 2018). The varying metrics of imputation quality used by different imputation methods make it slightly more difficult to compare the imputation methods with each other. However, there are several protocols published that can be followed to adequately impute and filter the results (Das, Abecasis, and Browning 2018; J. Chen et al. 2019).

Although genotype imputation can have many advantages for further analyses, it also has drawbacks. Genotype imputation errors can cause bias in further association analyses. Population stratification can introduce bias if the reference base ethnically differs from the imputation dataset. Studies have shown that populations of African American and Asian descent are more difficult to impute than those of Caucasian descent (Schurz et al. 2019). Small study and reference sample sizes, missing rates of higher than 50% and window sizes smaller than 500 SNPs all negatively affect imputation quality (B. Zhang et al. 2011). Moreover, there are certain areas in the genome which are more difficult to impute. For example, the major histocompatibility complex (MHC) on chromosome 6p21.3 is a region that is highly polymorphic, which results in highly varied haplotype combinations, making imputation more difficult (Naj 2019). At the same time, it is a region with many variants associated with numerous phenotypes of complex diseases and thus sparking research interest (Shiina et al. 2009). Another interesting issue is the potential bias of imputing data of a sample consisting only of cases if the reference base includes only controls. This will be discussed further in Chapter 2.1 by examining the effects of genotype imputation on the power of statistical tests for G×E interactions when only cases are used. It is therefore important to consider possible pitfalls when imputing data in order not to bring too much bias into further analyses.

There are still issues regarding the practicability of studies of G×G and G×E interactions that need to be systematically analysed. The effect of genotype imputation on the validity and power of statistical tests for G×E interactions in CO studies had not been discussed yet, but will be covered in Chapter 2.1. Moreover, due to the high statistical power the CO study design brings, it is interesting to analyse G×G interactions in this specific study design. Chapters 2.2 and 2.3 of my thesis will analyse, address and explore G×G interactions using large datasets of complex diseases cases.

Chapter 2

2 Results

2.1 The effect of genotype imputation on the validity and power of statistical tests for gene-environment interactions in case-only studies

2.1.1 Summary

The effects of genotype imputation in a case-only (CO) design have not been considered to date. The CO study design is a powerful approach to study gene-environment (G×E) interactions using only affected subjects. Genotype imputation uses a reference sample such as the Haplotype Reference Consortium (HRC) to predict genotypes at untyped loci. However, using healthy controls as a reference in a CO study may introduce systematic error, especially in regions of genetic main effects. Thus, in this paper I investigated the effects of imputation accuracy on the validity and power to detect G×E interactions in a CO design.

The data used comprised of 719 CD cases and 2491 healthy controls from Kiel, Germany along with 118 selected target SNPs (59 with main effects (MEs), 59 without ME in the Kiel sample). The healthy controls were only used to determine the MEs in the Kiel sample using the standard case-control study design. The analysis consisted of two main steps. In the first, the true genotypes of selected genotyped SNPs were masked and then imputed using the Michigan imputation server with HRC as a reference base and the results were compared with the true genotype. Levels of linkage disequilibrium (LD) to the nearest neighbouring SNP were varied in ten steps by pruning the data before imputation accordingly. In the second step, I simulated a G×E interaction effect 10 000 times on the target SNPs with MEs with an imputed minor allele frequency (MAF) of at least 0.005 and imputation missing rate of less than 0.2, totalling to 56 target SNPs. Thus, the statistical power as well as the bias in the beta of the G×E interaction could be analysed for SNPs with MEs and different MAFs.

One of the main findings is the notion that imputation bias may be introduced in areas with MEs. The results showed that imputation was most unreliable in gene regions with known MEs. These include *NOD2*, *NKD1* and *CYLD* genes, which are known to be important contributors to CD (Cleyne et al. 2014). Here, the estimated MAF based on imputed genotypes was more similar to MAF in controls than to the true MAF in the CO sample. The agreement between true and imputed genotypes decreased more drastically for SNPs located in gene regions with known MEs as the maximum LD threshold decreased. Moreover, the study shows, that given very low MAFs of less than 0.05, the imputation accuracy score for

SNPs of interest with ME is still high, even if the imputed MAFs differ from the CO sample's MAF. The simulation of environmental exposure led to the observation that in cases when the MAF is low and/or the LD to the nearest SNP is not high, the imputation of SNPs with MEs may lead to further power loss and underestimation of effects when analysing G×E interactions.

In conclusion, this paper describes constellations in which imputed data should be used with caution when testing for G×E interactions in CO studies. SNPs with MAF of lower than 0.05 and/or ME should be carefully handled, as an imputation bias may arise which further incriminates the analysis of G×E interactions. It all comes down to the fact that genotype imputation uses a reference base for the algorithm to impute missing genotypes - it is logical, that the quality of imputation depends on the compatibility of the reference base. Reference base is not only important for population stratification with regards to ethnicity, but to disease status as well. Cases genetically differ from the general population and thus genotype imputation is not as accurate as for controls, which in turn provides pitfalls in areas of MEs. In these areas there was not enough statistical power to identify G×E interactions due to the inaccuracy imposed by imputation. Therefore, when analysing G×E interactions, it must be kept in mind that if no interactions are found, this does not mean that there are not any. Rather, the issue may be that there was not enough statistical power to find G×E interactions, especially for imputed SNPs with low MAFs.

2.1.2 Publication

The Effect of Genotype Imputation on the Validity and Power of Statistical Tests for Gene-Environment Interactions in Case-Only Studies

Milda Aleknonytė-Resch¹, Silke Szymczak², Sandra Freitag-Wolf¹, Michael Krawczak¹, Astrid Dempfle^{1*}

1. Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany
2. Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, Germany

*corresponding author, contact details:

Prof. Dr. Astrid Dempfle, dempfle@medinfo.uni-kiel.de

Introduction

The genetic aetiology of most if not all common complex diseases such as, for example, cancer, diabetes and asthma is still poorly understood. General progress in this direction has been hampered by the fact that the diseases in question result from a large number of genetic and environmental factors, each with only a small effect upon disease risk. The consequent causative complexity is exacerbated further by a number of related phenomena (Manolio et al. 2009) including gene-gene (G×G) and gene-environment (G×E) interaction, among others.

The precise meaning of the word ‘interaction’ depends upon the context in which it is being used, either as a biological (causal) or as a statistical term (Cordell 2002; Dempfle et al. 2008). Biological interaction usually refers to the combined effect of two factors that interact physically or chemically, or that affect the same disease-relevant biological pathway (Yang and Khoury 1997). Statistical interaction, by contrast, is defined as the “departure from additivity of effects on a specific outcome scale” (Rothman et al. 2008). It is tantamount to so-called ‘effect modification’, meaning that the risk difference associated with one factor on a certain scale depends upon the presence or absence of the other risk factor. Ideally, statistical interaction points towards plausible biological interaction, but the two need not necessarily coincide (Cowman and Koyutürk 2017). Moreover, there is no such thing as a lack of statistical interaction because, whenever the effects of two risk factors are additive on one scale, this cannot hold true on any other scale. In the following, we will focus upon statistical interaction of risk factors on the logit scale, i.e. we shall deal with departures from the multiplicity of odds ratios (OR).

Genetic epidemiological studies of common complex diseases employ different designs and methods, and the case-control (CC) design has emerged as the ‘work horse’ in this context, particularly in the form of genome-wide association studies (GWAS) of single nucleotide polymorphisms (SNPs). For studies of G×E interaction, however, the case-only (CO) design has also received some attention (Piegorsch et al. 1994) because it provides several advantages

compared to the CC design (Gauderman 2002b; Kraft et al. 2007). First and foremost, only cases (i.e. patients affected by the disease of interest) are required, which obviates the oftentimes difficult identification and recruitment of suitable controls (Schulz and Grimes 2002). Second, to detect G×E, the CO study design entails a substantial gain in statistical power over a CC design, using the same number of cases (Gauderman 2002a). On the other hand, however, the reliability of CO studies of G×E hinges upon the validity of two critical assumptions, namely (i) that the disease of interest is sufficiently rare (i.e. has prevalence $\leq 5\%$, say) and (ii) that the two risk factors under study (genetic and environmental) are uncorrelated in the general population (Piegorsch et al. 1994). Although the last presumption may often seem justified *prima facie*, it still needs to be reviewed carefully from study to study. For example, some variants in genes associated with alcohol metabolism are known to be linked to alcohol consumption (Goldman et al. 2005), and even such a minor gene-environment association can lead to false positive results in CO studies of G×E (Albert 2001).

For some time now, researchers into the genetic basis of common complex disease have been trying to improve the evidential capacity of GWAS through genotype imputation (Marchini and Howie 2010). With this technique, genotypes of untyped SNPs are inferred from the genotypes of typed SNPs by way of exploiting population-level linkage disequilibrium (LD). Genotype imputation has since become a standard for GWAS, because it facilitates the harmonization of SNP panels, improves statistical power by increasing sample size, and allows greater genomic coverage in terms of the number and density of the SNPs considered (Naj 2019). Genotype imputation can be performed either offline or using web-based services such as the Michigan Imputation Server (Das et al. 2016) or the Sanger Imputation Server (The Haplotype Reference Consortium 2016). Notably, all software available for this purpose provides means to assess the quality of the genotype imputation, usually through the provision of an imputation quality score. The Michigan Imputation Server used in the present study, for example, generates an R^2 quality score that relates the empirical variance of the imputed genotypes to its expectation at Hardy-Weinberg equilibrium (Das et al. 2016).

Irrespective of the potential to improve evidential capacity, genotype imputation is still an error-prone technique that can cause bias in subsequent analyses. Thus, the presence of hidden population stratification, the use of an inappropriate imputation base and a lack of sufficient SNP coverage may all negatively affect imputation quality (Zhang et al. 2011; Das et al. 2018; Schurz et al. 2019). Moreover, the imputation quality achievable in a certain GWAS may still vary substantially along the human genome (Naj 2019). On the other hand, misclassification of genotypes is known to cause spurious gene-environment associations in CC studies (Wong et al. 2004), and Cheng and Lin (2009) demonstrated how genotype misclassification can reduce the power of both the CC and the CO design.

We previously examined the role of LD for the validity of CO studies of G×E (Yadav et al. 2015b, a), and subsequently developed means to allow for hidden population stratification in such studies (Yadav et al. 2015b, a). Extending this earlier work, we here present an investigation of how genotype imputation accuracy influences the validity and power of CO studies of G×E, an aspect that to our knowledge has not been studied in detail to date. In our simulation study, we paid specific attention to the fact that, in regions with genetic main effects on disease risk, haplotype frequencies, and hence LD, are bound to differ systematically between cases and

controls. We also considered different environmental exposure frequencies when simulating the G×E interactions in order to cover a broad range of realistic GWAS set-ups.

Methods

Data

The main goal of our study was to analyse under which conditions imputed genotypes allow reliable inference of the presence and size of G×E interactions, using a CO design. Even although our investigations were mostly simulation-based, we nevertheless chose to employ real SNP genotype data in order to ensure realistic haplotype patterns in our test samples (Kulle et al. 2005; Ramnarine et al. 2015). Systematically varying the parameters of interest, namely the G×E odds ratio (OR) and the environmental exposure frequency, individual exposure states were then randomly assigned to individuals according to their given SNP genotypes.

The data, which comprised 719 Crohn disease (CD) patients and 2491 healthy controls from Northern Germany, were kindly provided to us by the PopGen biobank (Krawczak et al. 2006). All individuals had been genotyped before for 156,499 SNPs, and the data coincided with the 'Germany, Kiel' set used by Yadav et al. (2017) in their global study of gene-smoking interactions. Hence, the data underlying the present study had been subjected to the same quality control measures as employed in the earlier study.

Analysis

Our analysis consisted of two steps: First, we masked the genotypes of a number of selected SNPs in the cases, followed by the imputation of the missing genotypes using a suitable imputation base (see below). Then, we compared the minor allele frequencies (MAFs) among true and imputed SNP genotypes to one another and to those of the Haplotype Reference Consortium (The Haplotype Reference Consortium 2016) European population (HRC), which served as the imputation base in our study. Finally, the imputed genotypes were compared to the true genotypes, using Cohen's kappa as a means to quantify the respective level of genotype concordance for each SNP. In the second step, we simulated binary environmental exposure states (1: exposed, 0: non-exposed) for the cases each time depending upon the presumed G×E odds ratio and the original genotype of the SNP under study (henceforth referred to as the 'target SNP'). Consideration of different target SNPs meant that we could study the effects of genotype imputation upon the validity of subsequent G×E interaction analyses under different scenarios regarding MAF and main effect OR.

All statistical analyses were carried out with R (v. 3.5.0) or PLINK2 (Chang et al. 2015), as appropriate. For statistical modelling, SNP genotypes (G) were encoded assuming a dominant G×E effect of the minor allele, i.e. G=1 for homozygous or heterozygous carriers of the minor allele, G=0 for homozygous carriers of the major allele. A dominant model was used here because it is capable of covering a wide range of plausible genotype-phenotype relationships (Guan et al. 2012).

The choice of SNPs for step 1 (i.e. genotype imputation) was based upon the respective MAF and the presence or absence of a main effect on CD risk. To this end, the disease ORs of SNPs were determined by way of a case-control logistic regression association analysis, adjusted for

the first 10 principal components of all SNPs that passed quality control so as to allow for possible population stratification (Price et al. 2006). SNPs with a disease association p value $\leq 10^{-5}$ were considered further and pruned according to the following similarity criteria: (i) OR difference ≤ 0.02 , (ii) MAF difference in cases ≤ 0.02 , and (iii) physical distance ≤ 15 kb. Pruning left 141 'independent' SNPs that were grouped into four MAF-defined categories: low (MAF <0.05), medium-low ($0.05 \leq \text{MAF} < 0.15$), medium-high ($0.15 \leq \text{MAF} < 0.25$) and high ($0.25 \geq \text{MAF}$). All main effect SNPs in the low (n=16) and medium-low (n=7) category were forwarded to imputation step 2 alongside 18 SNPs each, randomly chosen from the medium-high and high category. These 59 main effect SNPs were complemented by 59 randomly selected SNPs lacking a main effect, chosen according to the following matching criteria: (i) localization on the same chromosome as the respective main effect SNP, and (ii) a MAF difference to the matching SNP in cases ≤ 0.01 . A detailed list of the 118 target SNPs is provided in Supplementary Table 1. Only target SNPs with a main effect were forwarded to step 2 of the study (i.e. the simulation and analysis of G×E interaction). Here, however, we also excluded main effect SNPs with an imputed MAF < 0.005 or a missing rate > 0.2 so that the final number of target SNPs in step 2 equalled 56.

Imputation

Genotype imputation was carried out for the 118 target SNPs in 10 successive rounds, each time thinning further the set of SNPs underlying the imputation. While only the genotypes of the 118 target SNPs themselves were masked in the beginning, surrounding SNPs were sequentially LD-pruned by maintaining, in the n^{th} round (n=1 to 9), only SNPs with $r^2 < n \cdot 0.1$ to the target SNP. A script published by the Wellcome Centre for Human Genetics, Oxford, UK (<https://www.well.ox.ac.uk/~wrayner/tools/>) was used to prepare the datasets for genotype imputation with the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/>), selecting Quality control and Eagle v. 2.4 phasing for the latter. The HRC European data comprising 39,635,008 SNP genotypes from 32,470 samples (The Haplotype Reference Consortium 2016) served as the imputation base.

Simulation of G×E Interaction

G×E interaction analyses were simulated assuming two different values of the population-level environmental exposure frequency, namely 10% and 30%. Under a dominant model, G×E interaction manifests in cases via different exposure frequencies in carriers and non-carriers of the minor SNP allele. Hence, we simulated G×E interaction by assigning environmental exposure states to individuals depending also upon their respective (true) SNP genotype. The necessary genotype-specific probabilities of an environmental exposure were calculated in two steps: First, QUANTO (Gauderman 2002b) was used for each of the 56 main effect target SNPs to calculate the interaction OR that would be detectable with 80% statistical power, based upon the main effect OR and MAF of the SNP in question, the population-level exposure frequency and the case sample size (n=719). In the power calculations, a nominal significance level of 0.05 was assumed. QUANTO also requires a main effect OR for the environmental exposure, which was consistently set to 1.5. A summary of the resulting SNP genotype-specific exposure probabilities is provided in Supplementary Table 2. The null hypothesis of no G×E interaction, where all genotype-specific exposure probabilities equal the population-level environmental exposure frequency, was also simulated for each SNP to complement the corresponding G×E interaction analysis.

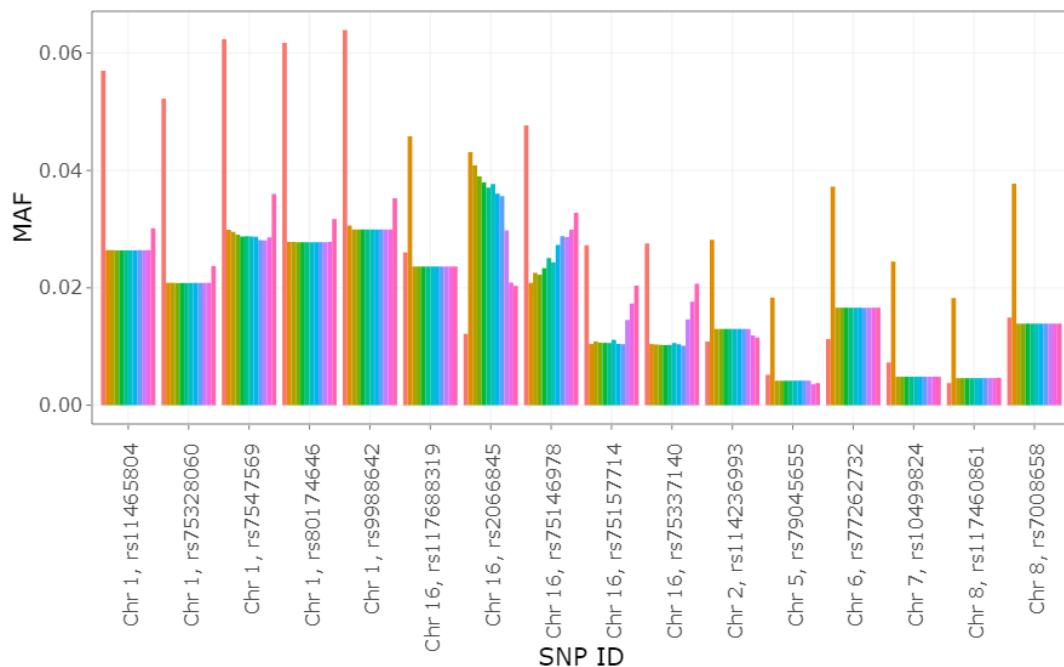
For each of the 56 main effect target SNPs, 10,000 replicates of the environmental exposure simulation were undertaken and the G×E interaction ORs determined for the 10 imputed genotype datasets (see 'Imputation'). Since a CO study design was used, the G×E interaction OR was estimated by logistic regression

$$\text{logit}\{P(E = 1)\} = \beta_0 + \beta_1 G$$

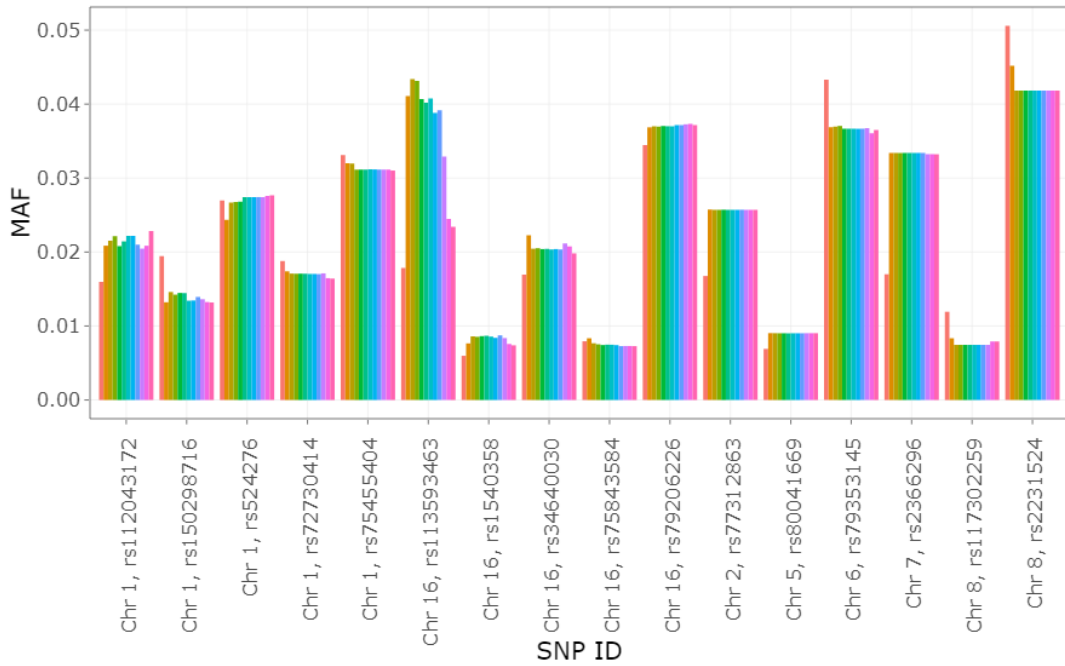
and the statistical significance (i.e. the p value) of $\{\beta_1 \neq 0\}$ was determined using a Wald test. Following Piergosch et al. (1994), no classical confounders such as age or sex were included in the regression model because their main effects cannot sensibly be modelled in a CO design.

Results

After analysing ten sets of varying possible maximum LD for 118 SNPs, our findings show that imputation works well for SNPs without ME, but not in all cases when MEs are present. As seen in Figure 1a, some SNPs with MEs, for example rs2066845, rs75146978, rs75157714 and rs75337140, exhibit a varying MAF, which increasingly resembles the MAF of the imputation reference population as the maximum LD threshold decreases. This observation can be seen not only in the low MAF group (Figure 1a), but in all MAF groups (Figures 1-3 in Appendix 3). A comparison of the MAF of the actual genotypes without ME of the CO data sample and the imputed datasets shows little variation, regardless of the MAF group as can be seen in Figure 1b and Figures 4-6 in Appendix 3. Regarding SNPs without ME, in most cases, the calculated Cohen's kappa coefficient showed little variation as the maximum possible LD level decreased (Figure 2b).

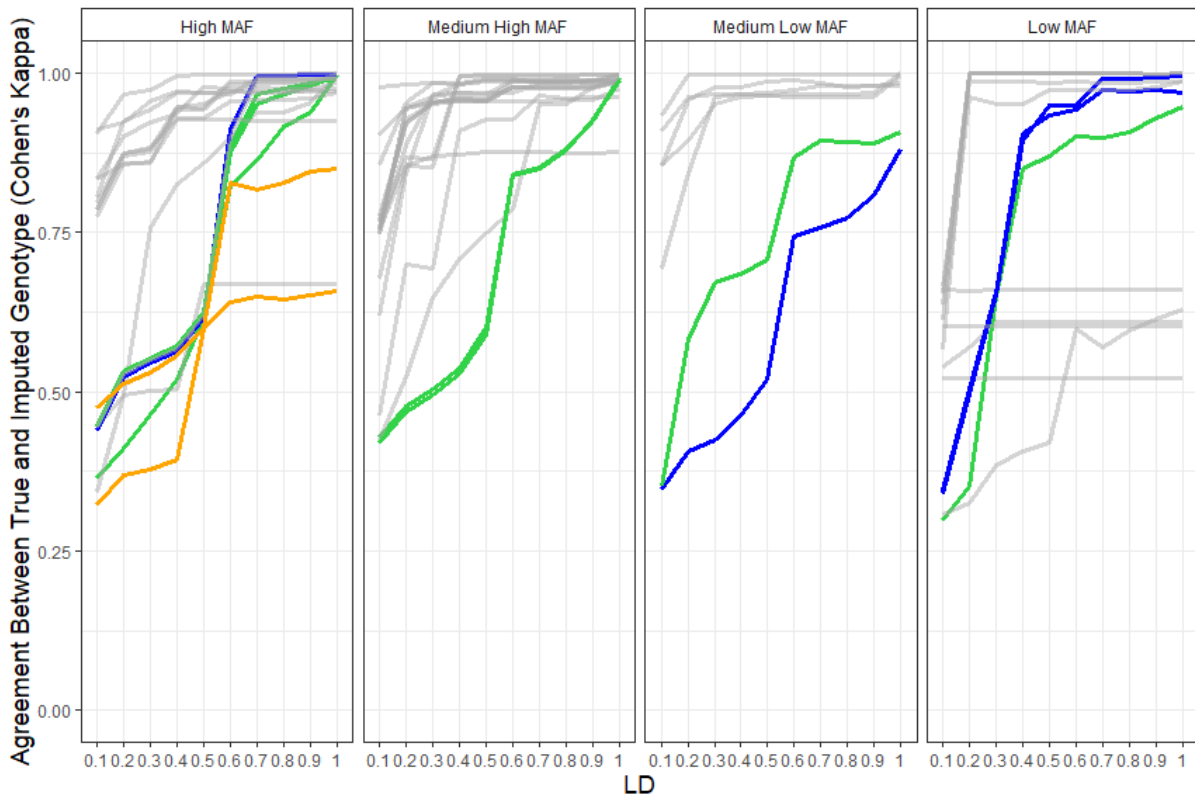


(a)



(b)

Figure 1: MAF of target SNPs. The histograms are grouped by target SNPs with low MAF with MEs (a) and without MEs (b). The first bar on the left of each group depicting the MAF of the HRC reference sample, which was used for imputation. The second column from the left portrays the SNP's of interest MAF in the CO data sample. Each following bar illustrates the SNPs' of interest MAF from the imputed datasets with LD pruning levels from 1 to 0.1 in decreasing order.



(a)

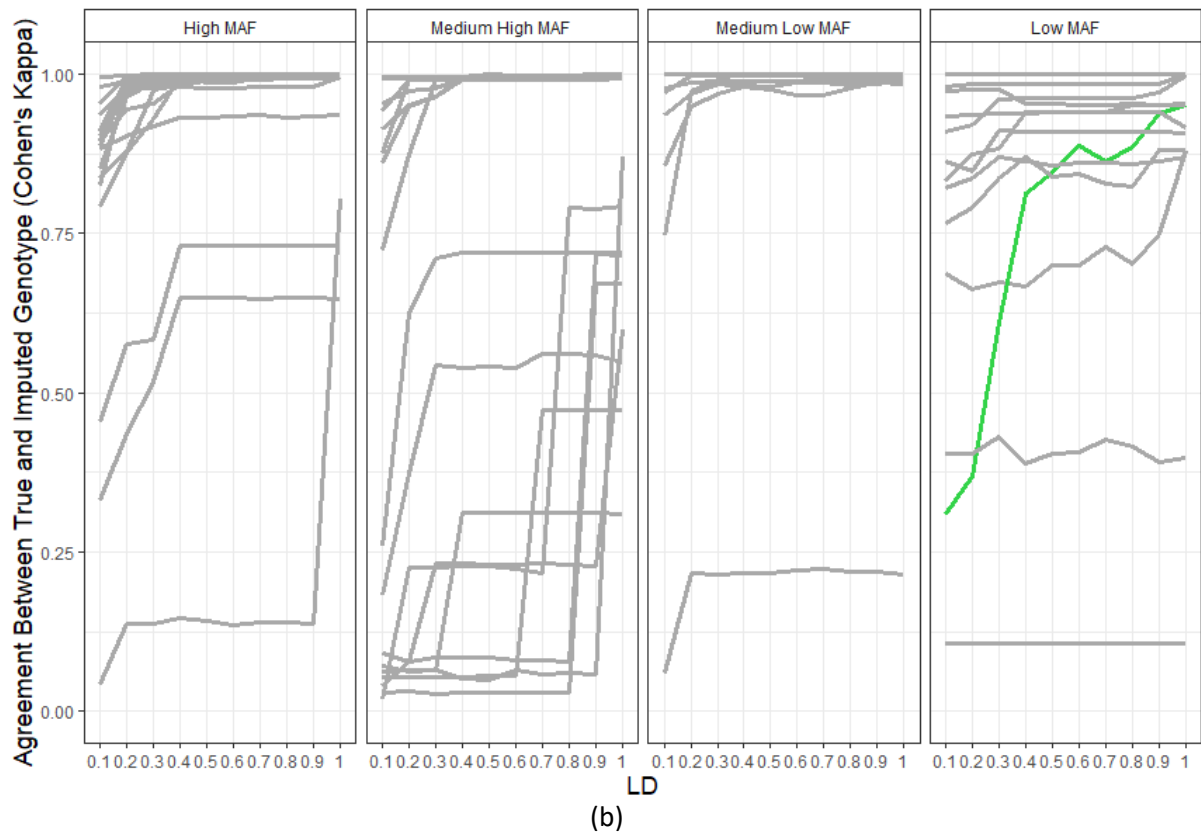


Figure 2: Kappa value of target SNPs (a) with and (b) without significant MEs calculated from our dataset with varying LD threshold of the highest possible LD. Green depicts SNPs in the *NOD2* gene region; orange: *NKD1* gene region and blue: *CYLD* gene region. Data is stratified by MAF groups, which are defined in the “Methods” section.

There are two cases in which the MAF of SNPs with MEs in the imputed datasets differs from the MAF in the CO sample. Firstly, when the MAF of the target SNP is less than 0.03. The MAF of the target SNP in the imputed datasets is imprecise regardless of the LD pruning level and resembles the MAF of the HRC reference sample. For example, as seen in Figure 1a, the MAF of rs79045655, rs10499824 and rs117460861 is about 0.02, however, the MAFs of the HRC and the imputed datasets are all closer to 0.005. Secondly, SNPs of interest with MEs in gene regions known to be strongly associated with CD such as *NOD2*, *NKD1* and *CYLD* (Cleyen et al. 2014) show a systematic decline in Cohen’s kappa as the maximum possible LD level threshold decreases (Figure 2a). While there is one SNP in the *NOD2* gene region present in target SNPs with low MAF group (MAF = 0.04) and no ME present, it had a calculated p-value of 3.34×10^{-4} and a ME OR of 1.82. As illustrated in Figures 1-3 in Appendix 3, regardless of the original MAF, for some SNPs with MEs, the difference between the imputed SNP MAF and the MAF of the reference sample decreases while the difference to the MAF in the CO sample increases as the LD pruning threshold decreases.

When considering imputed SNPs for further statistical analyses, the imputation accuracy score is used as a measure of quality control. Generally, the higher the imputation accuracy score is, the better the correspondence between the true and imputed MAF is expected. Scores of higher than 0.8 are considered adequate. However, our findings show, that given a MAF of less than 0.05 and present ME in the target SNP, a scenario can occur in which the estimated

imputation accuracy score calculated by the minimac4 algorithm is higher than 0.8, yet the agreement between true and imputed genotypes (Cohen's kappa) is low (Figure 3a).

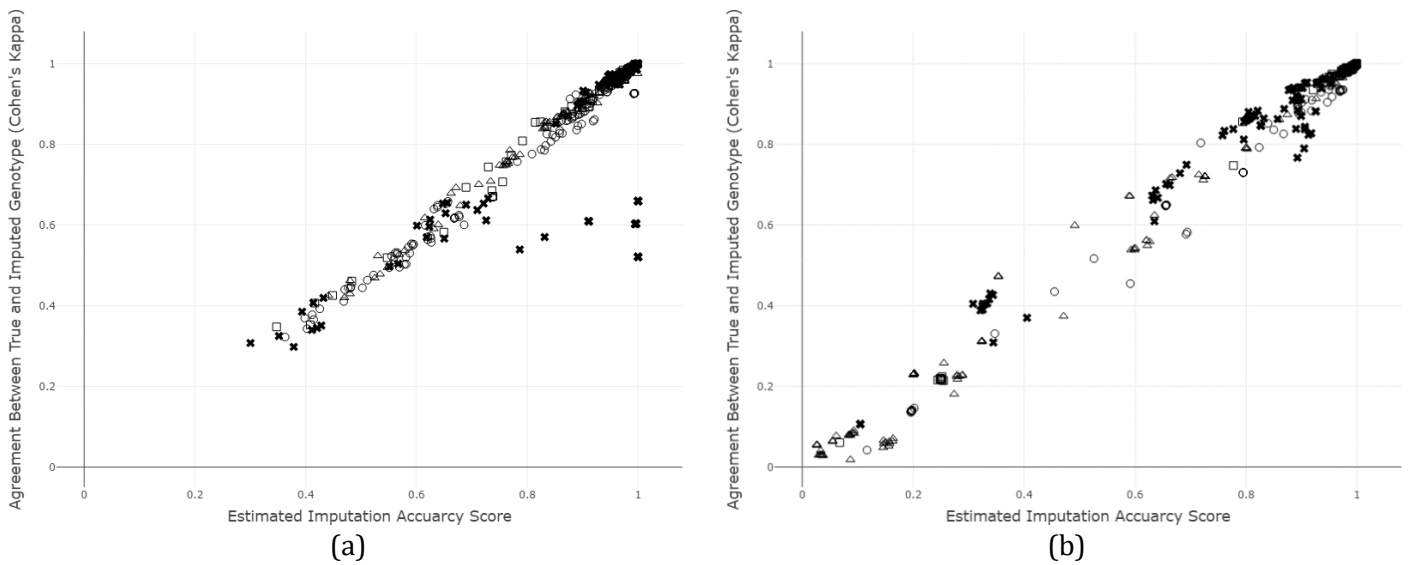


Figure 3: Estimated imputation accuracy score (R^2 from minimac4) against kappa in SNPs (a) with ME and (b) without ME by MAF group. MAF groups are depicted as follows: X - low (MAF less than 0.05), square - medium-low (MAF of 0.05-0.15), triangle - medium-high (MAF of 0.15-0.25) and circle - high (MAF of 0.25-0.5).

Our simulation of the G×E interaction effect in a CO study design using target SNPs with MEs can highlight three aspects regarding statistical power loss. First, given our sample size of 719 cases, a higher exposure frequency leads to higher statistical power to find G×E interactions (Figure 4). Second, statistical power depends on the MAF of the SNP, as seen in Figure 4, and Figure 1 in Appendix 4. Although our parameters were set to achieve 80% statistical power, the mean of the simulation results for SNPs with low and medium-low varied greatly from 80% (Figure 2 in Appendix 4). The interquartile range for low MAF SNPs was from 25-90% in most LD pruning threshold levels (Figure 2). In general, in our sample, a median statistical power of 80% was rarely achieved. Third, as can also be seen from Figure 4, regardless of the MAF group, there is a gradual power loss as the LD pruning levels decrease with a step-like decline when moving from LD pruning level of 0.6 to 0.5.

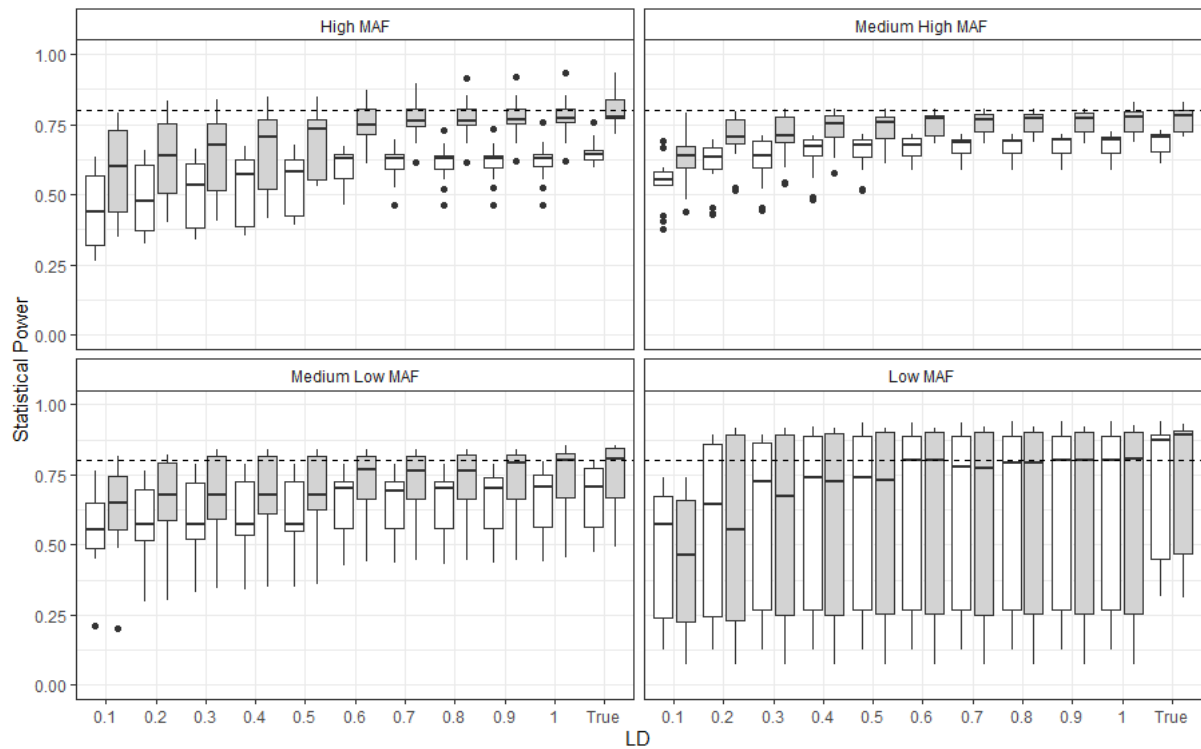


Figure 4: Boxplot of statistical power of the G×E interaction by LD threshold of the highest possible LD ranging from 0.1 to 1. “True” depicts the statistical power of the G×E interaction using data of the actual, not imputed SNP genotype. Boxes with grey (white) fill illustrate simulation results with environmental exposure frequency of 0.3 (0.1). Data is stratified by MAF groups, which are defined in the “Methods” section. Dashed horizontal line marks 80% statistical power.

Our analysis shows that a bias in the estimate of the G×E interaction beta coefficient can show up after imputation especially for SNPs with low MAF and the bias increases as the LD of the nearest SNP decreases. As seen in Figure 5, while there is a slight change in the difference in the beta coefficient in the medium and high MAF groups, the largest systematic underestimation is visible in the low MAF group.

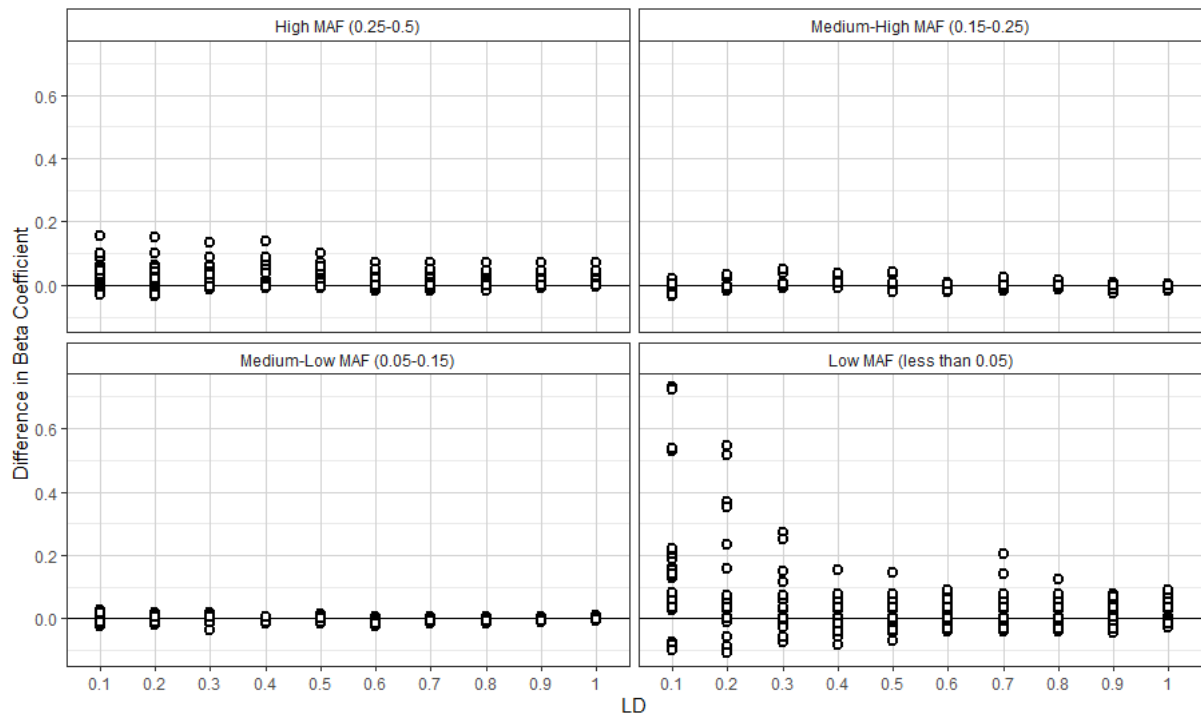


Figure 5: G×E interaction beta coefficient estimation bias calculated as the difference between true beta coefficient and that calculated from the imputed data grouped by SNP MAF.

Simulations were also conducted under the null hypothesis of no G×E interaction in order to assess the type I error rate of the imputation-based approach. The results suggest that the type I error rate was not systematically inflated, regardless of the level of LD pruning around the target SNPs (Figures 1-3, Appendix 5). On the contrary, the analysis was even found to be overly conservative in the low MAF group where the median type I error rate was well below 0.05 (Figure 2, Appendix 5). Finally, no differences were seen between the G×E interaction OR estimates obtained with true and imputed genotypes in any MAF category or at any level of LD pruning (Figure 4, Appendix 5).

Discussion

Using real genotype data on 719 CD patients with simulated G×E interaction effects, we compared the power of a G×E interaction test with true vs. imputed genotypes as well as the bias in the interaction beta coefficient estimation. We viewed realistic scenarios and therefore chose to analyse SNPs with varying MAFs, with and without ME present before imputation and after imputation with 10 different levels of maximum LD of the target SNP and surrounding SNPs. Given the different set-ups, the imputation accuracy score from the minimac4 algorithm was examined. Finally, environmental exposures were simulated 10 000 times and G×E interaction tests were calculated for each level of maximum possible LD for 56 SNPs of interest with MEs.

One of our main findings is the notion that imputation bias may be introduced in areas with MEs. Our results show, that imputation was most unreliable in gene regions with known MEs. These include *NOD2*, *NKD1* and *CYLD* genes, which are known to be important contributors to

CD (Cleyne et al. 2014). Regarding some SNPs of interest with MEs, their MAF gradually moves away from the MAF in the CO dataset and approaches the level of the MAF of the HRC reference dataset used for imputation as the maximum possible LD to the target SNP is decreased. There was one target SNP (rs113593463) present in the low MAF group without a ME which also belongs in the *NOD2* gene region. It showed a similar decline in Cohen's kappa as the other SNPs with MEs that were found in gene regions with known MEs. Since our definition of a present ME relied only on the calculations within our given dataset and a significance threshold of 5×10^{-5} , it must be noted that the p-value of rs113593463 ($p = 3.34 \times 10^{-4}$, ME OR = 1.82) does not fall far from our defined significance threshold. Given the target SNP's low MAF of 0.04, it is plausible that it may be statistically significant in other, larger datasets. This brings us to the conclusion that areas with underlying MEs are prone to introduce bias due to imputation as the maximum LD of surrounding SNPs decreases.

Moreover, our study shows, that given very low MAFs of less than 0.05, the imputation accuracy score for SNPs of interest with ME is still high, even if the imputed MAFs differ from the CO sample's MAF. This calls for caution when imputing SNPs with small MAFs in areas with known MEs. Furthermore, since not all ME areas are known, caution is always required when applying the CO study design, because without controls, ME cannot be estimated. Thus, the results of further analyses may include imputation bias and produce in particular false negative results, even if the imputation quality score is high.

Imputation bias is visible in further analysis of G×E interactions in the CO study design, as the results from our simulation demonstrate. The absolute difference in the G×E interaction OR increases as the LD of the neighbouring SNPs decreases when the MAF is less than 0.05. SNPs of interest with lower MAF achieve less statistical power. Thus, G×E interaction effects may remain hidden when the target SNPs are imputed. What is more, when examining the statistical power of G×E interactions of target SNPs with MEs, a sudden widening of the boxplot can be seen at the LD level of 0.5 in the high MAF target SNP group as well as an almost step-like decrease in statistical power in the low MAF target SNP group. The same tendency can be seen when examining the change in agreement between true and imputed genotypes of target SNPs (especially in gene regions with known MEs): there is a sudden drop in kappa when the LD threshold level changes from 0.6 to 0.5. Therefore, in certain cases, when SNPs with MEs are imputed in low LD regions, it may lead to further power loss and underestimation of effects when analysing G×E interactions.

Our study also highlights the fact that a large sample size is needed for G×E interaction studies with imputed data when investigating SNPs with low MAF and/or low environmental exposure frequencies. The statistical power in our simulation study was systematically lower in the low exposure frequency setting and a median statistical power of 80% was rarely achieved. This is partially due to our sample size of 719 cases, as given, say, a MAF of 0.021 and environmental exposure frequency of 10%, power calculations suggest that an interaction OR of 2.95 is enough to achieve 80% statistical power. Given a MAF of 0.03, in our case, a maximum of 15 cases could be carriers of the minor allele and be exposed to the environmental factor. Even if the probability of each case being exposed to the environmental factor is 24%, with only 15

cases, the chances of enough exposed cases for significant results is low. This is also seen by the varying statistical power of the true G×E interaction.

The number of genotyped SNPs is always a fraction of the genotype data available after imputation. These non-genotyped SNPs in the genome that are imputed may also be relevant for further analyses of G×E interaction and would be missed if one did not impute at all. However, the consequences of imputation should be considered when examining G×E interactions. It all comes down to the fact that genotype imputation uses a reference base for the algorithm to impute missing genotypes - it is logical, that the quality of imputation depends on the compatibility of the reference base with the CO sample. Reference base is not only important for population stratification with regards to ethnicity, but to disease status as well. Cases genetically differ from the general population (represented in the reference base) in regions with MEs. Thus, while genotype imputation will be accurate for controls, it will be systematically biased for cases in areas with MEs. This is well visible when using the CO study design to examine G×E interactions as it is more difficult to achieve reliable results using imputed data in areas with known MEs.

While the analysis of G×E interactions in our simulation worked well in numerous cases, there were scenarios in which it was problematic. These scenarios included target SNPs with low MAFs and regions not saturated by neighbouring SNPs. In these areas there was not enough statistical power to identify G×E interactions due to the inaccuracy imposed by imputation. Therefore, when analysing G×E interactions, it must be kept in mind that if no interactions are found, this does not mean that there are not any. Rather, the issue may be that there was not enough statistical power to find G×E interactions, especially for imputed SNPs with low MAFs.

Appendix 1: SNPs Used in Analysis

MAF Group	Chromosome	SNP with ME				Paired SNP without ME			
		BP Position	ID	ME OR	ME p Value	MAF	BP Position	ID	MAF
Low	1	67697069	rs75328060	0.39	1.61E-06	0.02	152191469	rs150298716	0.01
Low	1	67702526	rs11465804	0.43	1.64E-06	0.03	93342890	rs112043172	0.02
Low	1	67708155	rs80174646	0.40	1.81E-07	0.03	7810923	rs75455404	0.03
Low	1	67726104	rs9988642	0.44	8.67E-07	0.03	93499216	rs72730414	0.02
Low	1	67731368	rs7547569	0.44	8.57E-07	0.03	160737722	rs524276	0.02
Medium High	1	67607135	rs736197	0.63	4.75E-08	0.23	71929968	rs6703162	0.24
Medium High	1	67624304	rs12043240	0.62	1.98E-08	0.23	101543205	rs4617425	0.23
High	1	67661041	rs6656929	0.63	9.87E-08	0.27	205369075	rs1928442	0.28
High	1	67664840	rs58621044	0.63	1.05E-07	0.28	172854334	rs4916283	0.27
High	1	67665210	rs1569923	0.60	4.83E-09	0.31	92573670	rs11166163	0.31
High	1	67668970	rs6588249	1.55	4.32E-07	0.37	152599265	rs925976	0.38
High	1	67672765	rs7539625	1.61	3.26E-08	0.35	67744500	rs11209033	0.36
High	1	172838848	rs34884278	1.49	2.67E-06	0.34	198886404	rs1275161	0.34
Low	2	100576838	rs114236993	2.74	2.87E-06	0.03	181971974	rs77312863	0.03
Medium High	4	3467519	rs10022329	1.54	8.70E-06	0.16	102704134	rs6532967	0.16
Low	5	102135253	rs79045655	8.04	3.60E-09	0.02	96150110	rs80041669	0.01
Medium Low	5	40318419	rs7733749	0.64	3.08E-06	0.15	96004363	rs7704167	0.14
Medium Low	5	150174295	rs12655654	1.65	8.95E-06	0.11	26112192	rs9293221	0.11
Medium Low	5	150245724	rs10875560	1.66	3.85E-06	0.11	72493896	rs10036050	0.10
Medium High	5	40288878	rs4245975	0.66	9.31E-06	0.15	72484883	rs57125671	0.15
Medium High	5	150188072	rs10056694	1.55	5.20E-06	0.16	159849586	rs1895320	0.17
High	5	40323714	rs348595	0.68	6.34E-06	0.31	111682654	rs1019743	0.32
Low	6	34734915	rs77262732	2.24	7.82E-06	0.04	126797000	rs79353145	0.04
Medium High	6	32681992	rs3998158	0.64	2.09E-06	0.16	30388936	rs9261893	0.16
High	6	20809890	rs6902880	1.63	4.16E-08	0.41	29691019	rs1362126	0.41
High	6	28829486	rs209165	0.68	5.42E-06	0.25	29461914	rs1233490	0.26

MAF Group	Chromosome	SNP with ME				Paired SNP without ME			
		BP Position	ID	ME OR	ME p Value	MAF	BP Position	ID	MAF
Low	7	75171091	rs10499824	4.39	7.70E-09	0.02	50267315	rs2366296	0.03
Low	8	11006828	rs7008658	2.71	7.52E-08	0.04	141525327	rs2231524	0.05
Low	8	11128037	rs117460861	4.14	1.07E-06	0.02	129149288	rs117302259	0.01
High	12	113031474	rs233722	1.53	5.15E-06	0.46	40532151	rs937109	0.46
Medium Low	14	98409403	rs78558838	0.58	7.67E-06	0.07	81383427	rs7144634	0.06
Low	16	50756540	rs2066845	2.55	9.63E-08	0.04	50719743	rs113593463	0.04
Low	16	50762771	rs72796367	2.18	5.96E-07	0.05	75218899	rs79206226	0.04
Low	16	50810973	rs75337140	0.27	1.93E-06	0.01	11229329	rs1540358	0.01
Low	16	50826942	rs75157714	0.27	2.12E-06	0.01	30701362	rs75843584	0.01
Low	16	50846717	rs75146978	0.41	7.38E-06	0.02	30888012	rs34640030	0.02
Medium Low	16	50827601	rs2302759	0.57	4.45E-09	0.12	75370230	rs3844219	0.13
Medium Low	16	75244425	rs117688319	2.45	9.22E-08	0.05	75420854	rs74208577	0.06
Medium High	16	50751175	rs751271	0.55	6.33E-12	0.22	81821203	rs4584816	0.22
Medium High	16	50751972	rs13332952	0.53	3.65E-13	0.23	11296285	rs59785529	0.23
High	16	50565970	rs1990623	1.84	2.09E-12	0.26	11083069	rs56231421	0.27
High	16	50661273	rs9673419	1.49	3.29E-06	0.28	11281218	rs248831	0.29
High	16	50744624	rs2066842	1.89	3.96E-13	0.44	30172627	rs13331817	0.46
High	16	50752815	rs1861757	0.65	5.96E-07	0.34	11177824	rs12708715	0.34
High	16	50766127	rs3135499	0.63	1.32E-07	0.35	11254331	rs41367	0.35
High	16	50769563	rs718226	0.62	3.52E-08	0.36	11403753	rs28671554	0.36
High	16	50798929	rs2066851	0.63	1.61E-07	0.36	30934075	rs3813020	0.36
Medium Low	20	62308612	rs2738783	0.63	1.53E-06	0.14	48599561	rs4647955	0.13
Medium High	20	62203748	rs34681475	0.63	3.55E-07	0.18	1539350	rs2250199	0.18
Medium High	20	62219740	rs3810492	0.67	7.57E-06	0.19	4055167	rs16989193	0.19
Medium High	20	62315593	rs34894559	0.67	7.41E-06	0.17	30150077	rs6060002	0.17
Medium High	20	62342654	rs2750480	0.67	9.22E-06	0.17	57597645	rs6070696	0.17
Medium High	20	62343956	rs2315008	0.65	6.19E-07	0.24	10769074	rs6032951	0.23
Medium High	20	62348460	rs4809329	0.65	5.64E-07	0.24	36266789	rs6067117	0.24

MAF Group	Chromosome	SNP with ME			SNP without ME				
		BP Position	ID	ME OR	ME p Value	MAF	BP Position	ID	MAF
Medium High	20	62361737	rs2427530	0.67	8.40E-06	0.17	62291008	rs6089953	0.17
Medium High	20	62362563	rs6062509	0.67	2.22E-06	0.24	51318351	rs6013509	0.24
Medium High	20	62381979	rs6089970	0.62	2.85E-07	0.16	44757213	rs3765457	0.15
Medium High	21	16838572	rs35910543	0.66	1.53E-06	0.23	44487111	rs234711	0.24
High	21	16805220	rs1736135	0.66	1.28E-06	0.36	40486507	rs2037922	0.37

Appendix 2: exposure probabilities for major and minor allele carriers by SNP, and prevalence rate.

SNP ID	ME OR	Environmental Exposure Frequency	MAF in Cases	GxE Interaction OR for 80% Statistical Power	Genotype Exposure Probability	
					Minor Allele Carrier	Major Allele Carrier
rs75157714	0.27	0.1	0.01	8.10	0.46	0.10
rs75157714	0.27	0.3	0.01	6.10	0.72	0.30
rs75337140	0.27	0.1	0.01	8.10	0.46	0.10
rs75337140	0.27	0.3	0.01	6.10	0.72	0.30
rs117460861	4.14	0.1	0.02	2.10	0.19	0.10
rs117460861	4.14	0.3	0.02	1.80	0.43	0.30
rs79045655	8.04	0.1	0.02	1.85	0.17	0.10
rs79045655	8.04	0.3	0.02	1.65	0.41	0.30
rs75328060	0.39	0.1	0.02	4.45	0.32	0.10
rs75328060	0.39	0.3	0.02	3.45	0.59	0.29
rs75146978	0.41	0.1	0.02	4.30	0.31	0.10
rs75146978	0.41	0.3	0.02	3.35	0.58	0.29
rs10499824	4.39	0.1	0.02	1.95	0.18	0.10
rs10499824	4.39	0.3	0.02	1.70	0.42	0.30
rs11465804	0.43	0.1	0.03	3.85	0.29	0.09
rs11465804	0.43	0.3	0.03	3.00	0.55	0.29
rs80174646	0.40	0.1	0.03	3.80	0.28	0.09
rs80174646	0.40	0.3	0.03	3.00	0.55	0.29
rs114236993	2.74	0.1	0.03	2.05	0.18	0.10
rs114236993	2.74	0.3	0.03	1.75	0.42	0.30
rs7547569	0.44	0.1	0.03	3.55	0.27	0.09
rs7547569	0.44	0.3	0.03	2.80	0.54	0.29
rs9988642	0.44	0.1	0.03	3.50	0.27	0.09
rs9988642	0.44	0.3	0.03	2.80	0.54	0.29

SNP ID	ME OR	Environmental Exposure Frequency	MAF in Cases	GxE Interaction OR for 80% Statistical Power	Genotype Exposure Probability	
					Minor Allele Carrier	Major Allele Carrier
rs77262732	2.24	0.1	0.04	2.05	0.18	0.10
rs77262732	2.24	0.3	0.04	1.75	0.42	0.30
rs7008658	2.71	0.1	0.04	1.95	0.17	0.10
rs7008658	2.71	0.3	0.04	1.70	0.42	0.30
rs2066845	2.55	0.1	0.04	1.95	0.17	0.10
rs2066845	2.55	0.3	0.04	1.70	0.42	0.29
rs117688319	2.45	0.1	0.05	1.95	0.17	0.10
rs117688319	2.45	0.3	0.05	1.65	0.41	0.29
rs72796367	2.18	0.1	0.05	1.90	0.17	0.10
rs72796367	2.18	0.3	0.05	1.65	0.41	0.29
rs78558838	0.58	0.1	0.07	2.40	0.20	0.09
rs78558838	0.58	0.3	0.07	2.00	0.45	0.29
rs12655654	1.65	0.1	0.11	1.80	0.16	0.09
rs12655654	1.65	0.3	0.11	1.60	0.39	0.29
rs10875560	1.66	0.1	0.11	1.80	0.16	0.09
rs10875560	1.66	0.3	0.11	1.60	0.39	0.29
rs2302759	0.57	0.1	0.12	2.05	0.17	0.09
rs2302759	0.57	0.3	0.12	1.75	0.41	0.28
rs2738783	0.63	0.1	0.14	1.95	0.16	0.09
rs2738783	0.63	0.3	0.14	1.70	0.40	0.28
rs7733749	0.64	0.1	0.15	1.90	0.16	0.09
rs7733749	0.64	0.3	0.15	1.65	0.39	0.28
rs4245975	0.66	0.1	0.15	1.90	0.16	0.09
rs4245975	0.66	0.3	0.15	1.65	0.39	0.28
rs10022329	1.54	0.1	0.16	1.75	0.15	0.09
rs10022329	1.54	0.3	0.16	1.55	0.38	0.28

SNP ID	ME OR	Environmental Exposure Frequency	MAF in Cases	G×E Interaction OR for 80% Statistical Power	Genotype Exposure Probability	
					Minor Allele Carrier	Major Allele Carrier
rs6089970	0.62	0.1	0.16	1.90	0.16	0.09
rs6089970	0.62	0.3	0.16	1.65	0.39	0.28
rs10056694	1.55	0.1	0.16	1.75	0.15	0.09
rs10056694	1.55	0.3	0.16	1.55	0.38	0.28
rs3998158	0.64	0.1	0.16	1.90	0.16	0.09
rs3998158	0.64	0.3	0.16	1.65	0.39	0.28
rs2427530	0.67	0.1	0.17	1.85	0.15	0.09
rs2427530	0.67	0.3	0.17	1.60	0.39	0.28
rs2750480	0.67	0.1	0.17	1.85	0.15	0.09
rs2750480	0.67	0.3	0.17	1.60	0.39	0.28
rs34894559	0.67	0.1	0.17	1.85	0.15	0.09
rs34894559	0.67	0.3	0.17	1.60	0.39	0.28
rs34681475	0.63	0.1	0.18	1.85	0.15	0.09
rs34681475	0.63	0.3	0.18	1.60	0.39	0.28
rs3810492	0.67	0.1	0.19	1.85	0.15	0.09
rs3810492	0.67	0.3	0.19	1.60	0.38	0.28
rs751271	0.55	0.1	0.22	1.85	0.15	0.09
rs751271	0.55	0.3	0.22	1.60	0.38	0.28
rs13332952	0.53	0.1	0.23	1.85	0.15	0.09
rs13332952	0.53	0.3	0.23	1.60	0.38	0.28
rs12043240	0.62	0.1	0.23	1.80	0.15	0.09
rs12043240	0.62	0.3	0.23	1.55	0.37	0.28
rs35910543	0.66	0.1	0.23	1.80	0.15	0.09
rs35910543	0.66	0.3	0.23	1.55	0.37	0.28
rs736197	0.63	0.1	0.23	1.80	0.15	0.09
rs736197	0.63	0.3	0.23	1.55	0.37	0.28

SNP ID	ME OR	Environmental Exposure Frequency	MAF in Cases	G×E Interaction OR for 80% Statistical Power	Genotype Exposure Probability	
					Minor Allele Carrier	Major Allele Carrier
rs6062509	0.67	0.1	0.24	1.75	0.14	0.09
rs6062509	0.67	0.3	0.24	1.50	0.37	0.28
rs2315008	0.65	0.1	0.24	1.75	0.14	0.09
rs2315008	0.65	0.3	0.24	1.50	0.37	0.28
rs4809329	0.65	0.1	0.24	1.75	0.14	0.09
rs4809329	0.65	0.3	0.24	1.50	0.37	0.28
rs209165	0.68	0.1	0.25	1.75	0.14	0.09
rs209165	0.68	0.3	0.25	1.50	0.37	0.28
rs1990623	1.84	0.1	0.26	1.80	0.14	0.08
rs1990623	1.84	0.3	0.26	1.55	0.37	0.28
rs6656929	0.63	0.1	0.27	1.75	0.14	0.09
rs6656929	0.63	0.3	0.27	1.55	0.37	0.27
rs58621044	0.63	0.1	0.28	1.75	0.14	0.08
rs58621044	0.63	0.3	0.28	1.55	0.37	0.27
rs9673419	1.49	0.1	0.28	1.75	0.14	0.08
rs9673419	1.49	0.3	0.28	1.55	0.37	0.27
rs1569923	0.60	0.1	0.31	1.75	0.14	0.08
rs1569923	0.60	0.3	0.31	1.55	0.37	0.27
rs348595	0.68	0.1	0.31	1.75	0.14	0.08
rs348595	0.68	0.3	0.31	1.55	0.36	0.27
rs34884278	1.49	0.1	0.34	1.80	0.14	0.08
rs34884278	1.49	0.3	0.34	1.60	0.37	0.27
rs1861757	0.65	0.1	0.34	1.75	0.14	0.08
rs1861757	0.65	0.3	0.34	1.55	0.36	0.27
rs7539625	1.61	0.1	0.35	1.85	0.14	0.08
rs7539625	1.61	0.3	0.35	1.65	0.37	0.26

SNP ID	ME OR	Environmental Exposure Frequency	MAF in Cases	G×E Interaction OR for 80% Statistical Power	Genotype Exposure Probability	
					Minor Allele Carrier	Major Allele Carrier
rs3135499	0.63	0.1	0.35	1.75	0.13	0.08
rs3135499	0.63	0.3	0.35	1.55	0.36	0.27
rs718226	0.62	0.1	0.36	1.75	0.13	0.08
rs718226	0.62	0.3	0.36	1.55	0.36	0.27
rs1736135	0.66	0.1	0.36	1.75	0.13	0.08
rs1736135	0.66	0.3	0.36	1.55	0.36	0.27
rs2066851	0.63	0.1	0.36	1.75	0.13	0.08
rs2066851	0.63	0.3	0.36	1.55	0.36	0.27
rs6588249	1.55	0.1	0.37	1.85	0.14	0.08
rs6588249	1.55	0.3	0.37	1.65	0.37	0.26
rs6902880	1.63	0.1	0.41	1.95	0.14	0.08
rs6902880	1.63	0.3	0.41	1.70	0.37	0.25
rs2066842	1.89	0.1	0.44	2.05	0.14	0.07
rs2066842	1.89	0.3	0.44	1.80	0.37	0.25
rs233722	1.53	0.1	0.46	2.00	0.13	0.07
rs233722	1.53	0.3	0.46	1.75	0.36	0.25

Appendix 3: Comparison of the MAF of the CO data Sample and the Imputed Datasets by MAF group

The histograms are grouped by SNPs of interest, with the first bar on the left of each group depicting the MAF of the HRC reference sample, which was used for imputation. The second column from the left portrays the SNP's of interest MAF in the CO data sample. Each following bar illustrates the SNPs' of interest MAF from the imputed datasets with LD pruning levels from 1 to 0.1 in decreasing order.

With ME:

Figure 1: Medium-low MAF group

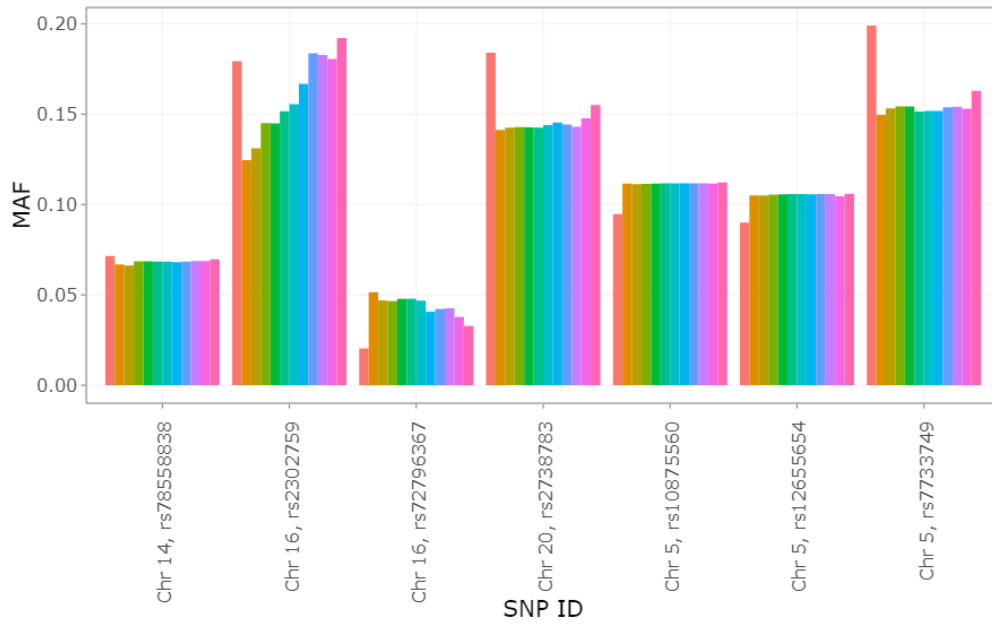


Figure 2: Medium-high MAF group

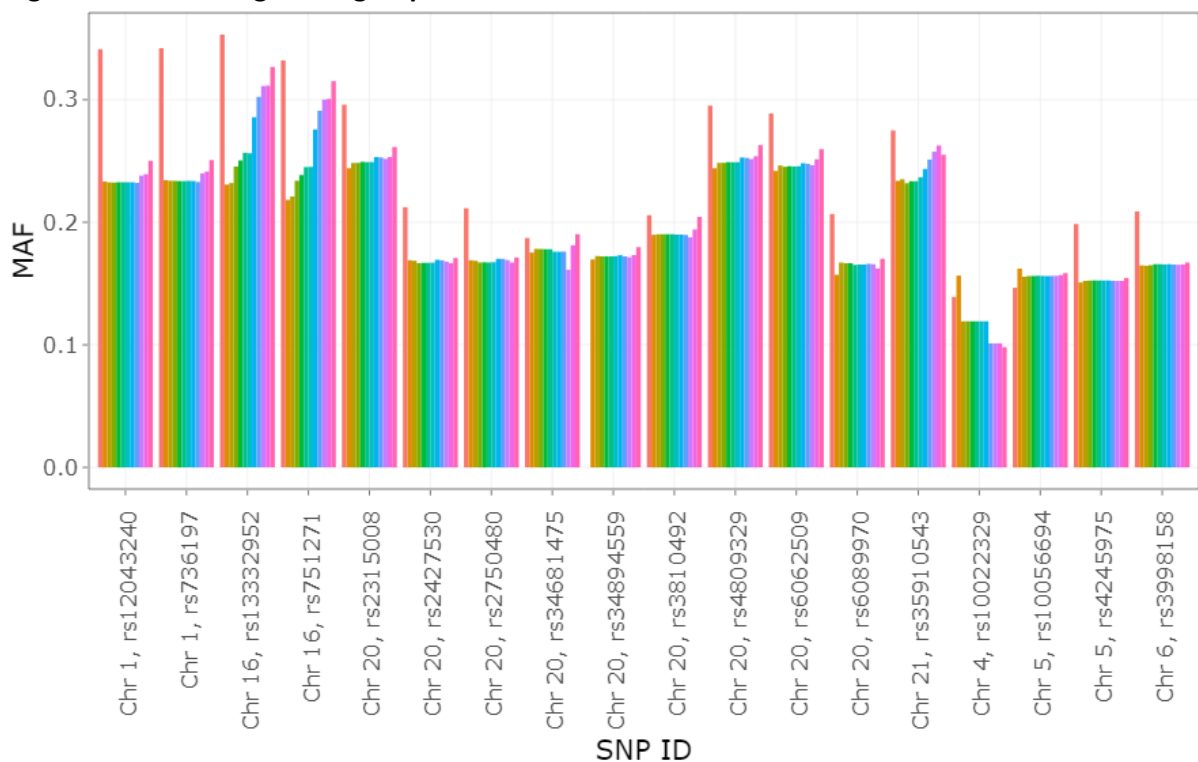
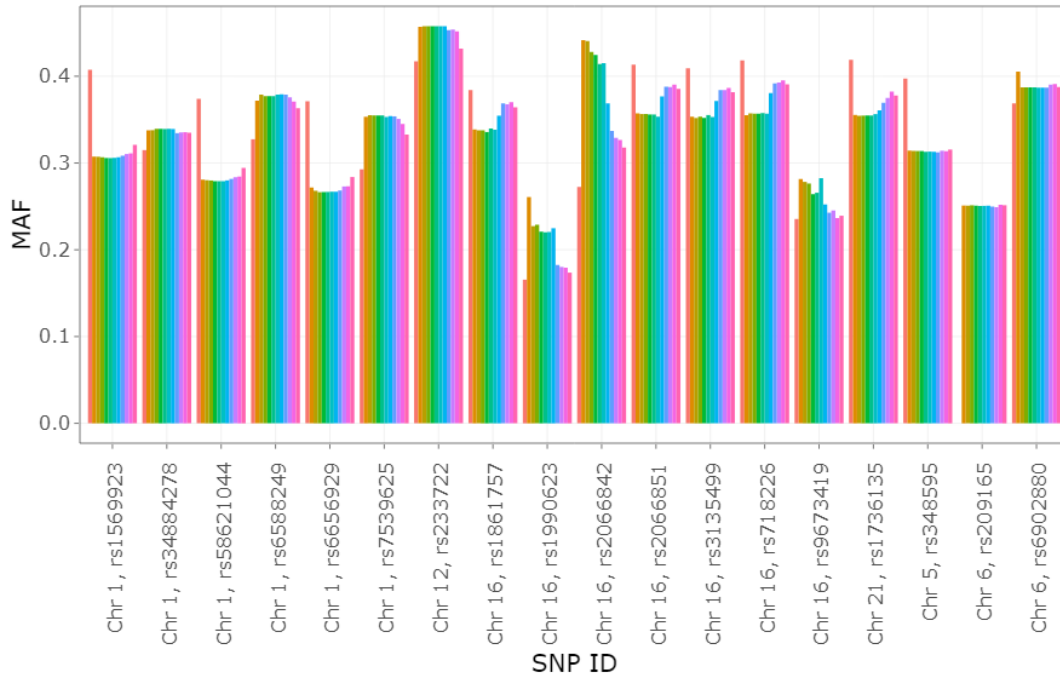


Figure 3: High MAF group



Without ME:

Figure 4: Medium-low MAF group

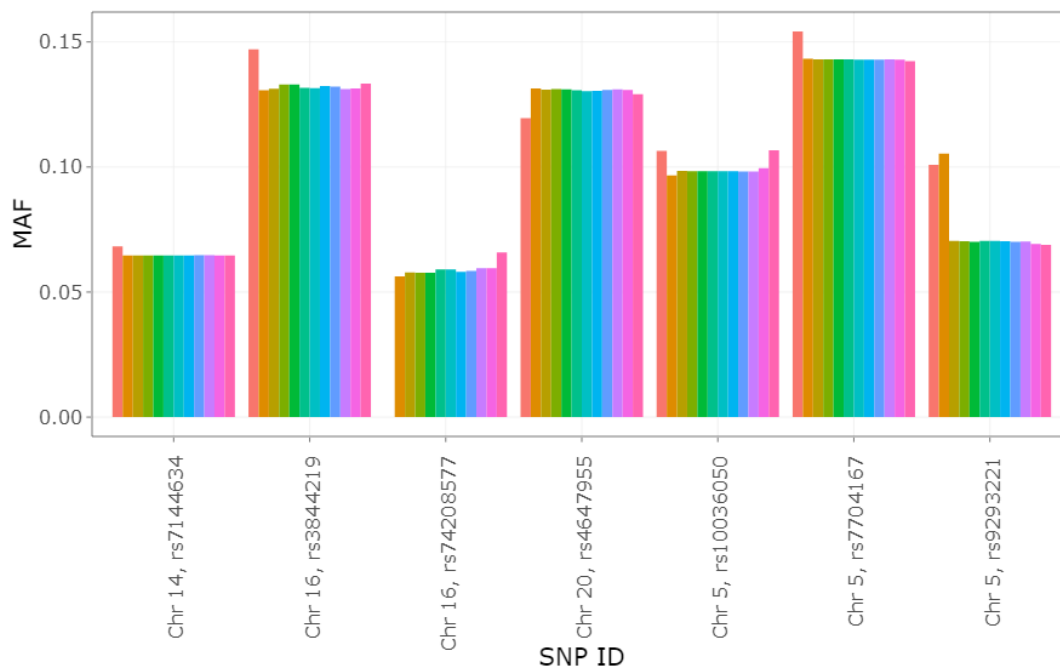


Figure 5: Medium-high MAF group

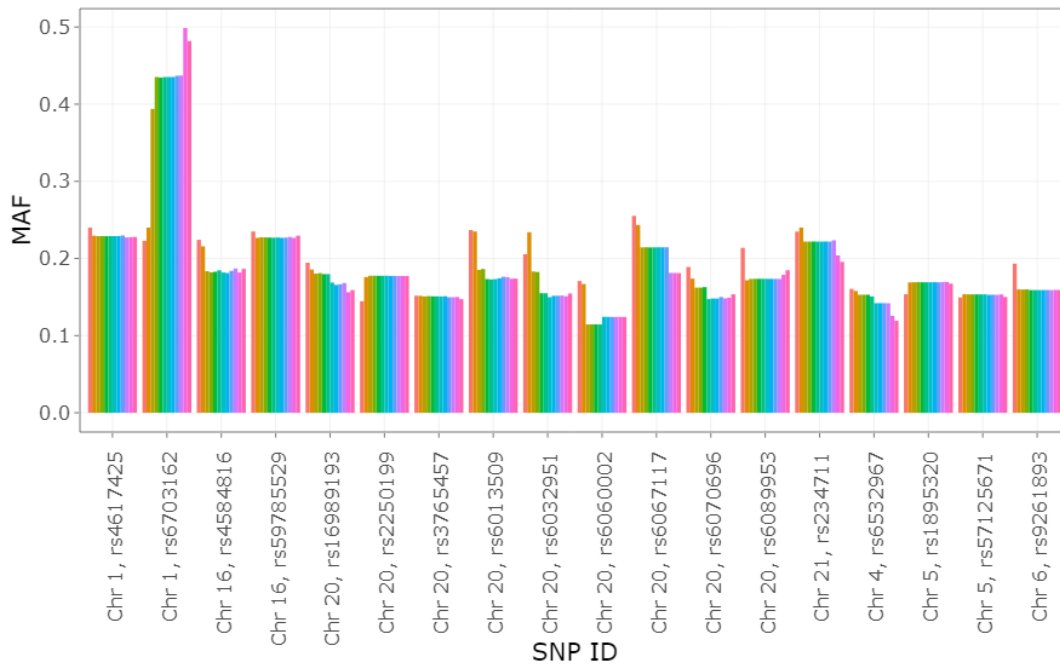
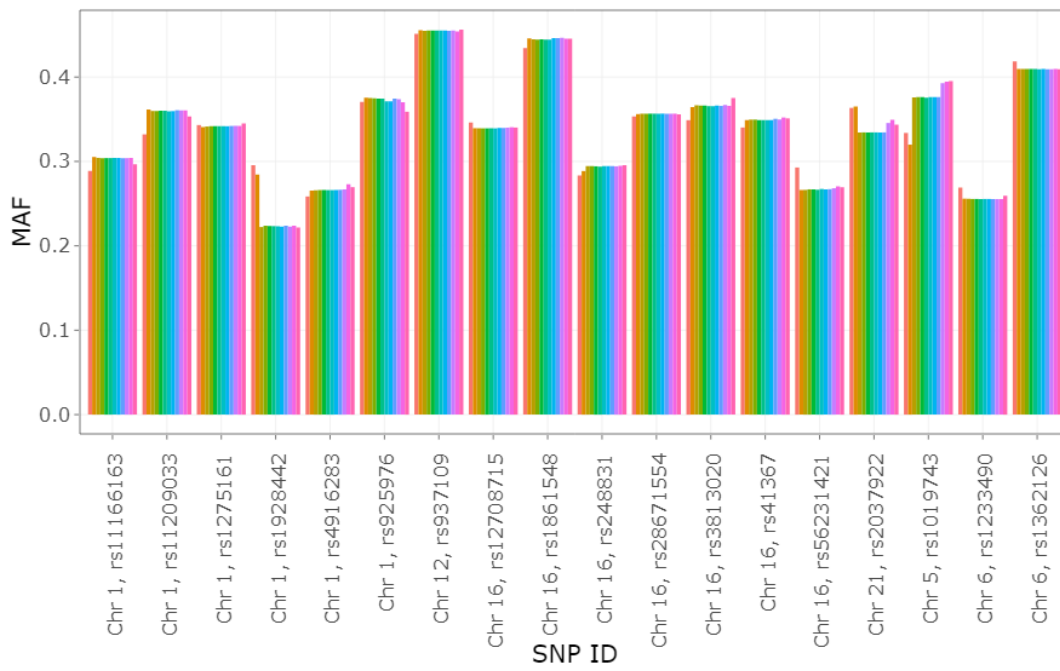


Figure 6: High MAF group



Appendix 4: Additional Figures depicting results of G×E interaction simulation.

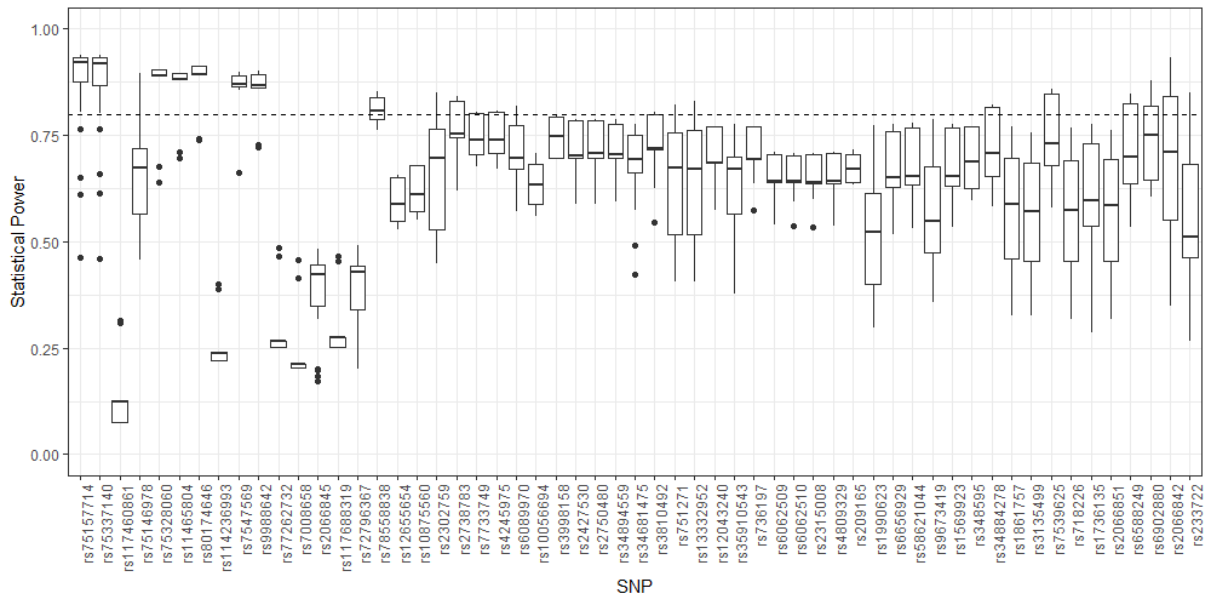


Figure 1: Boxplot of statistical power of G×E interaction by simulated SNP, ordered from left to right by increasing SNP MAF. Dashed horizontal line marks 80% statistical power.

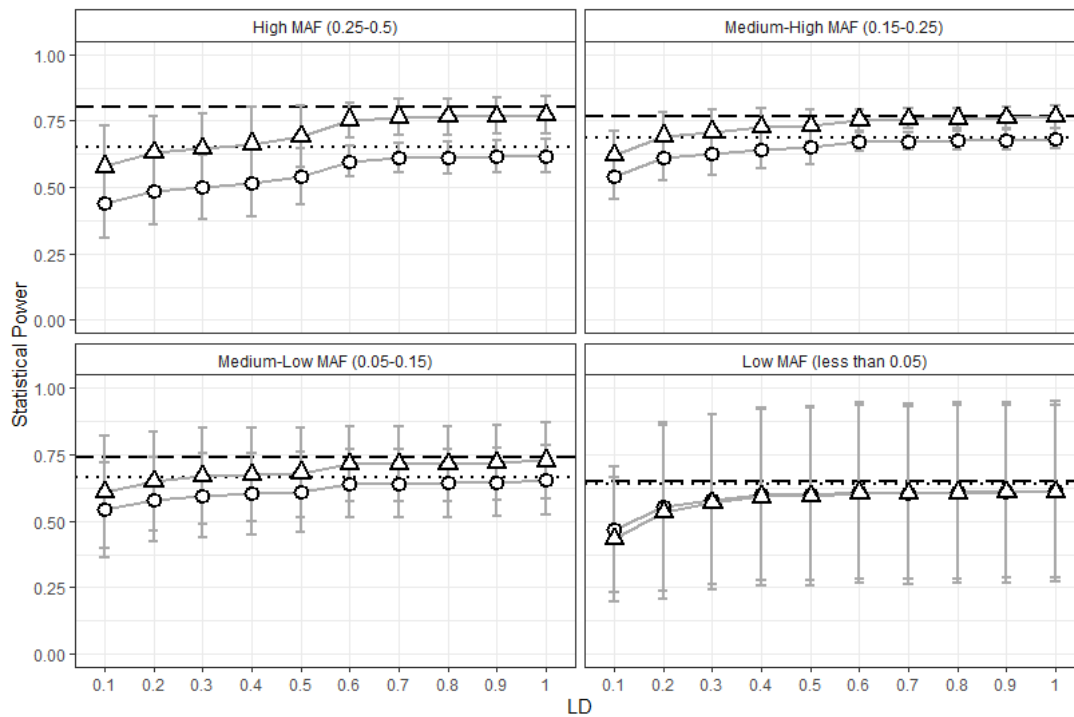


Figure 2: Mean statistical power of G×E interaction from imputed data by highest LD level of neighbouring SNP from 10 000 replications grouped by MAF. Circles depict means when the exposure frequency is equal to 0.1, triangles – 0.3. Dotted (dashed) line is the true statistical power of the genotyped, not imputed, SNPs for the 0.1 (0.3) exposure frequencies.

Appendix 5: H_0 simulation results

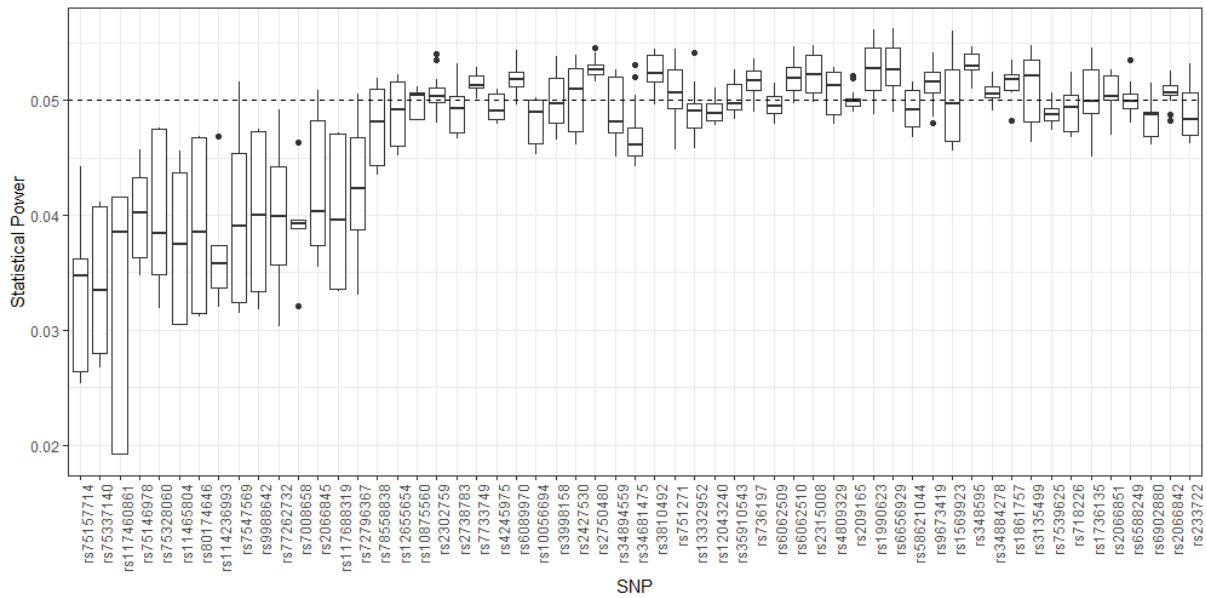


Figure 1: Boxplot of type I error rate of G×E interaction (under the hypothesis of no G×E interaction) by simulated SNP, ordered from left to right by increasing SNP MAF. Dashed horizontal line marks 5% type I error rate.

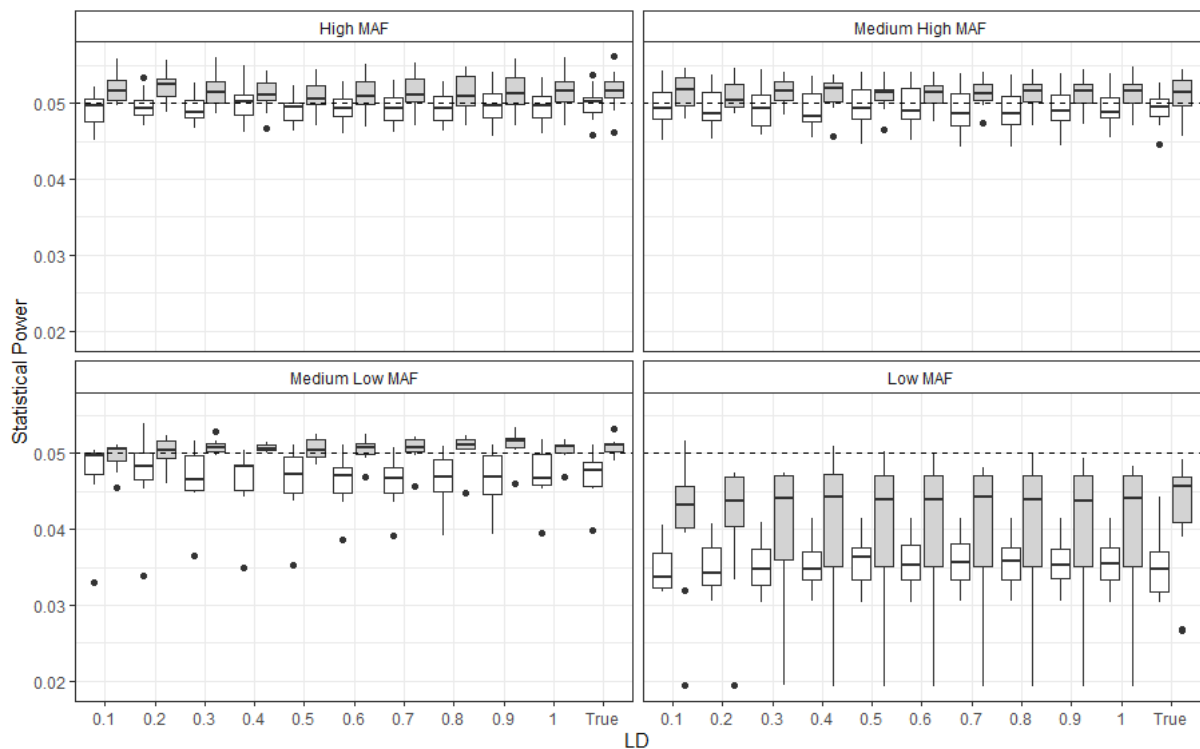


Figure 2: Boxplot of type I error rate of the G×E interaction (under the hypothesis of no G×E interaction) by LD threshold of the highest possible LD ranging from 0.1 to 1. “True” depicts the type I error rate of the G×E interaction using data of the actual, not imputed SNP genotype. Boxes with grey (white) fill illustrate simulation results with environmental exposure frequency of 0.3 (0.1). Data is stratified by MAF groups, which are defined in the “Methods” section. Dashed horizontal line marks 5% type I error rate.

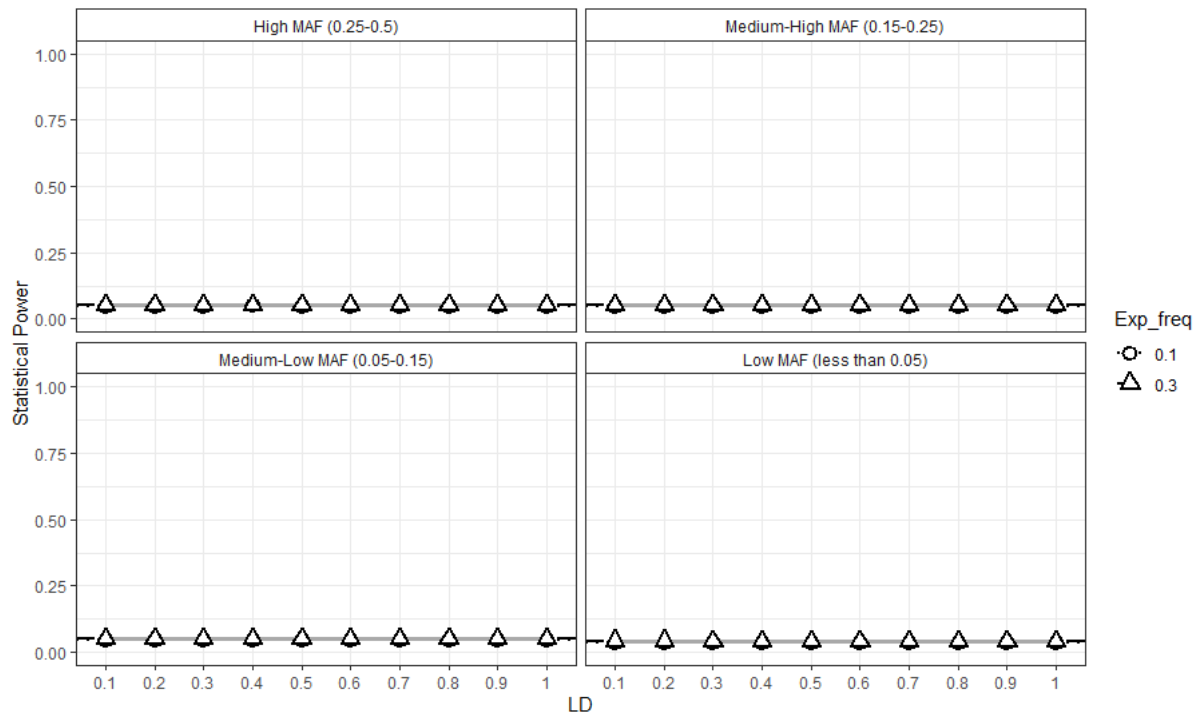


Figure 3: Mean type I error rate of G×E interaction (under the hypothesis of no G×E interaction) from imputed data by highest LD level of neighbouring SNP from 10 000 replications grouped by MAF. Circles depict means when the exposure frequency is equal to 0.1, triangles – 0.3. Dotted (dashed) line is the true statistical power of the genotyped, not imputed, SNPs for the 0.1 (0.3) exposure frequencies.

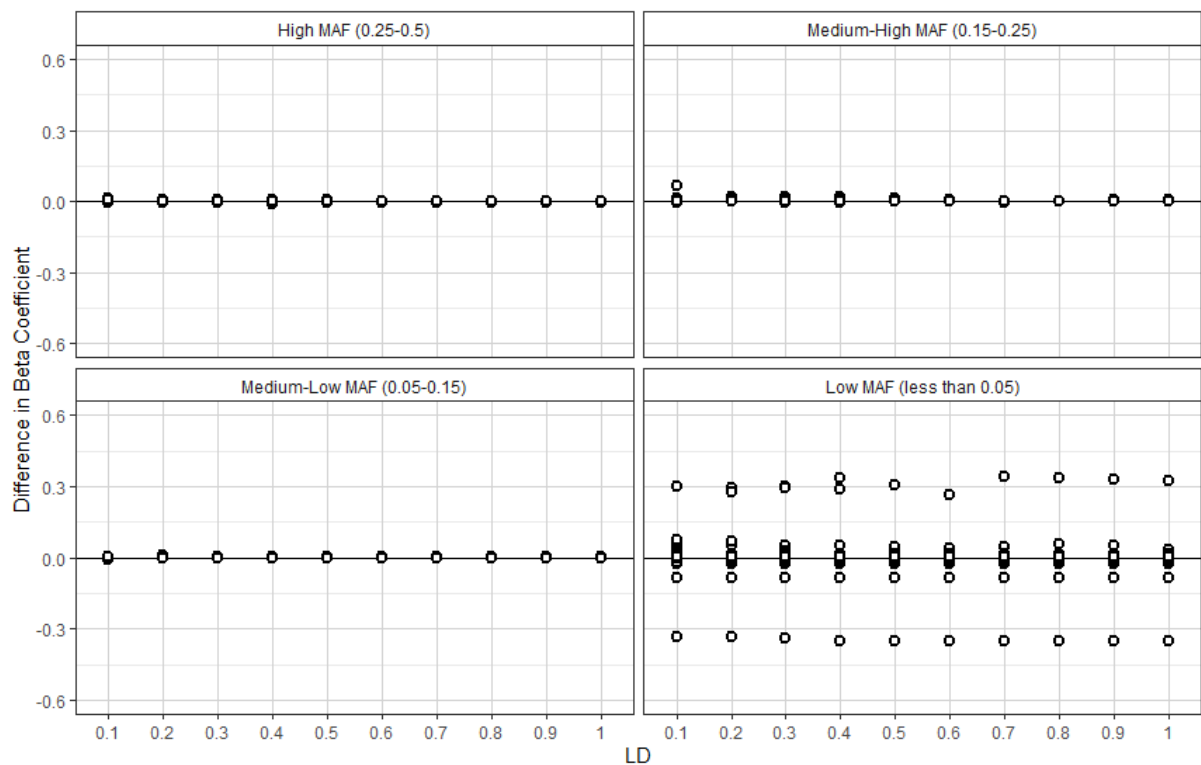


Figure 4: G×E interaction beta estimation bias (under the hypothesis of no G×E interaction) calculated as the difference between true beta and that calculated from the imputed data grouped by SNP MAF.

2.2 Case-Only Analysis of Gene-Gene Interactions in Inflammatory Bowel Disease

2.2.1 Summary

This paper focuses primarily on the practical implementation of the CO study design for analysing G×G interactions. The CO study design deemed to be attractive due to its high statistical power (in comparison to the CC study design, given the same number of cases) and no need for controls. However, issues such as LD, population stratification, cryptic relatedness, computational burden arising from many possible SNP interaction pairs as well as the multiple testing problem needed to be addressed. Moreover, the CO study design called for two assumptions to be fulfilled when analysing G×G interactions, namely, that (i) the disease of interest is sufficiently rare (i.e. has a prevalence of less than 5%) and (ii) that the two genes under study are uncorrelated in the general population. The (ii) assumption is problematic in the analysis of G×G interactions primarily due to LD, as two genes in close proximity to each other will have a higher correlation due to their high LD and thus not fulfil assumption (ii) for the CO only design to be valid. Thus, this paper presents a roadmap of methods needed for making a practical application of the CO study design for detecting G×G interactions possible. The developed methodology was exemplified on the largest available genotype dataset for IBD from the IIBDGC comprising of 16 636 CD and 12 888 UC cases from over 40 different centers.

In order to address the potential pitfalls of analysing G×G interactions in the CO study design, a short-list of SNPs was defined and principal components as covariates in the logistic regression and meta-analysis using random effects were performed. Focusing on SNPs with proven main effects seems plausible as they are more probable to exhibit G×G interaction effects. In the case of the diseases of interest, CD and UC, a short-list of 169 SNPs for CD and 156 SNPs for UC with genome-wide significance in the meta-analysis from Liu et al. (Liu et al. 2015) was considered. This significantly reduced the computational burden from over 6.6 billion possible pairs to 14 196 (12 090) for CD (UC). When more than one population is examined in a study, whether it is within the same study or from different centers, meta-analysis, in combination with PCA, is an appropriate way to address population stratification among centres (Price et al. 2006). Thus, in the IIBDGC dataset examined, PC were included on the center level and center-wise logistic regression analyses were combined by a random effects meta-analysis applying the Wald test. For multiple testing correction, a simulation approach in which the test-wise threshold was chosen in a two-step fashion was considered. The empirical null distribution (i.e. no interaction) of the meta-analysis Wald test statistic was estimated by a simulation approach. In this approach, first a number of SNPs, corresponding to the number used in the main analysis for each disease, was repeatedly drawn from the genome. Then, all possible interaction pairs were formed and the 5% quantile of the minimal p-values was determined as an alternative p-value threshold. However, this proved to be highly time-consuming and led to significance thresholds almost identical to those determined when applying usual Bonferroni correction based on all pairs of SNPs. This indicated that there was only very little residual association left if SNPs on different chromosomes and/or

chromosome arms were considered and center-wise PCs were included in the logistic regression models.

The independence assumption of the CO study design proved to be a challenge when investigating G×G interactions in the CO study design. Due to present LD structures, which could cause correlation in the general population, many SNP pairs would not fulfil the independence assumption. Thus, a few approaches were examined for checking the assumption: restriction of the G×G analysis to pairs of unlinked SNPs and the verification of the statistical independence of genotypes in a suitable reference database. The restriction of the G×G analysis to pairs of unlinked SNPs could be achieved by only considering SNPs on different chromosome arms. The empirical verification of independence was explored using either the given controls in the IIBDGC dataset or using the European super-population EUR of the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015) as a proxy for the controls. As seen in Figures 1a – 1c, there is almost a full overlap of SNP pairs that fulfil and violate the independence assumption when comparing the approaches of SNPs on different chromosome arms and verifying the independence using controls from the same dataset. Meanwhile, the situation differs if the EUR population from the 1000 Genomes Project is used as control proxy. This could be due to the fact that certain associations such as batch effects and dataset specific population stratification structures would not be reflected in the European dataset. Since using controls to validate the independence assumption would make one of the advantages of the CO study design, namely, no need for controls, redundant, the restriction of the G×G analysis to the easily applicable approach of SNPs on different chromosome arms seemed most promising.

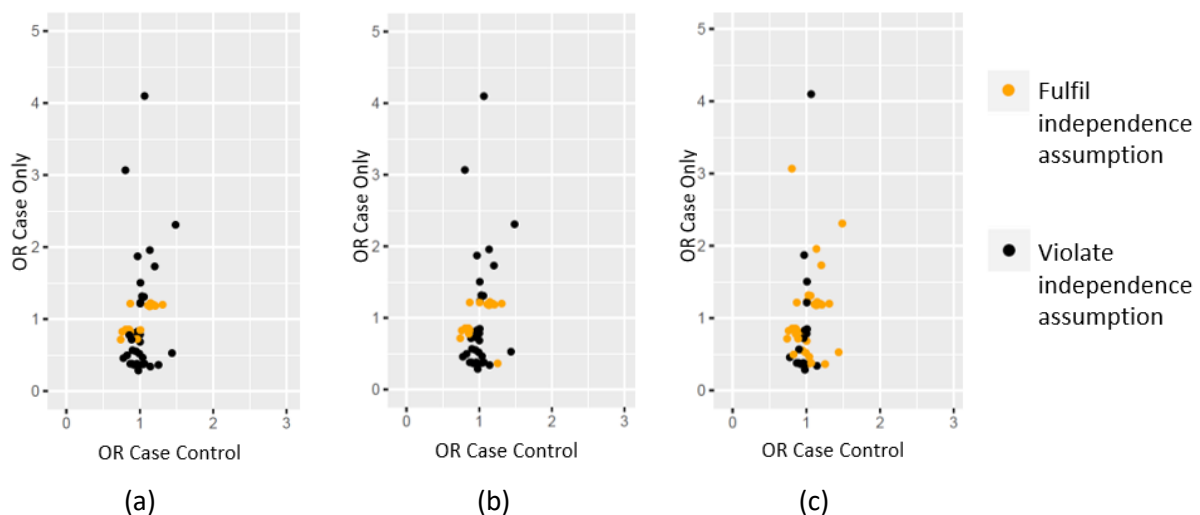


Figure 1: Independence assumption validation of SNP pairs with p -value of $< 5 \times 10^{-3}$ in the CO study design for CD. Three different approaches examined: (a) SNPs on different chromosome arms presumed to fulfil the independence assumption; (b) sample controls used as a reference base, (c) 1000 Genomes used as a reference base. Black points indicate violation, yellow – fulfilment of independence assumption.

After applying the mentioned aspects to the IBD dataset, a number of nominally significant ($p < 0.05$) G×G interactions were observed, yet none of these withstood Bonferroni multiple testing correction. One SNP pair stood out, namely rs26528 in the *IL27* gene and rs9297145 in

the *KPNA7* gene region. For CD, an interaction OR of 1.18 (95% CI: 1.10-1.27) and p-value of 7.75×10^{-6} was observed (significance threshold p-value was equal to 3.66×10^{-6}). This interaction could also be biologically plausible, as the *IL27* and *KPNA7* genes both play a role in NF- κ B signalling, a master regulator of pro- and anti-inflammatory processes in IBD.

In conclusion, in this paper I analysed the aspects that were necessary to consider in order to retain the validity and utility of the CO study design for analysing G×G interactions. An analysis method focusing on SNPs with known MEs on different chromosome arms, using PCs in the center-wise logistic regressions and meta analysing the results was established for future G×G interactions using the CO study design. Having applied these methods on the currently largest available IBD dataset, I had to conclude that such interactions are probably scarce for this specific disease of interest - IBD.

2.2.2 Publication

Case-only analysis of gene–gene interactions in inflammatory bowel disease

Milda Aleknonytė-Resch^a, Sandra Freitag-Wolf^a, The International Inflammatory Bowel Disease Genetics Consortium, Stefan Schreiber^b, Michael Krawczak^a and Astrid Dempfle^a

^aInstitute of Medical Informatics and Statistics, Kiel University, Kiel, Germany; ^bDepartment of Internal Medicine I, University Hospital Schleswig-Holstein, Kiel, Germany

ABSTRACT

Background: Gene–gene interactions ($G \times G$) potentially play a role in the etiology of complex human diseases, including inflammatory bowel disease (IBD), and may partially explain their ‘missing heritability’.

Methods: Using the largest genotype dataset available for IBD (16,636 Crohn’s disease (CD) and 12,888 ulcerative colitis (UC) cases) we analyzed $G \times G$ with the powerful case-only (CO) design. We studied 169 single nucleotide polymorphisms (SNPs) for CD (156 for UC), previously shown to be associated with the respective diseases. To ensure the validity of the CO design, we confined our analysis to pairs of unlinked SNPs. We used principal component analysis at the center level to adjust for possible causes of genotypic association other than $G \times G$, such as population stratification and genotyping batch effects. Results from center-wise logistic regression analyses were combined by a random effects meta-analysis.

Results: A number of nominally significant ($p < .05$) $G \times G$ interactions were observed, but none of these withstood the Bonferroni multiple testing correction. However, one SNP pair, comprising rs26528 in the *IL27* gene and rs9297145 in the *KPNA7* gene region was characterized by an interaction odds ratio of 1.18 (95% CI: 1.10–1.27) for CD and a p -value of 7.75×10^{-6} . Owing to the concurrent role of the *IL27* and *KPNA7* genes in NF- κ B signaling, a master regulator of pro- and anti-inflammatory processes in IBD, the observed interaction also has biological plausibility.

Conclusions: We were able to exemplify the utility of the CO design for analyzing $G \times G$, but had to recognize that such interactions are probably scarce for IBD.

ARTICLE HISTORY

Received 25 February 2020
Revised 25 June 2020
Accepted 26 June 2020

KEYWORDS

Epistasis; effect modification; Crohn’s disease; ulcerative colitis; gene–gene interaction

Introduction


Genome-wide association studies (GWAS) have been used extensively in the past to unravel the ‘genetic architecture’ of complex human diseases, i.e., to assess the population-wide level of statistical association between a disease of interest and the genotypes of single nucleotide polymorphisms (SNPs). This way, GWAS successfully identified potentially causal links between genes or gene regions and several medical conditions [1]. Early on, however, it became clear that the extent to which SNP associations can explain the heritability of complex human diseases is limited [2]. In addition to allelic and locus heterogeneity, particularly of rare variants, epistasis was suspected to hallmark the genetic etiology of most, if not all, of these traits [3,4].

Originally, the term ‘epistasis’ was used to refer to the ability of one or more genotypes of a gene, say A, to mask the phenotypic effects of another gene, B [5]. Over time, however, epistasis has become more or less synonymous of gene–gene interaction in general [6], where it is important to distinguish between biological and statistical interaction. The former is usually postulated when the gene products in

question share some common role in the disease etiology, i.e., if they either interact physically with one another or if they impede upon one and the same, disease-relevant biological pathway. Statistical interaction, on the other hand, is defined as the lack of additivity of the genotype-associated disease risk difference, measured on a particular scale (usually linear, log or logit). Notably, absence of statistical interaction on one scale implies the presence of interaction on all other scales, i.e., there is no such thing as a lack of statistical gene–gene interaction. Statistical interaction can also be interpreted as ‘effect modification’ in that the risk difference associated with a given genotype of gene A, scaled correspondingly (i.e., risk difference, relative risk, or odds ratio), depends upon the genotype of gene B. While certain types of biological interaction result in statistical interaction on a certain scale, the presence of statistical interaction does not necessarily imply the concurrent presence of any meaningful biological interaction [7].

In the following, we will deal with statistical gene–gene interaction ($G \times G$) between SNPs. These interactions pose a computational problem in practice, because the number of possible pairwise SNP combinations that need to be taken

CONTACT Astrid Dempfle  dempfle@medinfo.uni-kiel.de  Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany

 Supplemental data for this article can be accessed [here](#).

© 2020 Informa UK Limited, trading as Taylor & Francis Group

into consideration in a comprehensive search for $G \times G$ equals

$$\frac{n \times (n-1)}{2}$$

which is a quadratic function of the number n of SNPs under study. For example, the Illumina HumanOmni 2.5–8 chip used in many GWAS covers 2.5 million SNPs, which generates >3 trillion possible SNP pairs. Since statistical interaction is equivalent to effect modification, it may thus be advisable in searches for $G \times G$ initially to concentrate upon those SNPs that have a proven main effect and that are therefore more likely to exhibit $G \times G$ in the first place. Even though $G \times G$ without main effects is theoretically possible, to the best of our knowledge, it has never been demonstrated in reality in the context of a complex human disease. Moreover, statistical (usually generalized linear) models of genotype–phenotype relationships that include an interaction term, but no main effect terms, may still create substantial disease association of single SNPs in GWAS [8].

The case-only (CO) design is a powerful approach to detect pair-wise statistical interaction of disease risk factors, including $G \times G$. It has main advantages over the case-control (CC) design in that it obviates the need for proper controls and achieves greater statistical power with the same number of cases [9]. However, these advantages come at the price of requiring the validity of two assumptions, namely that the disease of interest is sufficiently rare (i.e., has prevalence $\leq 5\%$, say) and that the two risk factors under study are uncorrelated in the general population [10]. In gene–environment interaction studies, the second assumption is usually not critical because most environmental exposures are under minor, if any, genetic control. With $G \times G$, by contrast, linkage disequilibrium, population structure and cryptic relatedness all potentially induce pairwise genotype associations at the population level, which renders the practical utility of CO for $G \times G$ studies less straightforward. In addition, technical artifacts such as genotyping batch effects may also create spurious SNP–SNP associations among cases. Addressing such associations with the goal to retain the validity of the CO design as a means of $G \times G$ analysis thus appears well warranted.

Crohn's disease (CD) and ulcerative colitis (UC) are chronic inflammatory bowel diseases (IBD) that affect up to 0.3% and 0.5%, respectively, of the European population [11]. Clinically, CD may involve any part of the gastrointestinal tract. Bowel inflammation in CD is transmural and discontinuous, and may involve granulomas as well as intestinal or perianal fistulas. In UC, by contrast, inflammation is continuous, limited to rectal and colonic mucosal layers, and neither fistulas nor granulomas seem to occur [12]. In approximately 10% of IBD cases, no definitive classification as either CD or UC can be made, and considerable overlap has been noted for the two diseases in terms of their genetic risk loci, as identified by GWAS [13]. In any case, segregation analyses, twin studies and ethnic differences in familial risk have established both CD and UC as complex multifactorial disorders [14].

Previous studies of $G \times G$ in IBD were mostly confined to a few specific genomic regions, revealing weak interaction

between SNPs in the *IL12B* and *STAT4* genes for general IBD [15], between *JAK2* and *STAT3* [16], *ATG16L1* and *PTPN2* [17], *CARD8* and *NAPL3* [18] for CD, and between *ATG16L1* and *PTPN2* [17] and *JAK2* and *STAT3* [16] for UC. A more exhaustive search for $G \times G$ in IBD, undertaken by Zhang et al. [19], found nine weak interactions in the MHC region on chromosome 6. Notably, all of the above studies employed a CC design and, except for the Zhang et al. study [19], comprised less than 3000 individuals each.

In the present study, we examine $G \times G$ for IBD keeping in mind that the CO design provides increased statistical power to detect such interaction, compared to the CC design. In the course of this, we demonstrate how potential violations of the key CO design assumptions can be taken into consideration. Our study draws upon the largest currently available IBD dataset from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), not least because the same dataset was used successfully before to identify $G \times E$ interactions for CD and UC [20] and was also underlying the CC-based search for $G \times G$ by Zhang et al. [19]. As was noted above, $G \times G$ is not very likely in the absence of main effects so that it is plausible to focus searches for $G \times G$ for IBD on SNPs with a proven main effect on general IBD, CD or UC. Such a list was provided by Liu et al. [13] and, because it still represents the most up-to-date summary of genetic IBD associations available, it was used to guide our $G \times G$ analyses as well.

Materials and methods

Data

The data used in the present study originated from 29,524 IBD patients, i.e., patients diagnosed with either CD ($n = 16,636$) or UC ($n = 12,888$), and from 14,275 healthy controls. The data were provided to us by the IIBDGC, which comprises 48 centers in Europe, North America and Australia [21]. We only included individuals with full information available on disease status, center assignment and SNP genotypes (see below). Moreover, only centers that provided data on at least 10 patients were taken into consideration (46 centers for CD, 42 centers for UC, 48 centers for IBD). Some 19 centers also had genotype data available for controls. The final dataset widely overlapped with the data used by Yadav et al. [20] in their study of $G \times E$, and all additional data were subjected to the same quality control as applied there.

Genotyping with the Immuchip custom genotyping array (Illumina) was carried out in a varying number of batches at 11 different genotyping facilities, as described elsewhere [21]. In order to avoid zero counts of minor allele carriers, SNPs with a MAF < 0.05 were excluded, as were those with a genotyping rate ≤ 0.9 , leaving 115,123 SNPs for further analysis. For reasons alluded to above, we focused upon SNPs with a known main effect instead of undertaking a genome-wide search for $G \times G$. Thus, we confined the study to the 169 SNPs with MAF ≥ 0.05 and a nominally significant genotype–phenotype association ($p < 0.05$) as listed for CD (156 for UC, 201 for IBD) by Liu et al. [13]. This policy implied that 14,196 (CD), 12,090 (UC) and 20,100 (IBD)

potential interaction pairs, respectively, had to be taken into consideration in subsequent analyses. Owing to the clinical differences between CD and UC, we focused upon separate analyses of the two diseases throughout but, in an exploratory add-on, also investigated them jointly as IBD.

Statistical analysis

All statistical analyses were performed with either R (v. 3.5.0) or PLINK [22], as appropriate. For statistical modelling, genotypes (G) were encoded assuming a dominant effect of the minor allele, i.e., $G=1$ for homozygous or heterozygous carriers of the minor allele, $G=0$ for homozygous carriers of the major allele. A dominant model was chosen here, not only because of its wide coverage of biologically plausible genetic etiologies, but also because it provides greater statistical power for low minor allele frequencies in $G \times G$ interaction analyses, where numerous zero counts can be expected for at least one of the two homozygous genotypes.

In the standard CC design of $G \times G$, disease status D is treated as the response variable whereas genotypes G_1 and G_2 are treated as predictor variables, alongside an interaction effect G_1G_2 , i.e.,

$$\text{logit}\{P(D=1)\} = \theta_0 + \theta_1G_1 + \theta_2G_2 + \theta_3G_1G_2$$

In the CO design, the genotype of the first SNP (G_1) is treated as a predictor variable, with respective regression coefficient δ_1 representing the interaction effect, whereas the other genotype (G_2) is treated as the response variable, i.e.

$$\text{logit}\{P(G_2=1)\} = \delta_0 + \delta_1G_1$$

Following Piegorsch et al. [10], no classical confounders such as age or sex were included in the CO model because their main effects on D cannot sensibly be modelled when using a CO design. Moreover, according to the EMBL-EBI GWAS catalogue (<https://www.ebi.ac.uk/gwas/>), none of the candidate SNPs of interest in the present study is known to be associated with age, sex or smoking.

To allow for the heterogeneous origin of the IIBDGC data, $G \times G$ was first analyzed separately for each contributing center, yielding center-specific estimates of regression coefficient δ_1 . Then, a meta-analysis fitting a random effects model with inverse variance weights was carried out using R package metafor [23], followed by a Wald test to assess whether the average δ_1 , taken over centers, was significantly different from zero.

Spurious pair-wise correlation may arise between SNP genotypes when the population under study comprises distinct subpopulations with different allele frequencies [24]. Genotyping batch effects can create similar artefacts. To address this problem, we carried out principal component analysis (PCA) of cases at the IIBDGC study center level, because the top principal components (PCs) of SNP genotypes have been shown to adjust well, at least for population stratification, in previous genetic studies [25]. Our PCA comprised all 115,123 SNPs that passed quality control (see above). The top 10 PCs obtained were included in the logistic $G \times G$ model for large centers (>300 cases), five were

included for medium centers (100 to 300 cases) and three for small centers (<100 cases).

For the CO design to be valid, G_1 and G_2 have to be uncorrelated in the general population [10]. To some extent, this requirement can be fulfilled in two complementary ways, namely by

- i. restriction of the $G \times G$ analysis to pairs of unlinked SNPs, i.e., SNPs on different chromosomes or chromosome arms, or
- ii. verification of the statistical independence of genotypes in a suitable reference database.

The first approach ensures that there is at least no linkage disequilibrium between SNPs. Other causes of pairwise SNP genotype association, such as population stratification, cryptic relatedness or genotyping batch effects, would not be addressed this way but would have to be eliminated by the inclusion of a sufficiently large number of PCs in the statistical model. For exploring the second approach (i.e., empirical verification of independence), we used the European super-population EUR of the 1000 Genomes Project [26] as a proxy for our controls. The respective SNP genotype data were analyzed in exactly the same way as those of the cases from the IIBDGC (i.e., logistic regression model with PCA, random effects meta-analysis of the six EUR subpopulations, Wald test). A significance level of 0.05 was adopted in order to balance the false positive rate of the original $G \times G$ analysis against its efficacy. In principle, the validity of the independence assumption could have been required in the IIBDGC control data as well. However, even although we analyzed these data in the same way as the EUR data out of scientific interest, we abstained from considering the ensuing results as another filtering option for two reasons. First, confining subsequent $G \times G$ analyses to centers that provided both cases and controls would have reduced the number of cases from 16,636 to 8,776 for CD and from 12,888 to 6,327 for UC. Second, in reality, own controls would rarely be available to researchers adopting a CO design because this would eliminate the main advantage of the latter, namely that CO obviates the need for controls.

Two different approaches were followed to assess the statistical significance of pairwise SNP associations in the cases while limiting the family-wise error rate (FWER) of the respective association tests by 5%. First, Bonferroni correction was applied, dividing the conventional p value threshold of 0.05 by the number of SNP pairs considered for each disease in order to obtain a test-wise threshold. However, Bonferroni correction is overly conservative if many null hypotheses are incorrect or if the tests in question are not statistically independent. Therefore, we also adopted a supposedly less conservative approach in which the test-wise threshold was chosen in a two step fashion, drawing upon the empirical null distribution (i.e., no interaction) of the meta-analysis Wald test p values. To this end, 2,000 SNP sets of the same size (i.e., 169 for CD, 156 for UC) and genomic distribution (i.e., representation of chromosomal arms) as the two original SNP sets [13] were drawn at random from the 115,123 SNPs

that passed quality control. The random SNPs were also MAF-matched to the original SNPs within $\pm 3\%$, curtailed by the minimum and maximum MAF of the original SNPs. Next, $G \times G$ analyses were carried out for the random SNP sets in the same way as for the two original SNP sets and the smallest Wald test p value was recorded for each set. If all pairwise genotype associations in a random set of SNPs are spurious, at least one of the associations would be deemed statistically significant by reference to a p value threshold if, and only if, the minimum p value among all pairs falls below this threshold. Hence, the 5% quantile of the 2000 minimal p values was adopted as an alternative p value threshold for the original analyses.

Statistical power

The statistical power to detect $G \times G$ with a CO design was calculated using the Quanto software [27]. Quanto supports CO studies of $G \times G$, but was not specifically designed for this purpose. Therefore, the program specifications also include some parameters irrelevant for the CO design, such as disease prevalence and main effect ORs. The power calculations were carried out separately for CD and UC, adopting different sample sizes taken from the IBDGC dataset (16,636 CD cases, 12,888 UC cases) as well as published prevalence figures [11] (0.32% for CD, 0.51% for UC). The dominant main effect of the minor allele of a SNP was equated to the median OR (1.10 for CD, 1.08 for UC) obtained for the SNPs reported by Liu et al. [13]. The MAF was set to either 0.1 or 0.4 (i.e., to one end of a realistic range of MAFs), thereby creating three possible MAF combinations per SNP pair. Finally, a Bonferroni-corrected significance level of 3.66×10^{-6} and 4.32×10^{-6} was adopted for CD and UC, respectively (see below). The interaction OR was varied between 1.00 and 1.50 in steps of 0.05, covering a realistic range of possible effect sizes at sufficient resolution to be able to gauge the achievable statistical power. To compare the CO and CC design, Quanto was also used with the above settings to calculate the power of the CC design, assuming that equal numbers of cases and controls were available for analysis.

Results

Validity of the CO design for $G \times G$ analysis

One of the key prerequisites for the validity of the CO design for $G \times G$ analysis is the independence of SNP genotypes at the population level. There are various ways to ensure that this condition is fulfilled in practice. Thus, a CO study of $G \times G$ can exclude pairs of SNPs located on the same chromosome arm (criterion A), the validity of the independence assumption can be assessed in a suitable reference database such as, for example, 1000 Genomes ([26]; criterion B1), or own controls can be used to this end, if available (criterion B2).

In the present study, the proportion of SNP pairs that met one or the other non-independence criterion varied slightly

Table 1. Pairs of IBD-associated SNPs meeting different criteria for genotypic non-independence.

Disease	Criterion	N (%)
CD	A	519 (3.7)
	B1	473 (3.3)
	B2	622 (4.4)
UC	A	528 (4.4)
	B1	445 (3.7)
	B2	548 (4.5)

CD: Crohn's disease; UC: ulcerative colitis; N: number of SNP pairs; %: percentage of all possible SNP pairs (14,196 for CD, 12,090 for UC); A: location on the same chromosome arm; B1: nominally significant ($p \leq 0.05$) association in 1000 Genomes super-population EUR; B2: same as B1, but in IIBDGC controls.

between criteria and disease entities, ranging from 3.3% for B1 in CD to 4.5% for B2 in UC (Table 1). The vast majority of SNP pairs involved markers located on different chromosomes or chromosome arms (13,677 for CD; 11,562 for UC), and only a small proportion of these showed nominally significant association in the 1000 Genomes super-population EUR (429 pairs, or 3.2%, for CD; 417 pairs, or 3.6%, for UC). Interestingly, these results also imply that the proportion of SNP pairs that met criterion B1, but not criterion A, was as large as 90.7% for CD (429 of 473 pairs) and 93.7% (417 of 445 pairs) for UC. In other words, >90% of the nominally significant associations found in the 1000 Genomes data pertained to SNPs on different chromosomes or chromosome arms, and were therefore likely due to chance or technical or methodological shortcomings. Spurious associations are likely sample-specific and, hence, difficult to account for by screening a reference database such as the 1000 Genomes. In fact, using IIBDGC controls yielded similar proportions of significant associations between SNPs on different chromosomes or chromosome arms (578 of 622, or 92.9%, for CD; 492 of 548, or 89.8%, for UC) as the 1000 Genomes but the overlap with pairs meeting criterion B1 was low (27 of 622, or 4.4%, for CD; 28 of 548, or 5.1%, for UC). This means that only ~5% of the spurious associations in the IIBDGC controls would have been detected by reference to the 1000 Genomes data. In view of this, and because own controls are usually lacking in studies adopting a CO design, we chose to confine filtering of candidate SNP pairs to criterion A, leaving 13,677 and 11,562 pairs for further analysis for CD and UC, respectively.

Control of the family-wise error rate (FWER)

Bonferroni correction for the number of SNP pairs considered in our study (CD: 13,677, UC: 11,562) yielded p value thresholds of 3.66×10^{-6} and 4.32×10^{-6} for CD and UC, respectively. Since exhaustive pair formation from a given set of SNPs implies considerable overlap between pairs, Bonferroni correction may be overly conservative. This is because the Bonferroni method controls the FWER under the 'worst case' assumptions that all null hypotheses are correct and that the corresponding test statistics are independent. Therefore, we developed a simulation-based approach to take the overlap between SNP pairs and the potentially resulting lack of

independence into account (see Methods). For each disease, a corresponding number of SNPs was repeatedly drawn from the genome, all possible pairs were formed and the 5% quantile of the minimal p values was determined as an alternative p value threshold. However, the results of this approach (3.95×10^{-6} for CD, 4.95×10^{-6} for UC) were surprisingly close to the Bonferroni thresholds, indicating that only little residual association between SNPs on different chromosomes or chromosome arms was present in the IBDGC cases that had not been accounted for by the inclusion of PCs in the logistic regression models.

Statistical power

Given the same number of cases, the CO design was found to provide greater statistical power to detect $G \times G$ than the CC design under all parameter settings considered (Figure 1(a,b)). However, even with sample sizes like those commanded by the IBDGC (i.e., >16,000 cases for CD, >12,000 cases for UC), $\geq 80\%$ power is achieved upon Bonferroni correction (FWER < 0.05) only for interaction ORs exceeding between 1.2 and 1.35, depending upon MAF and disease. As will be shown in the following section, even the top $G \times G$ interactions unraveled here for the two diseases did not reach this level.

$G \times G$ interaction for IBD

None of the SNP pairs analyzed yielded a meta-analysis Wald test p value below the Bonferroni threshold (i.e., 3.66×10^{-6} for CD, 4.32×10^{-6} for UC) or the 5th percentile of the empirical minimum p value distribution (3.95×10^{-6} for CD, 4.95×10^{-6} for UC). Only 12 SNP pairs each for CD (Table 2) and UC (Table 3) yielded a nominal p value smaller than 0.001 (for a comprehensive list of interactions with $p < .005$, see Supplementary Tables 1(a) for CD, 1b for UC, and 1c for IBD combined). Moreover, the OR distributions among tested and simulated SNP pairs were found to be almost identical, thereby corroborating the impression that prominent $G \times G$, if any, is a rare phenomenon for both CD and UC (Supplementary Figure 1).

Nevertheless, one SNP pair clearly stood out in our analyses, comprising rs26528 in the *IL27* gene and rs9297145 in the *KPNA7* gene region, which was characterized by an interaction OR of 1.18 for CD. The corresponding p value of 7.75×10^{-6} was close to the Bonferroni threshold and marked the 8.25th percentile of the empirical minimal p value distribution. At the study center level, the interaction ORs for this pair ranged from 0.57 to 3.45 (see Supplementary Figure 2), and none of the ORs was found to be nominally significantly smaller than unity. Moreover, no significant heterogeneity between centers was observed (Cochrane's $Q = 38.67$, $p = .66$). The interaction of the two SNPs on CD risk was also nominally significant, and showed the same direction and magnitude, in a CC analysis (Supplementary Figure 3). Closer inspection of the actual genotype counts (Table 4) revealed that the genotypes of the two SNPs were clearly correlated in cases, but not in controls. Thus, while

homozygotes for both major alleles (AA/AA) and carriers of at least one minor allele of both SNPs (NG/NC) were over-presented among cases, the genotype counts observed among controls well matched their expected values, given statistical independence. The data also corroborated the appropriateness, for both SNPs, of a dominant model of the genotype-phenotype relationship in CD in that the crude disease ORs calculated from Table 4 equaled 1.39 (GG) and 1.15 (AG) for rs26528, and 1.15 (CC) and 1.10 (AC) for rs9297145. We also explored the *IL27* and *KPNA7* gene regions in more detail by examining $G \times G$ for all genotyped SNPs located within 30 kb upstream or downstream of the two original SNPs (68 for *IL27*, 5 for *KPNA7*). Of the 340 SNP pairs analyzed (Supplementary Table 2), however, only one (rs153109, rs9297145) yielded a slightly smaller p value (6.06×10^{-6}) than the original SNP pair.

For UC, the smallest interaction p value obtained equaled 1.43×10^{-4} (Supplementary Table 1b), which clearly missed the Bonferroni threshold and marked the 72nd percentile of the corresponding empirical distribution. Similarly, for IBD combined, the smallest interaction p value was 6.90×10^{-5} (Supplementary Table 1c) whereas the Bonferroni threshold equaled 2.70×10^{-6} . Finally, when the respective 519 (CD) and 528 (UC) pairs of SNPs located on the same chromosome arm were analyzed using a CC design, 20 (CD; 3.9%) and 25 (UC; 4.7%) showed nominally significant $G \times G$, but all p values clearly missed the Bonferroni thresholds of 9.63×10^{-5} (CD) and 9.47×10^{-5} (UC), respectively (Supplementary Tables 3a and 3b).

Discussion

Comprehensive analysis of $G \times G$ interaction potentially provides deeper insights into the genetic architecture of complex diseases because statistical interaction may point, directly or indirectly, towards biological interaction. Moreover, assessing the true magnitude of $G \times G$ is also an important step towards explaining some of the 'missing heritability' of these traits [2] because past estimates of total heritability probably have been significantly inflated, for example, by unaccounted $G \times G$ interaction [28]. Classically, $G \times G$ analysis is carried out using a CC design to determine whether the scaled risk difference between the genotypes of one variant depends upon the genotype of the other variant. However, when carried out at genome-wide level, covering millions of SNPs at a time, CC studies of $G \times G$ raise several methodological issues, including heterogeneous data quality and lack of statistical power. To some extent, adoption of a CO design may help to mitigate these problems. The CO design is not only more efficient than the CC design (i.e., it provides greater power with the same number of cases) but, by obviating the need for own controls, CO also avoids systematic differences in data quality due to, for example, differential DNA sampling, phenotyping or genotyping of cases and controls.

Although the CO design provides some important advantages over the CC design, it requires fulfilment of two key conditions for it to be statistically valid, namely that the

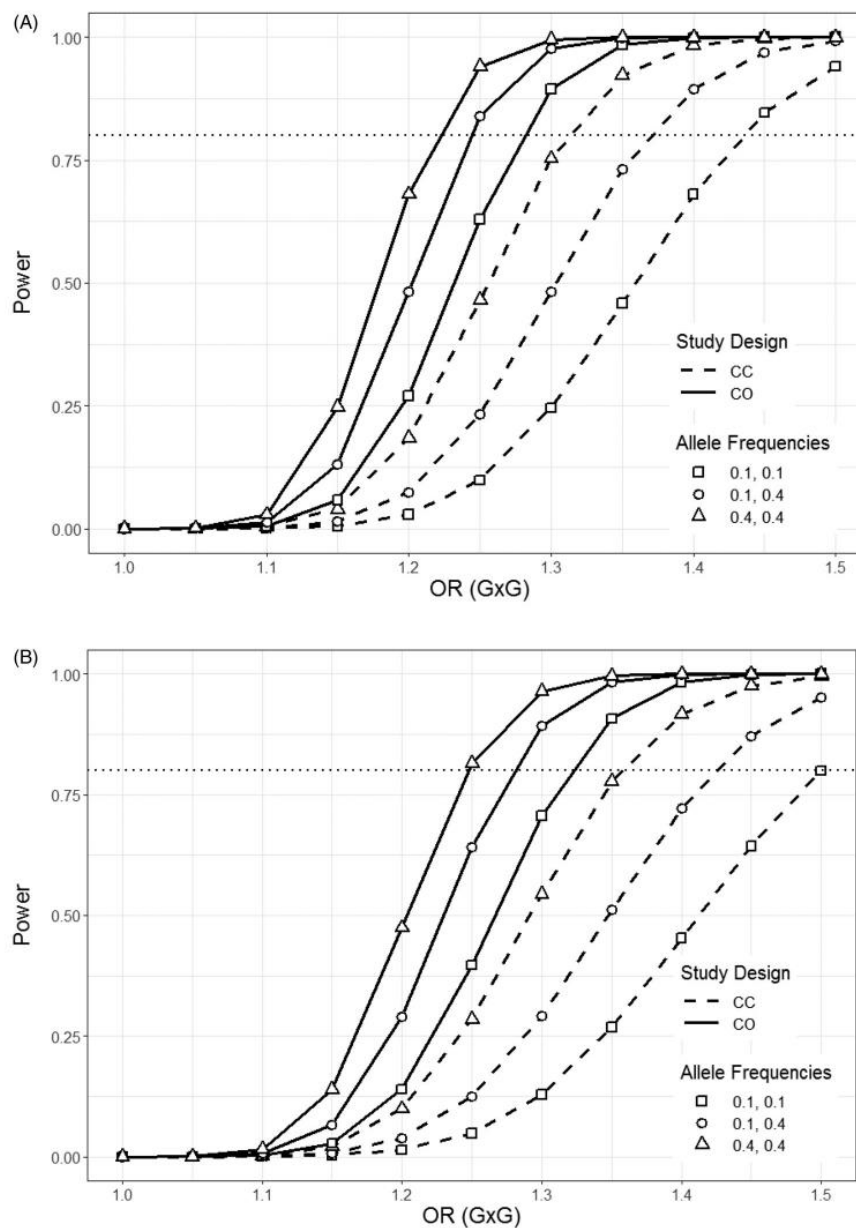


Figure 1. Power of the CO (solid lines) and CC (dashed lines) designs of $G \times G$ analysis (calculated using the Quanto software with parameter settings apt for (a) CD, Bonferroni-corrected significance level: 3.66×10^{-6} ; (b) UC; Bonferroni-corrected significance level: 4.32×10^{-6}). Pairs of SNP allele frequencies are marked by symbols: 0.1, 0.1 (square); 0.1, 0.4 (circle); 0.4, 0.4 (triangle). Dotted horizontal line marks 80% statistical power.

disease of interest is rare (i.e., has prevalence $<5\%$, say) and that the risk factors under study are uncorrelated in the general population [10]. These assumptions are usually met in studies of $G \times E$ interaction, but for $G \times G$, they may represent a major challenge due to the possibility of linkage disequilibrium, population stratification, cryptic relatedness and genotyping batch effects, among other factors. Here, we investigated two ways in which the independence requirement can be dealt with in $G \times G$ interaction studies, namely by restricting the analysis to pairs of genetic variants on different chromosome arms or by the assessment of a suitable reference database, such as the 1000 Genomes, or of own

controls. We examined the efficacy of both methods in a large set of SNP genotype data from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), comprising 16,636 CD cases and 12,888 UC cases (i.e., 29,524 IBD cases in total). As it turned out, after inclusion of a sensible number of top principal components in the underlying logistic model, only few, likely spurious, associations between SNPs on different chromosomes or chromosome arms were observed in the 1000 Genomes data super-population EUR or the IIBDGC controls. Since these associations showed little overlap between the two databases, and because own controls are usually not available in CO studies in the first place,

Table 2. Top G × G interactions ($p < 0.001$) between SNPs associated with CD.

SNP 1				SNP 2				Interaction		
ID	Gene	MAF	OR ^a	ID	Gene	MAF	OR ^a	OR ^b	SE	p
rs26528	<i>IL27</i>	0.47	1.13	rs9297145	<i>KPNA7</i>	0.28	1.08	1.18	0.04	7.75×10^{-6}
rs3740415	<i>MFSD13A</i>	0.45	0.95	rs12199775		0.06	0.89	0.81	0.05	7.19×10^{-5}
rs12946510	<i>GRB7, IKZF3</i>	0.49	1.13	rs17057051	<i>PTK2B</i>	0.30	0.94	0.87	0.04	1.19×10^{-4}
rs194749	<i>ZFP36L1</i>	0.24	1.08	rs314313	<i>EPHB4</i>	0.32	1.07	0.88	0.03	1.28×10^{-4}
rs11010067		0.39	1.14	rs7554511	<i>C1orf106</i>	0.24	0.86	0.88	0.04	3.88×10^{-4}
rs13407913	<i>ADCY3</i>	0.46	1.12	rs1748195	<i>DOCK7</i>	0.34	1.07	1.13	0.04	3.91×10^{-4}
rs3091315	<i>CCL2, CCL7</i>	0.25	0.87	rs921720	<i>LOC105375746</i>	0.36	0.89	0.89	0.03	4.86×10^{-4}
rs17057051	<i>PTK2B</i>	0.30	0.94	rs314313	<i>EPHB4</i>	0.32	1.07	1.12	0.03	5.03×10^{-4}
rs4768236	<i>LRRK2</i>	0.37	1.12	rs7773324	<i>DUSP22, IRF4</i>	0.38	0.92	1.12	0.03	5.75×10^{-4}
rs2945412	<i>KSR1</i>	0.38	0.89	rs568617	<i>FIBP</i>	0.20	1.09	0.89	0.03	8.44×10^{-4}
rs3740415	<i>MFSD13A</i>	0.45	0.95	rs11583043	<i>DPH5</i>	0.28	1.03	1.13	0.04	8.96×10^{-4}
rs259964	<i>ZNF831</i>	0.47	1.07	rs10065637	<i>ANKRD55</i>	0.20	0.91	1.13	0.04	9.01×10^{-4}

ID: SNP rs-number; MAF: minor allele frequency; OR: odds ratio; SE: standard error of interaction odds ratio; P: nominal p value of meta-analysis Wald test.

^aAllelic disease odds ratio of minor allele, according to Liu et al. [13].

^bInteraction odds ratio.

Table 3. Top G × G interactions ($p < 0.001$) between SNPs associated with UC (for details, see legend to Table 2).

SNP 1				SNP 2				Interaction		
ID	Gene	MAF	OR ^a	ID	Gene	MAF	OR ^a	OR ^b	SE	p
rs4722672		0.20	1.08	rs12568930	<i>ZBTB40</i>	0.15	0.88	1.18	0.04	1.43×10^{-4}
rs11742570		0.37	0.92	rs6716753	<i>SP140</i>	0.19	1.04	1.16	0.04	1.76×10^{-4}
rs7805114		0.40	0.89	rs10495903	<i>THADA</i>	0.14	1.07	0.82	0.05	2.42×10^{-4}
rs3764147	<i>LACC1</i>	0.24	1.04	rs653178	<i>ATXN2</i>	0.49	1.05	0.86	0.04	3.74×10^{-4}
rs17085007		0.20	1.14	rs3766606	<i>PARK7</i>	0.14	0.87	1.16	0.04	4.60×10^{-4}
rs3764147	<i>LACC1</i>	0.24	1.04	rs3740415	<i>MFSD13A</i>	0.45	0.94	0.87	0.04	7.00×10^{-4}
rs11083840	<i>PTGIR</i>	0.42	1.07	rs10185424		0.48	1.10	1.16	0.04	7.25×10^{-4}
rs2024092	<i>SBNO2</i>	0.22	1.05	rs28374715	<i>CHP1</i>	0.25	0.92	1.14	0.04	7.47×10^{-4}
rs6908425	<i>CDKAL1</i>	0.20	0.93	rs2651244		0.38	0.94	1.14	0.04	7.69×10^{-4}
rs13277237	<i>CCDC26</i>	0.45	1.06	rs2538470		0.38	1.07	1.15	0.04	8.08×10^{-4}
rs516246	<i>FUT2</i>	0.48	1.04	rs3742130	<i>GPR18</i>	0.21	0.93	0.87	0.04	8.67×10^{-4}
rs12946510	<i>GRB7, IKZF3</i>	0.49	1.14	rs4845604	<i>RORC</i>	0.13	0.85	0.85	0.05	9.79×10^{-4}

Table 4. G × G interaction on CD risk of SNPs rs26528 and rs9297145.

Group	rs9297145	rs26528			Total
		AA	AG	GG	
Cases	AA	2,465 (2,357.7)	1,690 (1,797.5)	353 (352.9)	4,508
	AC	4,168 (4,222.3)	3,277 (3,218.9)	628 (631.8)	8,073
	CC	1,834 (1,887.0)	1,488 (1,438.6)	286 (282.4)	3,608
	Total	8,467	6,455	1,267	16,189
Controls	AA	1,445 (1,428.1)	929 (949.0)	157 (153.9)	2,531
	AC	2,305 (2,318.9)	1,561 (1,540.4)	242 (249.7)	4,108
	CC	993 (997.0)	662 (662.6)	112 (107.4)	1,767
	Total	4,743	3,152	511	8,406

Given are the observed genotype counts in all centers combined (cases) or in centers that provided control data (controls). Values in parentheses are the genotype counts expected if SNP genotypes were uncorrelated within each group of individuals.

we concluded that location on different chromosome arms would have to suffice as a criterion for genotypic independence in practice, at least for genotype resources of similar provenience and quality as the IIBDGC data. Of course, pursuing such an approach implies that interactions within one and the same genomic region, such as those reported by Zhang et al. [19] for the MHC, would be systematically overlooked.

One major obstacle to G × G interaction studies aspiring to the control of the FWER is the quadratic relationship between the number of markers included in the study and the accruing number of interaction tests. This problem, which applies equally to the CO and the CC design, becomes virtually intractable when G × G is analyzed in a genome-wide fashion. We surmise that, under realistic assumptions

about the size of the underlying interaction ORs, sufficiently powered samples to sustain valid multiple testing correction are currently lacking for all complex diseases in humans. On the other hand, using Bonferroni correction to control the FWER may seem overly conservative in studies of pair-wise interactions owing to the inherent connection between candidate pairs and, hence, test statistics. However, as we found out in our CO-based study of G × G in IBD, the validity of such vindications cannot be taken for granted. Once we thoroughly excluded spurious pair-wise associations between SNPs, the FWER could not be controlled in any less conservative way than with Bonferroni. In other words, if there are no systematic causes of association left, such as LD, population stratification etc., then knowing the pair-wise genotypic correlations between SNPs A and B, and B and C, does not seem to predetermine the genotypic correlation between A and C.

Our investigation of the statistical power of the IIBDGC data confirmed the general notion that, with a given number of cases, the same statistical power is achievable for lower interaction ORs with a CO than with a CC design. Moreover, the IIBDGC data, which clearly represented the largest resource of SNP genotypes available for CD and UC, afforded the CO design enough statistical power (80%) after Bonferroni correction ($\text{FWER} < 0.05$) to detect even moderate interaction ORs of 1.2 to 1.3. However, not a single such interaction was detected between the candidate SNPs chosen for analysis here, neither for CD nor for UC. This

negative result suggests that $G \times G$ interactions of said magnitude are rare or even lacking for IBD, at least as far as SNPs with known main effects are concerned. We surmise that this conclusion is valid despite the fact that, in multicenter studies of $G \times G$, statistical power may be reduced by the heterogeneity of genetic background and environment (e.g., lifestyle, diet, smoking behavior), among others. We tried to minimize this potential for power loss by excluding small centers with less than 10 cases, and by choosing meta-analysis as the most efficient means to ensure statistical validity of our analyses. This notwithstanding, the negative findings reported here do not rule out the existence of $G \times G$ between, or with, other SNPs not yet included but, as was pointed out above, such interaction (i) is not very plausible to start with, (ii) has few, if any, real-world precedents and (iii) would be difficult to interpret biologically anyway. Of course, moderate or even strong $G \times G$ interaction is also conceivable between rare functional variants that are only tagged by the IBD-associated SNPs studied here but, as would be the case with smaller $G \times G$ effects between these SNPs themselves, such interaction would require even larger sample sizes to be detectable than were available here.

Although not formally statistically significant, one interesting interaction nevertheless emerged in our study for CD, namely between SNPs from the *IL27* (rs26528) and *KPNA7* (rs9297145) gene regions. The interaction OR of these two SNPs equaled 1.18 and was thus close to the detection limit suggested for the IIBDGC data by our own power investigations (see above). Moreover, the corresponding p value (7.75×10^{-6}) fell just short of the Bonferroni threshold and was an order of magnitude smaller than all other p values. Notably, the interaction ORs obtained for the two SNPs in a CC analysis had the same direction and magnitude as those from the CO analysis.

The *IL27* gene on human chromosome 16, encoding Interleukin 27, is widely known to play an important role in IBD, and there is even evidence for an *IL27* therapy being effective in IBD [29]. Interestingly, SNP rs26528 is also located in an enhancer (GH16J028503) that has seven target genes, including *IL27*. The other interacting SNP, rs9297145, is located in an intron of the *KPNA7* gene on chromosome 7. This gene encodes Karyopherin Subunit Alpha 7, a member of the importin alpha family involved in nuclear protein import. As was highlighted by Pumroy and Cingolani [30], importin expression is an important regulator of NF- κ B signaling, disruption of which is associated with various human conditions, including neurological and cardiovascular diseases. Most importantly, however, nucleocytoplasmic shuttling of NF- κ B plays a key role in IBD pathophysiology [31]. In our current understanding, IBD is characterized by an overwhelming expression of pro-inflammatory cytokines that are expressed in signaling pathways in the gut that are controlled by NF- κ B as a master regulator of gene transcription. In particular, NF- κ B is targeting the *IL27* gene promoter. Therefore, one plausible biological explanation of the statistical interaction observed here between SNPs rs26528 and rs9297145 could be genotype-dependent mutual tuning of the efficiency by which GH16J028503 and the NF- κ B

pathway control *IL27* gene expression. This interpretation would raise interest in *IL27* both as a biomarker and as a therapeutic target.

As was mentioned earlier, the interpretation of a statistical interaction depends upon the model used to describe the relationship between response and predictor variables. For example, if there is no interaction on the log scale then interaction necessarily exists on the linear scale, and vice versa. When investigating a particular disease, the question therefore arises as to whether its etiology comprises causal factors with predominantly additive or multiplicative effects. The answer, of course, depends upon the biological mechanism of pathogenesis, for which various models have been discussed in the literature [32,33]. In the so-called 'single hit' model, the disease in question is assumed to have several possible causes, each of which can be 'triggered' by only one risk factor. An example of this is provided by the dysfunction of independent, non-redundant components of a biochemical network. If the probability of triggering malfunction is sufficiently small for each individual risk factor, then the risk differences are approximately additive, i.e., the risk factors would not interact in a linear model, but would do so in a log-linear (or multiplicative) model. On the other hand, carcinogenic effects are commonly described by multistage models [34] based upon the assumption that any risk factor can independently cause a change of state towards disease (e.g., by heterozygosity for inactivating mutations in tumor suppressor genes). The disease then only occurs at the end of the causal chain, all stages of which must be traversed, so that the relative risks multiply and no interaction would exist in a log-linear model. The same holds for the so-called 'no-hit' model in which the non-acquisition of at least one of many possible protective factors triggers the disease.

In the present study, the effects of SNP genotypes on IBD risk were quantified by odds ratios, i.e., the underlying risk differences were measured on a logit scale. Since IBD is comparatively rare, odds ratios approximate to relative risks so that the arguments made above for multiplicative risk models apply to our logit-based analysis as well. In fact, impediment of the functions of *IL27* and NF- κ B seems to comply more with a 'single hit' than with a multistage model of causality in IBD which means that non-additivity of the corresponding genotypic risks, as observed here, appears plausible on a logit rather than a linear scale. On the other hand, the lack of additional substantial $G \times G$ on the logit scale suggests that the vast majority of other genetic risk factors affect IBD etiology via multistage or 'no-hit' models, rather than 'single hit' models, of causality.

In summary, using a large set of candidate SNP genotypes of IBD patients, we were able to exemplify use of the powerful CO design for $G \times G$ interaction analysis. To ensure fulfillment of the independence condition underlying the CO design, we confined our analysis to SNP pairs on different chromosome arms. Additional sources of systematic genotype association were accounted for by the inclusion of principal components in the statistical models used. Even with the CO design and a very large sample size, however, no statistically significant interactions were found in our

comprehensive $G \times G$ analysis of IBD after taking multiple testing into account. This notwithstanding, the study revealed one conspicuous interaction between variants in the *IL27* and *KPNA7* genes that links well with known etiological aspects of the disease, in this case the relevance of NF- κ B signaling for regulating pro-inflammatory and anti-inflammatory pathways. Nonetheless, similar to the CC-based findings by Zhang et al. [19] where significant interactions were confined to the MHC region, the overall outcome of our study was sobering and suggests that the missing heritability of IBD cannot be found primarily in $G \times G$ interactions. This insight, which in our case may trace primarily to the proven polygenic nature of IBD, calls for pursuing other avenues of research to fully understand the 'genetic architecture' of this and other human conditions, including the search for rare variants and the study of structural and epigenetic variation.

Acknowledgements

We thank Brittany Burmester for her assistance during her internship with us as well as Dr. Silke Szymczak and Olaf Junge for useful comments and technical support. A full list of members and affiliations of the International IBD Genetics Consortium is provided in the Supplementary Material.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Deutsche Forschungsgemeinschaft through the Research Training Group 1743 'Genes, Environment and Inflammation' [grant number GRK 1743, GRK 1743/2].

Data availability

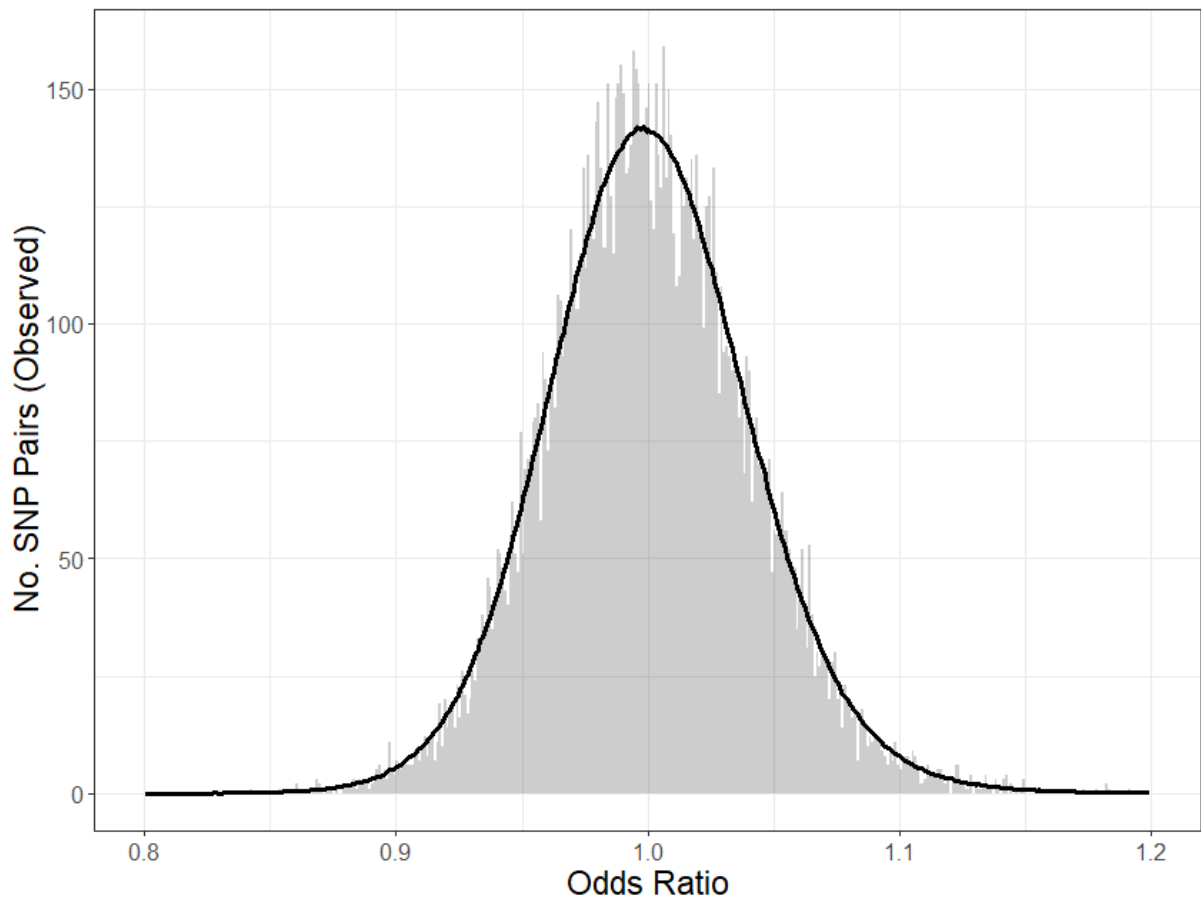
The data that support the findings of this study are available from the International IBD Genetics Consortium. Restrictions may apply to the availability of these data, which were used with the permission of the International IBD Genetics Consortium.

References

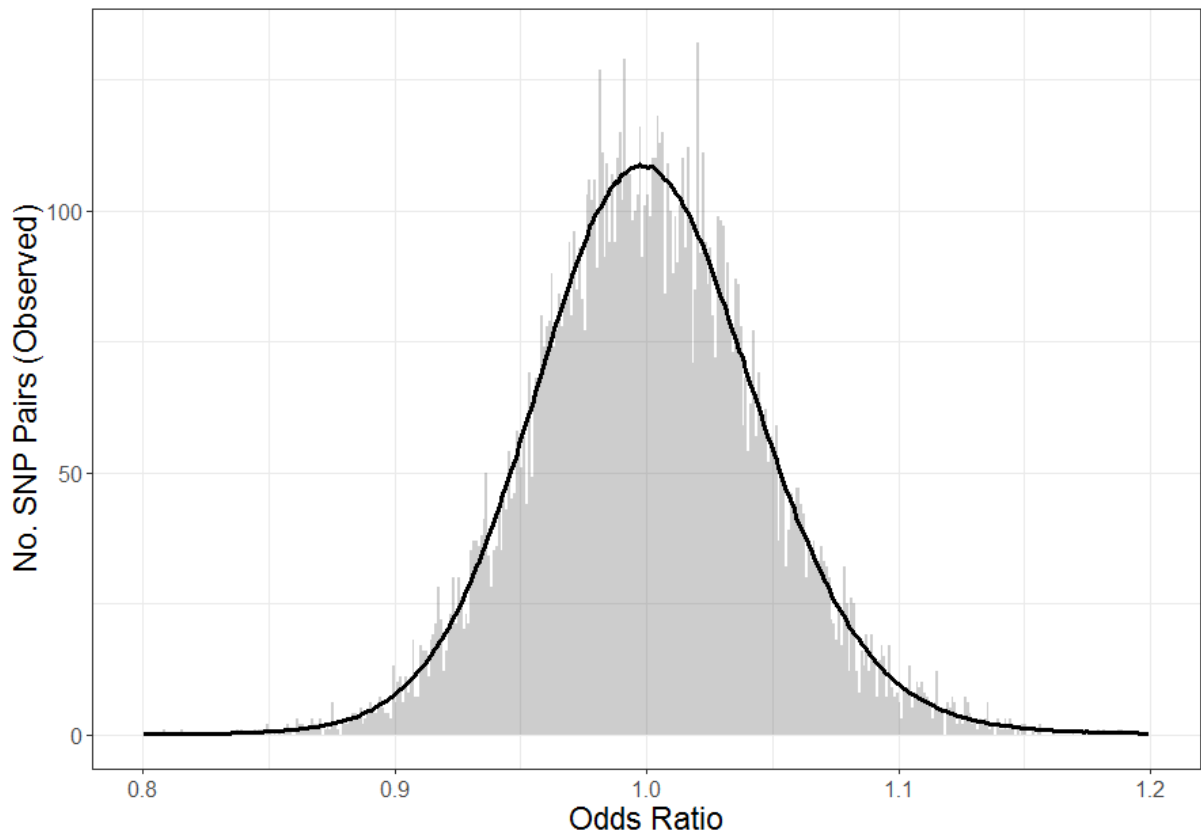
- [1] Buniello A, MacArthur J, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 2019; 47(D1):D1005–12.
- [2] Maher B. Personal genomes: the case of the missing heritability. *Nature.* 2008;456(7218):18–21.
- [3] Carlborg Ö, Haley CS. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet.* 2004;5(8):618–625.
- [4] Wei W-H, Hemani G, Haley CS. Detecting epistasis in human complex traits. *Nat Rev Genet.* 2014;15(11):722–733.
- [5] Bateson W, Mendel G. Mendel's principles of heredity/by W. Bateson. Cambridge: University Press; 1909.
- [6] Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet.* 2008; 9(11):855–867.
- [7] Greenland S. Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology.* 2009;20(1):14–17.
- [8] Culverhouse R, Suarez BK, Lin J, et al. A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet.* 2002;70(2):461–471.
- [9] Gauderman WJ. Sample size requirements for association studies of gene–gene interaction. *Am J Epidemiol.* 2002;155(5):478–484.
- [10] Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med.* 1994;13(2): 153–162.
- [11] Ng SC, Shi HY, Hamidi N, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet (Lond Engl).* 2017;390(10114):2769–2778.
- [12] Panaccione R. Mechanisms of inflammatory bowel disease. *Gastroenterol Hepatol (NY).* 2013;9(8):529–532.
- [13] Liu JZ, van Sommeren S, Huang H, International IBD Genetics Consortium, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet.* 2015;47(9):979–986.
- [14] Cho JH, Brant SR. Recent insights into the genetics of inflammatory bowel disease. *Gastroenterology.* 2011;140(6):1704–1712.e2.
- [15] Glas J, Seiderer J, Wagner J, et al. Analysis of *IL12B* gene variants in inflammatory bowel disease. *PLOS One.* 2012;7(3):e34349.
- [16] Polgar N, Csonge V, Szabo M, et al. Investigation of *JAK2*, *STAT3* and *CCR6* polymorphisms and their gene–gene interactions in inflammatory bowel disease: investigation of *JAK2*, *STAT3* and *CCR6* polymorphisms. *Int J Immunogenet.* 2012;39(3):247–252.
- [17] Glas J, Wagner J, Seiderer J, et al. *PTPN2* gene variants are associated with susceptibility to both Crohn's disease and ulcerative colitis supporting a common genetic disease background. *PLOS One.* 2012;7(3):e33682.
- [18] Roberts RL, Topless R, Phipps-Green AJ, et al. Evidence of interaction of *CARD8* rs2043211 with *NALP3* rs35829419 in Crohn's disease. *Genes Immun.* 2010;11(4):351–356.
- [19] Zhang J, Wei Z, Cardinale CJ, et al. Multiple epistasis interactions within MHC are associated with ulcerative colitis. *Front Genet.* 2019;10:257.
- [20] Yadav P, Ellinghaus D, Rémy G, et al. Genetic factors interact with tobacco smoke to modify risk for inflammatory bowel disease in humans and mice. *Gastroenterology.* 2017;153(2):550–565.
- [21] Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature.* 2012;491(7422):119–124.
- [22] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559–575.
- [23] Viechtbauer W. Conducting meta-analyses in R with the metafor Package. *J Stat Softw.* 2010;36(3):1–48.
- [24] Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet.* 2003;361(9357):598–604.
- [25] Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38(8):904–909.
- [26] The 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature.* 2015;526(7571):68–74.
- [27] Gauderman WJ. Sample size requirements for matched case-control studies of gene–environment interaction. *Stat Med.* 2002; 21(1):35–50.
- [28] Zuk O, Hechter E, Sunyaev SR, et al. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA.* 2012;109(4):1193–1198.
- [29] Andrews C, McLean MH, Durum SK. Interleukin-27 as a novel therapy for inflammatory bowel disease: a critical review of the literature. *Inflamm Bowel Dis.* 2016;22(9):2255–2264.
- [30] Pumroy RA, Cingolani G. Diversification of importin- α isoforms in cellular trafficking and disease states. *Biochem J.* 2015;466(1): 13–28.

- [31] McDaniel DK, Eden K, Ringel VM, et al. Emerging roles for noncanonical NF- κ B signaling in the modulation of inflammatory bowel disease pathobiology. *Inflamm Bowel Dis.* 2016;22(9):2265–2279.
- [32] Rothman KJ. Synergy and antagonism in cause–effect relationships. *Am J Epidemiol.* 1974;99(6):385–388.
- [33] Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol.* 1991;44(3):221–232.
- [34] Sarasin A. An overview of the mechanisms of mutagenesis and carcinogenesis. *Mutat Res.* 2003;544(2-3):99–106.

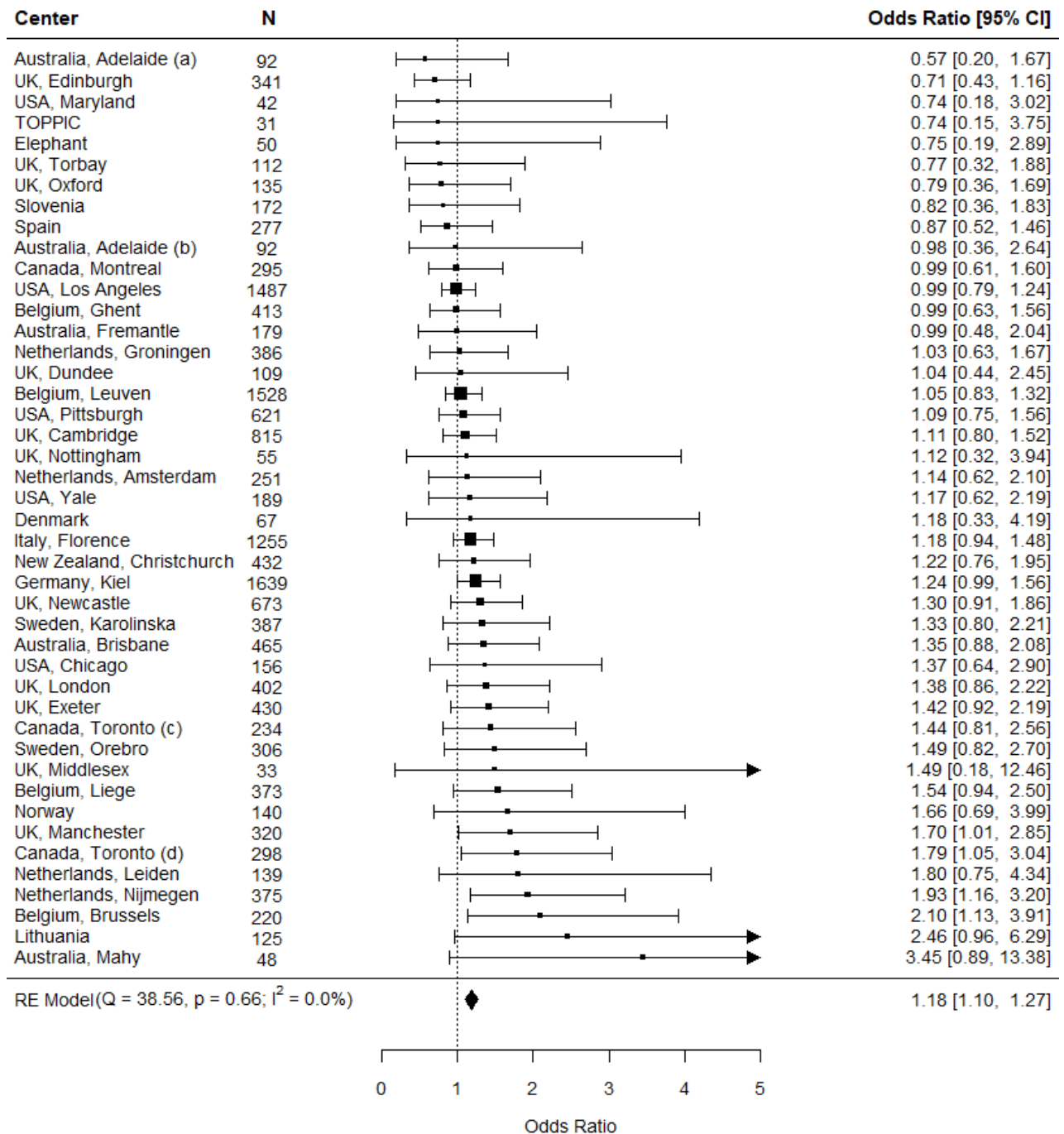
Supplementary Figures



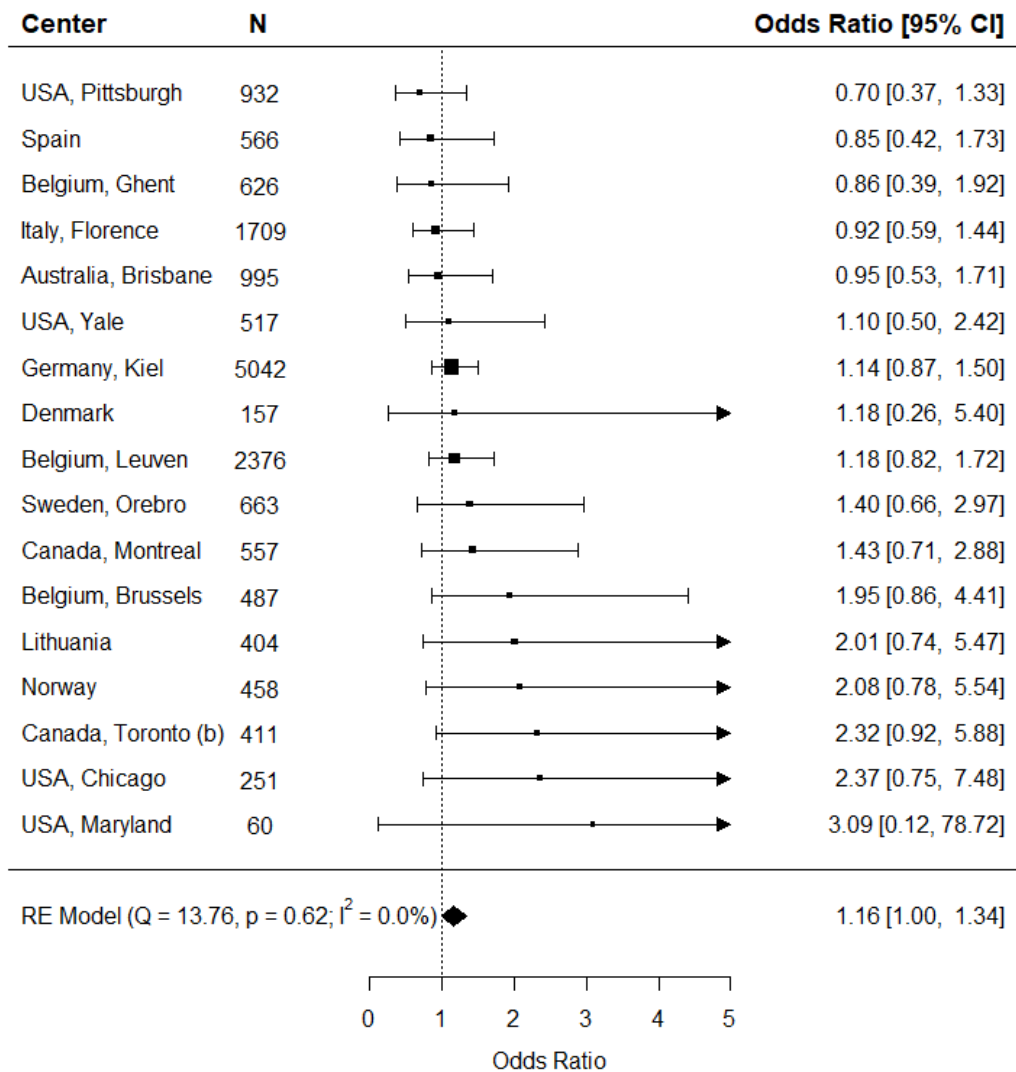
Supplementary Figure 1a: Distribution of interaction odds ratios observed among candidate SNP pairs (grey bars) in CD and as expected from a random subset of SNP pairs ($n=2 \times 10^7$, black curve).



Supplementary Figure 1b: Distribution of interaction odds ratios observed among candidate SNP pairs (grey bars) in UC and as expected from a random subset of SNP pairs ($n=2 \times 10^7$, black curve).



Supplementary Figure 2: Forest plot of CO interaction odds ratio for SNPs rs26528 and rs9297145 in CD. N: number of cases; CI: confidence interval; RE Model: random effects meta-analysis model; Q: Cochran's Q; p: p value of heterogeneity test (Cochran's Q); I²: I-squared index for heterogeneity. (a) Royal Adelaide Hospital. (b) Flinders Medical Centre. (c) National Institute of Diabetes, Digestive and Kidney Diseases, IBD Genetics Consortium. (d) University of Toronto.



Supplementary Figure 3: Forest plot of CC interaction odds ratio for SNPs rs26528 and rs9297145 in CD. N: number of cases; CI: confidence interval; RE Model: random effects meta-analysis model; Q: Cochran's Q; p: p value of heterogeneity test (Cochran's Q); I²: I-squared index for heterogeneity. (b) University of Toronto.

Supplementary Tables

Supplementary Table 1a: Top G×G interactions (p<0.005) between SNPs associated with CD (CO design)

ID: SNP rs number; MAF: minor allele frequency; OR: odds ratio; SE: standard error of interaction odds ratio; P: nominal p value of meta-analysis Wald test; \$: allelic disease odds ratio of minor allele, according to Liu et al. (2015); &: interaction odds ratio.

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs26528	<i>IL27</i>	0.47	1.13	rs9297145	<i>KPNA7</i>	0.28	1.08	1.18	0.04	7.75×10 ⁻⁶
rs3740415	<i>MFSD13A</i>	0.45	0.95	rs12199775		0.06	0.89	0.81	0.05	7.19×10 ⁻⁵
rs12946510	<i>GRB7. IKZF3</i>	0.49	1.13	rs17057051	<i>PTK2B</i>	0.30	0.94	0.87	0.04	1.19×10 ⁻⁴
rs194749	<i>ZFP36L1</i>	0.24	1.08	rs314313	<i>EPHB4</i>	0.32	1.07	0.88	0.03	1.28×10 ⁻⁴
rs11010067		0.39	1.14	rs7554511	<i>C1orf106</i>	0.24	0.86	0.88	0.04	3.88×10 ⁻⁴
rs13407913	<i>ADCY3</i>	0.46	1.12	rs1748195	<i>DOCK7</i>	0.34	1.07	1.13	0.04	3.91×10 ⁻⁴
rs3091315	<i>CCL2. CCL7</i>	0.25	0.87	rs921720	<i>LOC105375746</i>	0.36	0.89	0.89	0.03	4.86×10 ⁻⁴
rs17057051	<i>PTK2B</i>	0.30	0.94	rs314313	<i>EPHB4</i>	0.32	1.07	1.12	0.03	5.03×10 ⁻⁴
rs4768236	<i>LRRK2</i>	0.37	1.12	rs7773324	<i>DUSP22. IRF4</i>	0.38	0.92	1.12	0.03	5.75×10 ⁻⁴
rs2945412	<i>KSR1</i>	0.38	0.89	rs568617	<i>FIBP</i>	0.20	1.09	0.89	0.03	8.44×10 ⁻⁴
rs3740415	<i>MFSD13A</i>	0.45	0.95	rs11583043	<i>DPH5</i>	0.28	1.03	1.13	0.04	8.96×10 ⁻⁴
rs259964	<i>ZNF831</i>	0.47	1.07	rs10065637	<i>ANKRD55</i>	0.20	0.91	1.13	0.04	9.01×10 ⁻⁴
rs6087990		0.41	1.05	rs12199775		0.06	0.89	0.84	0.05	1.06×10 ⁻³
rs13126505	<i>BANK1</i>	0.08	1.20	rs7555082		0.12	1.13	1.19	0.05	1.07×10 ⁻³
rs1893217	<i>PTPN2</i>	0.18	1.18	rs17057051	<i>PTK2B</i>	0.30	0.94	1.12	0.03	1.12×10 ⁻³

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR ^s	ID	Gene	MAF	OR ^s	OR ^{&}	SE	P
rs915286		0.43	0.94	rs12718244	<i>SPATA48</i>	0.42	1.08	0.88	0.04	1.16×10 ⁻³
rs6087990		0.41	1.05	rs11554257		0.15	1.16	0.88	0.04	1.19×10 ⁻³
rs6856616		0.08	1.10	rs10495903	<i>THADA</i>	0.14	1.13	1.21	0.06	1.28×10 ⁻³
rs1042058	<i>MAP3K8</i>	0.39	0.93	rs113010081	<i>LOC105377068</i>	0.11	1.06	0.87	0.04	1.28×10 ⁻³
rs10781499	<i>CARD9</i>	0.46	1.18	rs6716753	<i>SP140</i>	0.21	1.14	0.89	0.04	1.37×10 ⁻³
rs7954567	<i>LTBR</i>	0.34	1.09	rs11741861	<i>ZNF300</i>	0.11	1.33	0.88	0.04	1.39×10 ⁻³
rs516246	<i>FUT2</i>	0.50	1.12	rs10865331		0.41	1.08	0.88	0.04	1.39×10 ⁻³
rs11554257		0.15	1.16	rs2816958	<i>NR5A2</i>	0.11	0.96	1.17	0.05	1.40×10 ⁻³
rs1569328	<i>FOS</i>	0.15	0.90	rs7954567	<i>LTBR</i>	0.34	1.09	1.13	0.04	1.59×10 ⁻³
rs3740415	<i>MFSD13A</i>	0.45	0.95	rs2503322	<i>RSPO3</i>	0.45	0.94	0.89	0.04	1.61×10 ⁻³
rs1569328	<i>FOS</i>	0.15	0.90	rs7555082		0.12	1.13	1.14	0.04	1.78×10 ⁻³
rs724016	<i>ZBTB38</i>	0.44	1.06	rs2382817	<i>TMBIM1</i>	0.42	1.07	0.89	0.04	1.79×10 ⁻³
rs10995235	<i>ZNF365</i>	0.18	1.05	rs314313	<i>EPHB4</i>	0.32	1.07	1.11	0.03	1.81×10 ⁻³
rs13204742		0.14	1.15	rs2816958	<i>NR5A2</i>	0.11	0.96	1.15	0.04	1.85×10 ⁻³
rs4246905	<i>TNFSF15</i>	0.26	0.87	rs11741861	<i>ZNF300</i>	0.11	1.33	1.13	0.04	1.87×10 ⁻³
rs11554257		0.15	1.16	rs4976646	<i>RGS14</i>	0.36	1.07	1.14	0.04	2.03×10 ⁻³
rs2382817	<i>TMBIM1</i>	0.42	1.07	rs11583043	<i>DPH5</i>	0.28	1.03	1.11	0.03	2.07×10 ⁻³
rs212388	<i>LOC105378083</i>	0.43	1.11	rs7608910	<i>PUS10</i>	0.43	1.13	1.12	0.04	2.10×10 ⁻³
rs10865331		0.41	1.08	rs3806308	<i>RNF186</i>	0.36	0.97	1.11	0.04	2.26×10 ⁻³
rs224090		0.45	1.14	rs12722515	<i>IL2RA</i>	0.14	0.88	0.88	0.04	2.26×10 ⁻³

SNP-1					SNP-2					Interaction		
ID	Gene	MAF	OR ^s		ID	Gene	MAF	OR ^s		OR ^{&}	SE	P
rs6708413	<i>IL18RAP</i>	0.26	1.12		rs670523	<i>RIT1</i>	0.35	1.08		0.90	0.03	2.38×10 ⁻³
rs1077773	<i>KCCAT333</i>	0.46	0.97		rs6856616		0.08	1.10		0.86	0.05	2.41×10 ⁻³
rs8005161	<i>GPR65</i>	0.10	1.17		rs1250546	<i>ZMIZ1</i>	0.38	0.90		0.88	0.04	2.50×10 ⁻³
rs4743820	<i>LINC00484</i>	0.29	0.95		rs2538470		0.38	1.08		0.91	0.03	2.50×10 ⁻³
rs2836878		0.25	0.90		rs4246905	<i>TNFSF15</i>	0.26	0.87		0.91	0.03	2.52×10 ⁻³
rs2155219	<i>EMSY</i>	0.45	1.19		rs12199775		0.06	0.89		0.84	0.06	2.55×10 ⁻³
rs2538470		0.38	1.08		rs2382817	<i>TMBIM1</i>	0.42	1.07		0.90	0.04	2.71×10 ⁻³
rs16967103		0.21	1.11		rs653178	<i>ATXN2</i>	0.48	1.07		1.12	0.04	2.84×10 ⁻³
rs2226628	<i>LOC107984361</i>	0.29	1.03		rs17622378	<i>C5orf56</i>	0.47	1.21		0.89	0.04	2.92×10 ⁻³
rs7954567	<i>LTBR</i>	0.34	1.09		rs395157	<i>OSMR</i>	0.48	1.10		1.11	0.04	2.96×10 ⁻³
rs13300483		0.27	1.14		rs7011507		0.12	0.93		1.12	0.04	3.19×10 ⁻³
rs9297145	<i>KPNA7</i>	0.28	1.08		rs7773324	<i>DUSP22.IRF4</i>	0.38	0.92		0.89	0.04	3.24×10 ⁻³
rs6087990		0.41	1.05		rs1505992		0.28	0.82		1.12	0.04	3.41×10 ⁻³
rs11230563	<i>CD6</i>	0.33	0.92		rs11583043	<i>DPH5</i>	0.28	1.03		0.90	0.04	3.45×10 ⁻³
rs7954567	<i>LTBR</i>	0.34	1.09		rs1505992		0.28	0.82		0.91	0.03	3.51×10 ⁻³
rs7282490		0.42	1.13		rs4256159	<i>LOC105376976</i>	0.16	1.13		1.14	0.05	3.54×10 ⁻³
rs913678		0.33	0.95		rs9358372	<i>CDKAL1</i>	0.39	1.08		0.91	0.03	3.66×10 ⁻³
rs11054935	<i>DUSP16</i>	0.28	1.03		rs12199775		0.06	0.89		1.16	0.05	3.71×10 ⁻³
rs11641184	<i>LITAF</i>	0.50	1.08		rs7438704	<i>SLAIN2</i>	0.33	0.92		1.14	0.05	3.72×10 ⁻³
rs2823286		0.26	0.87		rs12942547	<i>STAT3</i>	0.39	0.90		1.10	0.03	3.72×10 ⁻³

SNP-1				SNP-2				Interaction			
ID	Gene	MAF	ID	Gene	MAF	ID	Gene	MAF	ID	Gene	
rs7165170	<i>CRTC3</i>	0.17	0.93	rs653178	<i>ATXN2</i>	0.48	1.07	0.89	0.04	4.00×10 ⁻³	
rs9525625	<i>LINC02341</i>	0.50	1.08	rs2227551	<i>PLAU</i>	0.24	0.91	0.90	0.04	4.06×10 ⁻³	
rs2823286		0.26	0.87	rs314313	<i>EPHB4</i>	0.32	1.07	0.91	0.03	4.12×10 ⁻³	
rs259964	<i>ZNF831</i>	0.47	1.07	rs1077773	<i>KCCAT333</i>	0.46	0.97	1.12	0.04	4.15×10 ⁻³	
rs727563	<i>ACO2</i>	0.23	1.10	rs1847472	<i>BACH2</i>	0.33	0.92	1.11	0.04	4.17×10 ⁻³	
rs1292053	<i>TUBD1</i>	0.47	1.10	rs566416		0.23	0.94	0.90	0.04	4.21×10 ⁻³	
rs12627970		0.21	1.12	rs1456896		0.29	0.91	0.90	0.04	4.22×10 ⁻³	
rs925255	<i>FOSL2</i>	0.42	0.90	rs3024505	<i>IL10</i>	0.18	1.18	0.90	0.04	4.25×10 ⁻³	
rs2155219	<i>EMSY</i>	0.45	1.19	rs11010067		0.39	1.14	1.11	0.04	4.40×10 ⁻³	
rs6908425	<i>CDKAL1</i>	0.19	0.90	rs113010081	<i>LOC105377068</i>	0.11	1.06	0.88	0.04	4.40×10 ⁻³	
rs4246905	<i>TNFSF15</i>	0.26	0.87	rs11742570		0.33	0.78	0.91	0.03	4.46×10 ⁻³	
rs17780256	<i>SLC39A11</i>	0.19	0.95	rs6679677	<i>PHTF1</i>	0.08	0.83	1.15	0.05	4.62×10 ⁻³	
rs913678		0.33	0.95	rs11229555		0.24	0.93	1.10	0.03	4.62×10 ⁻³	
rs7282490		0.42	1.13	rs4656958		0.30	0.94	1.10	0.03	4.66×10 ⁻³	
rs224090		0.45	1.14	rs1505992		0.28	0.82	0.90	0.04	4.77×10 ⁻³	
rs7236492	<i>NFATC1</i>	0.13	0.91	rs12199775		0.06	0.89	1.22	0.07	4.77×10 ⁻³	
rs727563	<i>ACO2</i>	0.23	1.10	rs17293632	<i>SMAD3</i>	0.26	1.14	1.10	0.03	4.82×10 ⁻³	
rs11083840	<i>PTGIR</i>	0.41	1.03	rs7517810		0.27	1.14	1.11	0.04	4.86×10 ⁻³	
rs564349	<i>ERGIC1</i>	0.33	1.05	rs3766606	<i>PARK7</i>	0.15	0.90	1.11	0.04	4.93×10 ⁻³	

Supplementary Table 1b: Top G×G interactions ($p < 0.005$) between SNPs associated with UC (CO design)
 ID: SNP rs number; MAF: minor allele frequency; OR: odds ratio; SE: standard error of interaction odds ratio; P: nominal p value of meta-analysis Wald test; \$: allelic disease odds ratio of minor allele. according to Liu et al. (2015); &: interaction odds ratio.

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR ^{\$}	ID	Gene	MAF	OR ^{\$}	OR ^{&}	SE	P
rs4722672		0.20	1.08	rs12568930	ZBTB40	0.15	0.88	1.18	0.04	1.43×10^{-4}
rs11742570		0.37	0.92	rs6716753	SP140	0.19	1.04	1.16	0.04	1.76×10^{-4}
rs7805114		0.40	0.89	rs10495903	THADA	0.14	1.07	0.82	0.05	2.42×10^{-4}
rs3764147	LACC1	0.24	1.04	rs653178	ATXN2	0.49	1.05	0.86	0.04	3.74×10^{-4}
rs17085007		0.20	1.14	rs3766606	PARK7	0.14	0.87	1.16	0.04	4.6×10^{-4}
rs3764147	LACC1	0.24	1.04	rs3740415	MFSD13A	0.45	0.94	0.87	0.04	7.00×10^{-4}
rs11083840	PTGIR	0.42	1.07	rs10185424		0.48	1.10	1.16	0.04	7.25×10^{-4}
rs2024092	SBN02	0.22	1.05	rs28374715	CHP1	0.25	0.92	1.14	0.04	7.47×10^{-4}
rs6908425	CDKAL1	0.20	0.93	rs2651244		0.38	0.94	1.14	0.04	7.69×10^{-4}
rs13277237	CCDC26	0.45	1.06	rs2538470		0.38	1.07	1.15	0.04	8.08×10^{-4}
rs516246	FUT2	0.48	1.04	rs3742130	GPR18	0.21	0.93	0.87	0.04	8.67×10^{-4}
rs12946510	GRB7, IKZF3	0.49	1.14	rs4845604	RORC	0.13	0.85	0.85	0.05	9.79×10^{-4}
rs17293632	SMAD3	0.25	1.08	rs564349	ERGIC1	0.33	1.06	0.87	0.04	1.01×10^{-3}
rs11168249	HDAC7	0.48	1.06	rs6667605	LOC100996583	0.47	0.92	1.16	0.05	1.33×10^{-3}
rs11010067		0.37	1.08	rs3806308	RNF186	0.33	0.84	1.13	0.04	1.52×10^{-3}
rs13204048	SLC22A23	0.37	0.97	rs3024505	IL10	0.19	1.25	1.13	0.04	1.64×10^{-3}
rs516246	FUT2	0.48	1.04	rs11230563	CD6	0.34	0.93	1.15	0.05	1.67×10^{-3}
rs26528	IL27	0.47	1.06	rs1182188	GNA12	0.27	0.90	0.88	0.04	1.77×10^{-3}

SNP-1					SNP-2					Interaction		
ID	Gene	MAF	OR ^s		ID	Gene	MAF	OR ^s		OR ^{&}	SE	P
rs941823	LINC00598	0.22	0.90		rs9297145	KPNA7	0.27	1.07		1.12	0.04	1.77×10 ⁻³
rs12946510	GRB7, IKZF3	0.49	1.14		rs13126505	BANK1	0.07	1.09		0.82	0.06	1.81×10 ⁻³
rs907611	LSP1	0.33	1.08		rs9313808		0.15	0.88		0.88	0.04	1.88×10 ⁻³
rs2227551	PLAU	0.25	0.96		rs1042058	MAP3K8	0.39	0.95		0.89	0.04	1.92×10 ⁻³
rs9297145	KPNA7	0.27	1.07		rs10065637	ANKRD55	0.21	0.96		1.12	0.04	1.92×10 ⁻³
rs174537	MYRF	0.33	1.03		rs2457996		0.10	0.90		1.16	0.05	1.96×10 ⁻³
rs13126505	BANK1	0.07	1.09		rs6856616		0.08	1.10		1.27	0.08	2.00×10 ⁻³
rs2024092	SBN02	0.22	1.05		rs7554511	C1orf106	0.24	0.85		0.89	0.04	2.00×10 ⁻³
rs12778642	HHEX	0.45	1.06		rs6716753	SP140	0.19	1.04		1.14	0.04	2.02×10 ⁻³
rs7236492	NFATC1	0.14	0.95		rs1505992		0.31	0.95		0.88	0.04	2.08×10 ⁻³
rs4812833		0.45	0.90		rs11230563	CD6	0.34	0.93		1.13	0.04	2.15×10 ⁻³
rs6863411	NDFIP1	0.36	0.94		rs17229285	LOC105373831	0.47	1.10		1.14	0.04	2.16×10 ⁻³
rs3764147	LACC1	0.24	1.04		rs3851228	TRAF3IP2-AS1	0.08	1.20		1.18	0.06	2.33×10 ⁻³
rs6716753	SP140	0.19	1.04		rs3024505	IL10	0.19	1.25		0.88	0.04	2.40×10 ⁻³
rs11010067		0.37	1.08		rs13300218	NOTCH1	0.10	0.87		1.17	0.05	2.50×10 ⁻³
rs3172494	IP6K2	0.12	1.11		rs1517352	STAT4	0.39	0.93		0.87	0.04	2.61×10 ⁻³
rs10761659		0.43	0.89		rs6667605	LOC100996583	0.47	0.92		0.88	0.04	2.65×10 ⁻³
rs314313	EPHB4	0.32	1.06		rs4976646	RGS14	0.36	1.08		1.12	0.04	2.68×10 ⁻³
rs3742130	GPR18	0.21	0.93		rs7134472	LOC107984526	0.42	1.17		1.13	0.04	2.78×10 ⁻³
rs2226628	LOC107984361	0.29	1.03		rs10495903	THADA	0.14	1.07		1.15	0.05	2.80×10 ⁻³

SNP-1					SNP-2					Interaction		
ID	Gene	MAF	OR ^s		ID	Gene	MAF	OR ^s		OR ^{&}	SE	P
rs17736589	CYTH1	0.22	1.09		rs28374715	CHP1	0.25	0.92		1.12	0.04	2.90×10 ⁻³
rs10065637	ANKRD55	0.21	0.96		rs6716753	SP140	0.19	1.04		0.89	0.04	2.97×10 ⁻³
rs13300218	NOTCH1	0.10	0.87		rs13277237	CCDC26	0.45	1.06		0.86	0.05	3.00×10 ⁻³
rs4812833		0.45	0.90		rs1728785	ZFP90	0.22	0.93		0.88	0.04	3.00×10 ⁻³
rs7282490		0.41	1.11		rs17771967		0.44	1.07		1.13	0.04	3.05×10 ⁻³
rs13277237	CCDC26	0.45	1.06		rs3851228	TRAF3IP2-AS1	0.08	1.20		0.85	0.06	3.09×10 ⁻³
rs6863411	NDFIP1	0.36	0.94		rs2651244		0.38	0.94		0.89	0.04	3.12×10 ⁻³
rs3851228	TRAF3IP2-AS1	0.08	1.20		rs4692386		0.39	0.95		0.85	0.05	3.17×10 ⁻³
rs6856616		0.08	1.10		rs12568930	ZBTB40	0.15	0.88		1.18	0.06	3.18×10 ⁻³
rs559928		0.17	0.92		rs9313808		0.15	0.88		0.87	0.05	3.30×10 ⁻³
rs1363907	ERAP2	0.43	1.06		rs2382817	TMBIM1	0.42	1.07		1.14	0.05	3.32×10 ⁻³
rs11054935	DUSP16	0.29	1.08		rs4380874	DLD	0.44	1.14		0.89	0.04	3.41×10 ⁻³
rs174537	MYRF	0.33	1.03		rs9297145	KPNA7	0.27	1.07		1.11	0.04	3.46×10 ⁻³
rs17736589	CYTH1	0.22	1.09		rs9358372	CDKAL1	0.38	1.04		0.89	0.04	3.53×10 ⁻³
rs10781499	CARD9	0.45	1.14		rs6908425	CDKAL1	0.20	0.93		1.13	0.04	3.55×10 ⁻³
rs727563	AC02	0.21	1.03		rs9358372	CDKAL1	0.38	1.04		0.87	0.05	3.56×10 ⁻³
rs921720	LOC105375746	0.38	0.96		rs3774937	NFKB1	0.35	1.10		1.12	0.04	3.68×10 ⁻³
rs6087990		0.42	1.06		rs3091315	CCL2. CCL7	0.26	0.94		1.12	0.04	3.73×10 ⁻³
rs395157	OSMR	0.49	1.09		rs3749171	GPR35	0.20	1.15		1.14	0.04	3.75×10 ⁻³
rs2361755		0.07	0.91		rs2382817	TMBIM1	0.42	1.07		0.85	0.06	3.77×10 ⁻³

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR ^s	ID	Gene	MAF	OR ^s	OR ^{&}	SE	P
rs6740462		0.25	0.94	rs10798069	PLA2G4A	0.48	0.97	1.13	0.04	3.78×10 ⁻³
rs1893217	PTPN2	0.18	1.13	rs3774937	NFKB1	0.35	1.10	0.89	0.04	3.81×10 ⁻³
rs4743820	LINC00484	0.28	0.92	rs3766606	PARK7	0.14	0.87	0.89	0.04	3.89×10 ⁻³
rs9868809	CELSR3	0.12	1.16	rs3766606	PARK7	0.14	0.87	1.17	0.05	3.92×10 ⁻³
rs2413583		0.15	0.87	rs11741861	ZNF300	0.09	1.13	1.17	0.05	4.06×10 ⁻³
rs11083840	PTGIR	0.42	1.07	rs1728785	ZFP90	0.22	0.93	1.13	0.04	4.12×10 ⁻³
rs9297145	KPNA7	0.27	1.07	rs3851228	TRAF3IP2-AS1	0.08	1.20	0.86	0.05	4.28×10 ⁻³
rs1250546	ZMIZ1	0.40	0.96	rs4722672		0.20	1.08	0.89	0.04	4.33×10 ⁻³
rs7282490		0.41	1.11	rs174537	MYRF	0.33	1.03	0.89	0.04	4.43×10 ⁻³
rs7404095	PRKCB	0.41	0.93	rs11743851	CDC42SE2	0.38	1.06	0.89	0.04	4.45×10 ⁻³
rs7282490		0.41	1.11	rs2226628	LOC107984361	0.29	1.03	1.13	0.04	4.47×10 ⁻³
rs12946510	GRB7, IKZF3	0.49	1.14	rs2382817	TMBIM1	0.42	1.07	1.14	0.04	4.47×10 ⁻³
rs2836878		0.23	0.80	rs13204742		0.13	1.04	1.13	0.04	4.49×10 ⁻³
rs1728785	ZFP90	0.22	0.93	rs12778642	HHEX	0.45	1.06	0.89	0.04	4.51×10 ⁻³
rs9313808		0.15	0.88	rs2073505	HGFAC	0.09	1.09	1.17	0.06	4.55×10 ⁻³
rs11054935	DUSP16	0.29	1.08	rs7097656	TSPAN14	0.19	0.94	1.12	0.04	4.70×10 ⁻³

Supplementary Table 1c: Top G×G interactions ($p < 0.005$) between SNPs associated with IBD (CO design)
 ID: SNP rs number; MAF: minor allele frequency; OR: odds ratio; SE: standard error of interaction odds ratio; P: nominal p value of meta-analysis Wald test; \$: allelic disease odds ratio of minor allele. according to Liu et al. (2015); &: interaction odds ratio.

SNP-1					SNP-2					Interaction		
ID	Gene	MAF	OR ^{\$}	OR ^{&}	ID	Gene	MAF	OR ^{\$}	OR ^{&}	OR ^{&}	SE	P
rs6556412	LOC285626	0,36	1,13	1,08	rs1260326	GCKR	0,44	1,08	1,11	1,03	1,03	6,90×10 ⁻⁵
rs2945412	KSR1	0,39	0,95	1,08	rs6716753	SP140	0,20	1,08	0,89	1,03	1,03	1,55×10 ⁻⁴
rs11083840	PTGIR	0,42	1,05	1,09	rs10185424		0,48	1,09	1,11	1,03	1,03	1,57×10 ⁻⁴
rs3091315		0,25	0,90	0,92	rs921720	LOC105375746	0,37	0,92	0,91	1,02	1,02	1,95×10 ⁻⁴
rs1569328		0,15	0,92	1,06	rs212388	LOC105378083	0,41	1,06	0,90	1,03	1,03	2,05×10 ⁻⁴
rs7240004		0,37	0,94	0,92	rs3742130	GPR18	0,21	0,92	1,10	1,03	1,03	2,65×10 ⁻⁴
rs6087990		0,41	1,05	0,90	rs3091315		0,25	0,90	1,10	1,03	1,03	3,17×10 ⁻⁴
rs727563		0,22	1,06	1,11	rs17293632	SMAD3	0,25	1,11	1,09	1,02	1,02	3,54×10 ⁻⁴
rs1077773	KCCAT333	0,46	0,95	1,10	rs6856616		0,08	1,10	0,88	1,04	1,04	3,58×10 ⁻⁴
rs4409764		0,47	1,18	0,85	rs7554511	C1orf106	0,24	0,85	0,91	1,03	1,03	4,68×10 ⁻⁴
rs941823	LINC00598	0,23	0,92	0,95	rs616597	NFKBIZ	0,22	0,95	1,10	1,03	1,03	4,87×10 ⁻⁴
rs2503322	RSPO3	0,45	0,96	1,06	rs2111485		0,42	1,06	0,91	1,03	1,03	5,43×10 ⁻⁴
rs653178	ATXN2	0,49	1,06	1,06	rs907611	LSP1	0,33	1,06	1,10	1,03	1,03	5,86×10 ⁻⁴
rs13126505	BANK1	0,08	1,14	0,94	rs4692386		0,39	0,94	0,89	1,04	1,04	7,14×10 ⁻⁴
rs7404095	PRKCB	0,41	0,95	1,05	rs3774937	NFKB1	0,33	1,05	0,92	1,03	1,03	7,25×10 ⁻⁴
rs4246905	TNFSF15	0,26	0,88	0,84	rs11742570		0,35	0,84	0,92	1,02	1,02	7,40×10 ⁻⁴
rs259964	ZNF831	0,47	1,07	0,84	rs11742570		0,35	0,84	1,09	1,03	1,03	9,65×10 ⁻⁴
rs17057051	PTK2B	0,30	0,94	1,07	rs314313	×10PHB4	0,32	1,07	1,08	1,02	1,02	9,72×10 ⁻⁴

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR ^s	ID	Gene	MAF	OR ^s	OR ^{&}	SE	P
rs1893217	PTPN2	0,18	1,15	rs17057051	PTK2B	0,30	0,94	1,09	1,03	1,03×10 ⁻³
rs26528	IL27	0,47	1,10	rs9297145	KPNA7	0,28	1,07	1,09	1,03	1,08×10 ⁻³
rs26528	IL27	0,47	1,10	rs907611	LSP1	0,33	1,06	1,09	1,03	1,08×10 ⁻³
rs4246905	TNFSF15	0,26	0,88	rs11741861	ZNF300	0,10	1,22	1,11	1,03	1,12×10 ⁻³
rs915286		0,43	0,94	rs12718244	SPATA48	0,42	1,08	0,92	1,03	1,13×10 ⁻³
rs26528	IL27	0,47	1,10	rs1842076		0,27	0,92	0,92	1,03	1,19×10 ⁻³
rs16967103		0,21	1,07	rs10185424		0,48	1,09	0,91	1,03	1,19×10 ⁻³
rs7011507		0,12	0,92	rs4692386		0,39	0,94	0,91	1,03	1,19×10 ⁻³
rs2945412	KSR1	0,39	0,95	rs12722515	IL2RA	0,14	0,91	1,10	1,03	1,24×10 ⁻³
rs174537	MYRF	0,34	1,06	rs13204742		0,14	1,10	1,09	1,03	1,24×10 ⁻³
rs913678		0,32	0,93	rs7097656	TSPAN14	0,18	0,91	0,92	1,03	1,42×10 ⁻³
rs13300218	NOTCH1	0,09	0,86	rs13277237	CCDC26	0,45	1,06	0,90	1,03	1,43×10 ⁻³
rs11742570		0,35	0,84	rs6716753	SP140	0,20	1,08	1,08	1,03	1,44×10 ⁻³
rs921720	LOC105375746	0,37	0,92	rs11741861	ZNF300	0,10	1,22	1,11	1,03	1,46×10 ⁻³
rs11554257		0,15	1,15	rs2816958	NR5A2	0,11	0,91	1,11	1,03	1,55×10 ⁻³
rs4768236	LRRK2	0,36	1,08	rs7773324		0,38	0,94	1,08	1,03	1,68×10 ⁻³
rs2284553	IFNGR2	0,39	0,94	rs1847472	BACH2	0,33	0,93	1,09	1,03	1,75×10 ⁻³
rs7404095	PRKCB	0,41	0,95	rs10495903	THADA	0,14	1,11	1,11	1,03	1,75×10 ⁻³
rs2361755		0,07	0,87	rs10185424		0,48	1,09	0,88	1,04	1,78×10 ⁻³
rs925255	FOSL2	0,42	0,93	rs3024505		0,18	1,22	0,92	1,03	1,81×10 ⁻³

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR ^s	ID	Gene	MAF	OR ^s	OR ^{&}	SE	P
rs17736589	CYTH1	0,22	1,05	rs7555082		0,12	1,10	1,11	1,03	1,89×10 ⁻³
rs616597	NFKBIZ	0,22	0,95	rs6724516		0,27	0,94	0,93	1,02	1,99×10 ⁻³
rs11054935	DUSP16	0,28	1,05	rs7097656	TSPAN14	0,18	0,91	1,08	1,03	2,10×10 ⁻³
rs7954567	LTBR	0,33	1,05	rs2073505	HGFAC	0,09	1,10	0,90	1,03	2,20×10 ⁻³
rs11641184	LITAF	0,50	1,08	rs194749		0,24	1,06	0,92	1,03	2,36×10 ⁻³
rs12199775		0,06	0,90	rs4845604	RORC	0,13	0,88	1,14	1,04	2,38×10 ⁻³
rs224090		0,44	1,10	rs12722515	IL2RA	0,14	0,91	0,91	1,03	2,45×10 ⁻³
rs6087990		0,41	1,05	rs12199775		0,06	0,90	0,89	1,04	2,57×10 ⁻³
rs516246	FUT2	0,49	1,08	rs3742130	GPR18	0,21	0,92	0,92	1,03	2,64×10 ⁻³
rs7954567	LTBR	0,33	1,05	rs11741861	ZNF300	0,10	1,22	0,91	1,03	2,70×10 ⁻³
rs2155219		0,46	1,16	rs4703855		0,28	0,93	1,08	1,03	2,75×10 ⁻³
rs4768236	LRRK2	0,36	1,08	rs9868809	C×10LSR3	0,12	1,15	0,91	1,03	2,87×10 ⁻³
rs10065637	ANKRD55	0,20	0,93	rs6716753	SP140	0,20	1,08	0,93	1,03	2,94×10 ⁻³
rs3740415	MFSD13A	0,45	0,95	rs2503322	RSPO3	0,45	0,96	0,92	1,03	3,03×10 ⁻³
rs194749		0,24	1,06	rs4380874		0,42	1,08	0,93	1,03	3,04×10 ⁻³
rs4243971		0,43	0,95	rs3197999	MST1	0,32	1,18	1,09	1,03	3,09×10 ⁻³
rs13277237	CCDC26	0,45	1,06	rs3024505		0,18	1,22	1,08	1,03	3,10×10 ⁻³
rs3853824	C17orf67	0,34	0,94	rs6908425	CDKAL1	0,20	0,92	0,93	1,03	3,24×10 ⁻³
rs1042058	MAP3K8	0,39	0,94	rs7517810		0,26	1,06	1,08	1,03	3,40×10 ⁻³
rs17694108		0,30	1,09	rs6740462		0,24	0,92	0,93	1,02	3,49×10 ⁻³

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR ^s	ID	Gene	MAF	OR ^s	OR ^{&}	SE	P
rs194749		0,24	1,06	rs254560	C5orf66	0,41	1,06	1,08	1,03	3,52×10 ⁻³
rs8005161	GPR65	0,10	1,15	rs6556412	LOC285626	0,36	1,13	0,91	1,03	3,60×10 ⁻³
rs16967103		0,21	1,07	rs10761659		0,42	0,86	0,93	1,03	3,67×10 ⁻³
rs11879191	CDC37	0,18	0,89	rs3024505		0,18	1,22	0,93	1,03	3,92×10 ⁻³
rs11641184	LITAF	0,50	1,08	rs9525625	LINC02341	0,49	1,05	0,91	1,03	3,98×10 ⁻³
rs17736589	CYTH1	0,22	1,05	rs3749171	GPR35	0,19	1,12	0,93	1,03	3,99×10 ⁻³
rs2361755		0,07	0,87	rs6651252	LINC00824	0,12	0,91	1,13	1,04	4,28×10 ⁻³
rs4812833		0,47	0,96	rs11230563	CD6	0,33	0,92	1,08	1,03	4,40×10 ⁻³
rs11741861	ZNF300	0,10	1,22	rs4845604	RORC	0,13	0,88	0,90	1,04	4,42×10 ⁻³
rs4743820	LINC00484	0,28	0,94	rs2538470		0,38	1,07	0,93	1,03	4,43×10 ⁻³
rs11064881	CIT	0,08	1,10	rs11010067		0,38	1,11	0,90	1,04	4,46×10 ⁻³
rs3740415	MFSD13A	0,45	0,95	rs12199775		0,06	0,90	0,89	1,04	4,56×10 ⁻³
rs10781499	CARD9	0,46	1,17	rs921720	LOC105375746	0,37	0,92	0,93	1,03	4,57×10 ⁻³
rs10065637	ANKRD55	0,20	0,93	rs3172494	IP6K2	0,12	1,09	1,09	1,03	4,79×10 ⁻³
rs7011507		0,12	0,92	rs11583043	DPH5	0,28	1,06	1,08	1,03	4,88×10 ⁻³
rs10051722		0,29	0,94	rs3749171	GPR35	0,19	1,12	1,07	1,03	4,98×10 ⁻³

Supplementary Table 2: G×G interactions between SNPs 30 kb upstream or downstream from IL27 and KPNA7 gene regions.

ID: SNP rs number; MAF: minor allele frequency; OR: interaction odds ratio; SE: standard error of interaction odds ratio; P: nominal p value of meta-analysis Wald test.

SNP-1			SNP-2			Interaction		
ID	Gene	MAF	ID	Gene	MAF	OR	SE	P
rs153109	IL27	0.46	rs9297145	KPNA7	0.27	1.18	0.04	6.06E-06
rs26528	IL27	0.46	rs9297145	KPNA7	0.27	1.18	0.04	6.39E-06
rs40837	IL27	0.46	rs9297145	KPNA7	0.27	1.17	0.04	1.13E-05
rs40836	IL27	0.46	rs9297145	KPNA7	0.27	1.17	0.04	1.57E-05
rs40834	IL27	0.46	rs9297145	KPNA7	0.27	1.17	0.04	1.71E-05
rs28698667	IL27	0.42	rs9297145	KPNA7	0.27	1.17	0.04	1.76E-05
rs4787458	IL27	0.38	rs9297145	KPNA7	0.27	1.16	0.04	2.67E-05
rs56354901	IL27	0.34	rs9297145	KPNA7	0.27	1.16	0.03	2.90E-05
rs181209	IL27	0.34	rs9297145	KPNA7	0.27	1.15	0.03	4.90E-05
rs151301	IL27	0.41	rs9297145	KPNA7	0.27	1.15	0.04	5.89E-05
rs240702	IL27	0.42	rs9297145	KPNA7	0.27	1.16	0.04	6.71E-05
rs34835	IL27	0.42	rs9297145	KPNA7	0.27	1.15	0.04	6.78E-05
rs153106	IL27	0.42	rs9297145	KPNA7	0.27	1.16	0.04	6.89E-05
rs181206	IL27	0.33	rs9297145	KPNA7	0.27	1.14	0.03	7.63E-05
rs62034325	IL27	0.38	rs9297145	KPNA7	0.27	1.15	0.04	7.87E-05
rs151181	IL27	0.41	rs9297145	KPNA7	0.27	1.15	0.04	8.55E-05
rs181207	IL27	0.34	rs9297145	KPNA7	0.27	1.14	0.03	9.33E-05
rs62034319	IL27	0.42	rs9297145	KPNA7	0.27	1.16	0.04	9.77E-05
rs4788083	IL27	0.42	rs9297145	KPNA7	0.27	1.16	0.04	9.79E-05
rs4788084	IL27	0.42	rs9297145	KPNA7	0.27	1.15	0.04	1.02E-04
rs62034323	IL27	0.42	rs9297145	KPNA7	0.27	1.15	0.04	1.17E-04
rs2106480	IL27	0.42	rs9297145	KPNA7	0.27	1.15	0.04	1.23E-04
rs4788085	IL27	0.42	rs9297145	KPNA7	0.27	1.15	0.04	1.30E-04
rs28772958	IL27	0.42	rs9297145	KPNA7	0.27	1.15	0.04	1.61E-04
rs180743	IL27	0.41	rs9297145	KPNA7	0.27	1.15	0.04	1.92E-04
rs149271	IL27	0.41	rs9297145	KPNA7	0.27	1.15	0.04	2.00E-04
rs151174	IL27	0.41	rs9297145	KPNA7	0.27	1.15	0.04	2.21E-04
rs180744	IL27	0.41	rs9297145	KPNA7	0.27	1.15	0.04	2.34E-04
rs151179	IL27	0.41	rs9297145	KPNA7	0.27	1.14	0.04	3.46E-04
rs4787458	IL27	0.38	rs4729516	KPNA7	0.37	1.14	0.04	4.44E-04
rs1074631	IL27	0.37	rs9297145	KPNA7	0.27	1.13	0.04	6.92E-04
rs3785354	IL27	0.37	rs9297145	KPNA7	0.27	1.13	0.04	7.79E-04
rs26528	IL27	0.46	rs4729516	KPNA7	0.37	1.14	0.04	9.30E-04
rs153109	IL27	0.46	rs4729516	KPNA7	0.37	1.14	0.04	1.04E-03
rs62034325	IL27	0.38	rs4729516	KPNA7	0.37	1.13	0.04	1.29E-03
rs151181	IL27	0.41	rs4729516	KPNA7	0.37	1.14	0.04	1.29E-03
rs153106	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	1.34E-03
rs240702	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	1.43E-03
rs151301	IL27	0.41	rs4729516	KPNA7	0.37	1.13	0.04	1.61E-03
rs62034319	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	1.62E-03
rs40837	IL27	0.46	rs4729516	KPNA7	0.37	1.13	0.04	1.63E-03
rs181207	IL27	0.34	rs4729516	KPNA7	0.37	1.12	0.03	1.70E-03
rs40836	IL27	0.46	rs4729516	KPNA7	0.37	1.13	0.04	1.76E-03

SNP-1			SNP-2			Interaction		
ID	Gene	MAF	ID	Gene	MAF	OR	SE	P
rs40834	IL27	0.46	rs4729516	KPNA7	0.37	1.13	0.04	1.81E-03
rs4788084	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	1.87E-03
rs4788083	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	1.97E-03
rs26528	IL27	0.46	rs10953281	KPNA7	0.37	1.12	0.04	1.98E-03
rs34835	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	2.13E-03
rs153109	IL27	0.46	rs10953281	KPNA7	0.37	1.12	0.04	2.14E-03
rs2106480	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	2.16E-03
rs62034323	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	2.22E-03
rs4788085	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	2.33E-03
rs28772958	IL27	0.42	rs4729516	KPNA7	0.37	1.13	0.04	2.35E-03
rs4787458	IL27	0.38	rs10953281	KPNA7	0.37	1.11	0.03	2.35E-03
rs181209	IL27	0.34	rs4729516	KPNA7	0.37	1.11	0.03	2.36E-03
rs56354901	IL27	0.34	rs4729516	KPNA7	0.37	1.12	0.04	2.36E-03
rs71389552	IL27	0.40	rs9297145	KPNA7	0.27	0.90	0.03	2.38E-03
rs231976	IL27	0.45	rs9297145	KPNA7	0.27	0.90	0.03	2.38E-03
rs28698667	IL27	0.42	rs10953281	KPNA7	0.37	1.12	0.04	2.65E-03
rs28698667	IL27	0.42	rs4729516	KPNA7	0.37	1.12	0.04	2.88E-03
rs40837	IL27	0.46	rs10953281	KPNA7	0.37	1.11	0.04	3.28E-03
rs75227850	IL27	0.04	rs2395022	KPNA7	0.05	1.34	0.10	3.47E-03
rs40836	IL27	0.46	rs10953281	KPNA7	0.37	1.11	0.04	3.49E-03
rs149271	IL27	0.41	rs4729516	KPNA7	0.37	1.12	0.04	3.63E-03
rs180743	IL27	0.41	rs4729516	KPNA7	0.37	1.12	0.04	3.84E-03
rs180744	IL27	0.41	rs4729516	KPNA7	0.37	1.12	0.04	3.92E-03
rs40834	IL27	0.46	rs10953281	KPNA7	0.37	1.11	0.04	3.96E-03
rs151174	IL27	0.41	rs4729516	KPNA7	0.37	1.12	0.04	4.58E-03
rs181206	IL27	0.33	rs4729516	KPNA7	0.37	1.10	0.03	4.63E-03
rs151181	IL27	0.41	rs10953281	KPNA7	0.37	1.11	0.04	4.68E-03
rs62034325	IL27	0.38	rs10953281	KPNA7	0.37	1.10	0.03	4.74E-03
rs151179	IL27	0.41	rs4729516	KPNA7	0.37	1.12	0.04	4.94E-03
rs1074631	IL27	0.37	rs4729516	KPNA7	0.37	1.11	0.04	4.95E-03
rs240702	IL27	0.42	rs10953281	KPNA7	0.37	1.11	0.04	5.07E-03
rs153106	IL27	0.42	rs10953281	KPNA7	0.37	1.11	0.04	5.09E-03
rs151301	IL27	0.41	rs10953281	KPNA7	0.37	1.11	0.04	5.20E-03
rs3785354	IL27	0.37	rs4729516	KPNA7	0.37	1.11	0.04	5.49E-03
rs117488909	IL27	0.01	rs2395022	KPNA7	0.05	1.51	0.15	6.20E-03
rs62034319	IL27	0.42	rs10953281	KPNA7	0.37	1.11	0.04	6.30E-03
rs4788084	IL27	0.42	rs10953281	KPNA7	0.37	1.10	0.04	6.68E-03
rs4788083	IL27	0.42	rs10953281	KPNA7	0.37	1.11	0.04	6.75E-03
rs34835	IL27	0.42	rs10953281	KPNA7	0.37	1.11	0.04	6.94E-03
rs4788085	IL27	0.42	rs10953281	KPNA7	0.37	1.10	0.04	7.23E-03
rs2106480	IL27	0.42	rs10953281	KPNA7	0.37	1.10	0.04	7.42E-03
rs2034874	IL27	0.04	rs2395022	KPNA7	0.05	1.33	0.11	7.43E-03
rs72791888	IL27	0.02	rs10274646	KPNA7	0.12	1.30	0.10	7.56E-03
rs62034323	IL27	0.42	rs10953281	KPNA7	0.37	1.10	0.04	7.68E-03
rs1074629	IL27	0.04	rs2395022	KPNA7	0.05	1.32	0.10	7.71E-03
rs28772958	IL27	0.42	rs10953281	KPNA7	0.37	1.10	0.04	8.04E-03
rs80344130	IL27	0.04	rs2395022	KPNA7	0.05	1.32	0.10	8.26E-03
rs1074631	IL27	0.37	rs10953281	KPNA7	0.37	1.09	0.03	9.33E-03

SNP-1			SNP-2			Interaction		
ID	Gene	MAF	ID	Gene	MAF	OR	SE	P
rs3785354	IL27	0.37	rs10953281	KPNA7	0.37	1.09	0.03	9.60E-03
rs76764777	IL27	0.04	rs2395022	KPNA7	0.05	1.31	0.10	9.94E-03
rs111720341	IL27	0.07	rs9297145	KPNA7	0.27	1.12	0.05	1.21E-02
rs34833	IL27	0.08	rs2395022	KPNA7	0.05	1.21	0.08	1.30E-02
rs40833	IL27	0.49	rs9297145	KPNA7	0.27	0.91	0.04	1.34E-02
rs149271	IL27	0.41	rs10953281	KPNA7	0.37	1.10	0.04	1.44E-02
rs180744	IL27	0.41	rs10953281	KPNA7	0.37	1.10	0.04	1.48E-02
rs180743	IL27	0.41	rs10953281	KPNA7	0.37	1.10	0.04	1.57E-02
rs151174	IL27	0.41	rs10953281	KPNA7	0.37	1.10	0.04	1.59E-02
rs71389552	IL27	0.40	rs4729516	KPNA7	0.37	0.92	0.04	1.63E-02
rs231976	IL27	0.45	rs4729516	KPNA7	0.37	0.92	0.04	1.65E-02
rs56354901	IL27	0.34	rs10953281	KPNA7	0.37	1.08	0.03	1.65E-02
rs181207	IL27	0.34	rs10953281	KPNA7	0.37	1.08	0.03	1.72E-02
rs151179	IL27	0.41	rs10953281	KPNA7	0.37	1.10	0.04	1.76E-02
rs181209	IL27	0.34	rs10953281	KPNA7	0.37	1.08	0.03	1.91E-02
rs55690101	IL27	0.01	rs10274646	KPNA7	0.12	1.45	0.16	2.22E-02
rs231976	IL27	0.45	rs10953281	KPNA7	0.37	0.92	0.04	2.69E-02
rs55690101	IL27	0.01	rs4729516	KPNA7	0.37	0.75	0.13	2.70E-02
rs153107	IL27	0.12	rs2395022	KPNA7	0.05	1.15	0.06	2.90E-02
rs79317729	IL27	0.04	rs2395022	KPNA7	0.05	1.24	0.10	2.96E-02
rs61738759	IL27	0.06	rs9297145	KPNA7	0.27	1.12	0.05	3.00E-02
rs111720341	IL27	0.07	rs10953281	KPNA7	0.37	1.11	0.05	3.54E-02
rs61738759	IL27	0.06	rs10953281	KPNA7	0.37	1.12	0.05	3.57E-02
rs181206	IL27	0.33	rs10953281	KPNA7	0.37	1.07	0.03	3.74E-02
rs40835	IL27	0.07	rs2395022	KPNA7	0.05	1.18	0.08	3.97E-02
rs75738775	IL27	0.04	rs2395022	KPNA7	0.05	1.22	0.10	3.99E-02
rs55690101	IL27	0.01	rs10953281	KPNA7	0.37	0.78	0.12	4.25E-02
rs17855750	IL27	0.04	rs9297145	KPNA7	0.27	1.12	0.06	4.53E-02
rs71389552	IL27	0.40	rs10953281	KPNA7	0.37	0.93	0.03	4.70E-02
rs78510287	IL27	0.00	rs4729516	KPNA7	0.37	0.60	0.26	4.96E-02
rs72791888	IL27	0.02	rs2395022	KPNA7	0.05	1.28	0.13	5.23E-02
rs40833	IL27	0.49	rs4729516	KPNA7	0.37	0.92	0.04	5.49E-02
rs111720341	IL27	0.07	rs4729516	KPNA7	0.37	1.10	0.05	5.50E-02
rs17855750	IL27	0.04	rs2395022	KPNA7	0.05	1.21	0.10	5.76E-02
rs78510287	IL27	0.00	rs2395022	KPNA7	0.05	2.04	0.38	6.09E-02
rs61738759	IL27	0.06	rs4729516	KPNA7	0.37	1.11	0.05	6.28E-02
rs79317729	IL27	0.04	rs9297145	KPNA7	0.27	1.11	0.06	6.45E-02
rs45542432	IL27	0.00	rs2395022	KPNA7	0.05	3.33	0.66	7.03E-02
rs75738775	IL27	0.04	rs9297145	KPNA7	0.27	1.11	0.06	7.76E-02
rs79077329	IL27	0.03	rs4729516	KPNA7	0.37	0.87	0.08	8.29E-02
rs79077329	IL27	0.03	rs10953281	KPNA7	0.37	0.87	0.08	8.31E-02
rs34395441	IL27	0.05	rs9297145	KPNA7	0.27	0.91	0.06	8.42E-02
rs40833	IL27	0.49	rs10953281	KPNA7	0.37	0.93	0.04	8.60E-02
rs118134830	IL27	0.04	rs2395022	KPNA7	0.05	1.19	0.10	8.70E-02
rs76007774	IL27	0.04	rs2395022	KPNA7	0.05	1.20	0.11	9.13E-02
rs76662516	IL27	0.03	rs10953281	KPNA7	0.37	0.87	0.08	9.16E-02
rs76007774	IL27	0.04	rs9297145	KPNA7	0.27	1.10	0.06	9.26E-02
rs76662516	IL27	0.03	rs4729516	KPNA7	0.37	0.87	0.08	9.32E-02

SNP-1			SNP-2			Interaction		
ID	Gene	MAF	ID	Gene	MAF	OR	SE	P
rs118134830	IL27	0.04	rs9297145	KPNA7	0.27	1.10	0.06	9.47E-02
rs117488909	IL27	0.01	rs4729516	KPNA7	0.37	0.84	0.10	9.92E-02
rs55690101	IL27	0.01	rs2395022	KPNA7	0.05	1.40	0.21	1.12E-01
rs117488909	IL27	0.01	rs10953281	KPNA7	0.37	0.86	0.10	1.28E-01
rs40833	IL27	0.49	rs2395022	KPNA7	0.05	0.92	0.06	1.45E-01
rs78510287	IL27	0.00	rs10274646	KPNA7	0.12	1.43	0.25	1.55E-01
rs117481765	IL27	0.03	rs2395022	KPNA7	0.05	1.18	0.12	1.56E-01
rs79077329	IL27	0.03	rs9297145	KPNA7	0.27	0.90	0.07	1.56E-01
rs34395441	IL27	0.05	rs10953281	KPNA7	0.37	0.92	0.06	1.57E-01
rs231976	IL27	0.45	rs2395022	KPNA7	0.05	0.92	0.06	1.69E-01
rs116954433	IL27	0.00	rs10274646	KPNA7	0.12	3.50	0.92	1.74E-01
rs118134830	IL27	0.04	rs10274646	KPNA7	0.12	0.90	0.08	1.75E-01
rs76007774	IL27	0.04	rs10274646	KPNA7	0.12	0.90	0.08	1.80E-01
rs153107	IL27	0.12	rs10274646	KPNA7	0.12	0.93	0.06	1.89E-01
rs76662516	IL27	0.03	rs9297145	KPNA7	0.27	0.91	0.07	1.90E-01
rs78510287	IL27	0.00	rs10953281	KPNA7	0.37	0.72	0.25	1.96E-01
rs28698667	IL27	0.42	rs2395022	KPNA7	0.05	1.08	0.06	2.02E-01
rs76764777	IL27	0.04	rs10953281	KPNA7	0.37	0.92	0.06	2.03E-01
rs117481765	IL27	0.03	rs9297145	KPNA7	0.27	0.92	0.07	2.08E-01
rs40836	IL27	0.46	rs10274646	KPNA7	0.12	0.94	0.05	2.20E-01
rs80344130	IL27	0.04	rs10953281	KPNA7	0.37	0.92	0.06	2.22E-01
rs76764777	IL27	0.04	rs4729516	KPNA7	0.37	0.92	0.07	2.28E-01
rs117481765	IL27	0.03	rs10953281	KPNA7	0.37	0.92	0.07	2.36E-01
rs40837	IL27	0.46	rs10274646	KPNA7	0.12	0.94	0.05	2.38E-01
rs34395441	IL27	0.05	rs2395022	KPNA7	0.05	1.13	0.11	2.42E-01
rs34395441	IL27	0.05	rs4729516	KPNA7	0.37	0.93	0.06	2.49E-01
rs80344130	IL27	0.04	rs4729516	KPNA7	0.37	0.93	0.07	2.50E-01
rs45542432	IL27	0.00	rs10274646	KPNA7	0.12	1.64	0.43	2.53E-01
rs151233	IL27	0.14	rs10274646	KPNA7	0.12	0.95	0.04	2.57E-01
rs153107	IL27	0.12	rs9297145	KPNA7	0.27	1.04	0.04	2.57E-01
rs34835	IL27	0.42	rs2395022	KPNA7	0.05	1.07	0.06	2.58E-01
rs40834	IL27	0.46	rs10274646	KPNA7	0.12	0.94	0.05	2.59E-01
rs45542432	IL27	0.00	rs10953281	KPNA7	0.37	0.64	0.41	2.67E-01
rs79077329	IL27	0.03	rs2395022	KPNA7	0.05	1.16	0.13	2.68E-01
rs76662516	IL27	0.03	rs2395022	KPNA7	0.05	1.15	0.13	2.84E-01
rs117488909	IL27	0.01	rs10274646	KPNA7	0.12	1.13	0.12	2.86E-01
rs153109	IL27	0.46	rs10274646	KPNA7	0.12	0.95	0.05	2.86E-01
rs117481765	IL27	0.03	rs4729516	KPNA7	0.37	0.93	0.07	2.91E-01
rs17855750	IL27	0.04	rs10953281	KPNA7	0.37	1.07	0.06	2.91E-01
rs45542432	IL27	0.00	rs4729516	KPNA7	0.37	0.65	0.41	2.96E-01
rs1074629	IL27	0.04	rs10953281	KPNA7	0.37	0.94	0.06	3.03E-01
rs26528	IL27	0.46	rs10274646	KPNA7	0.12	0.95	0.05	3.05E-01
rs111720341	IL27	0.07	rs2395022	KPNA7	0.05	1.08	0.08	3.09E-01
rs151234	IL27	0.14	rs10274646	KPNA7	0.12	0.96	0.04	3.09E-01
rs75738775	IL27	0.04	rs10274646	KPNA7	0.12	0.91	0.09	3.22E-01
rs151227	IL27	0.13	rs10274646	KPNA7	0.12	0.95	0.05	3.23E-01
rs1074629	IL27	0.04	rs4729516	KPNA7	0.37	0.94	0.07	3.25E-01
rs40833	IL27	0.49	rs10274646	KPNA7	0.12	1.04	0.04	3.26E-01

SNP-1			SNP-2			Interaction		
ID	Gene	MAF	ID	Gene	MAF	OR	SE	P
rs76764777	IL27	0.04	rs9297145	KPNA7	0.27	0.94	0.06	3.31E-01
rs75227850	IL27	0.04	rs9297145	KPNA7	0.27	1.06	0.06	3.33E-01
rs79317729	IL27	0.04	rs10953281	KPNA7	0.37	1.06	0.06	3.33E-01
rs117488909	IL27	0.01	rs9297145	KPNA7	0.27	0.91	0.10	3.37E-01
rs2034874	IL27	0.04	rs10953281	KPNA7	0.37	0.94	0.06	3.38E-01
rs40835	IL27	0.07	rs10274646	KPNA7	0.12	0.94	0.07	3.39E-01
rs2034874	IL27	0.04	rs4729516	KPNA7	0.37	0.94	0.07	3.62E-01
rs180743	IL27	0.41	rs2395022	KPNA7	0.05	1.05	0.06	3.72E-01
rs180744	IL27	0.41	rs10274646	KPNA7	0.12	0.96	0.05	3.75E-01
rs80344130	IL27	0.04	rs9297145	KPNA7	0.27	0.95	0.06	3.76E-01
rs231974	IL27	0.13	rs10274646	KPNA7	0.12	0.96	0.05	3.85E-01
rs151174	IL27	0.41	rs10274646	KPNA7	0.12	0.96	0.05	3.92E-01
rs17855750	IL27	0.04	rs10274646	KPNA7	0.12	0.93	0.08	4.00E-01
rs180743	IL27	0.41	rs10274646	KPNA7	0.12	0.96	0.05	4.05E-01
rs79317729	IL27	0.04	rs10274646	KPNA7	0.12	0.93	0.08	4.12E-01
rs26528	IL27	0.46	rs2395022	KPNA7	0.05	1.05	0.06	4.14E-01
rs62034323	IL27	0.42	rs10274646	KPNA7	0.12	0.96	0.05	4.18E-01
rs40835	IL27	0.07	rs9297145	KPNA7	0.27	1.04	0.05	4.22E-01
rs75227850	IL27	0.04	rs10953281	KPNA7	0.37	1.05	0.07	4.23E-01
rs4788084	IL27	0.42	rs10274646	KPNA7	0.12	0.96	0.05	4.27E-01
rs151174	IL27	0.41	rs2395022	KPNA7	0.05	1.05	0.06	4.27E-01
rs153106	IL27	0.42	rs10274646	KPNA7	0.12	0.96	0.05	4.28E-01
rs28698667	IL27	0.42	rs10274646	KPNA7	0.12	0.96	0.05	4.28E-01
rs153109	IL27	0.46	rs2395022	KPNA7	0.05	1.05	0.06	4.31E-01
rs75738775	IL27	0.04	rs10953281	KPNA7	0.37	1.05	0.06	4.34E-01
rs4788083	IL27	0.42	rs10274646	KPNA7	0.12	0.96	0.05	4.36E-01
rs151301	IL27	0.41	rs2395022	KPNA7	0.05	1.04	0.06	4.39E-01
rs240702	IL27	0.42	rs10274646	KPNA7	0.12	0.97	0.05	4.40E-01
rs40836	IL27	0.46	rs2395022	KPNA7	0.05	1.05	0.06	4.40E-01
rs153107	IL27	0.12	rs10953281	KPNA7	0.37	1.03	0.04	4.42E-01
rs40837	IL27	0.46	rs2395022	KPNA7	0.05	1.05	0.06	4.43E-01
rs62034319	IL27	0.42	rs10274646	KPNA7	0.12	0.97	0.05	4.43E-01
rs151181	IL27	0.41	rs2395022	KPNA7	0.05	1.04	0.06	4.44E-01
rs1074629	IL27	0.04	rs9297145	KPNA7	0.27	0.95	0.06	4.54E-01
rs149271	IL27	0.41	rs10274646	KPNA7	0.12	0.97	0.05	4.61E-01
rs4788085	IL27	0.42	rs10274646	KPNA7	0.12	0.97	0.05	4.62E-01
rs151227	IL27	0.13	rs4729516	KPNA7	0.37	0.97	0.05	4.66E-01
rs28772958	IL27	0.42	rs10274646	KPNA7	0.12	0.97	0.05	4.70E-01
rs180744	IL27	0.41	rs2395022	KPNA7	0.05	1.04	0.06	4.71E-01
rs149271	IL27	0.41	rs2395022	KPNA7	0.05	1.04	0.06	4.71E-01
rs71389552	IL27	0.40	rs10274646	KPNA7	0.12	1.03	0.04	4.75E-01
rs40834	IL27	0.46	rs2395022	KPNA7	0.05	1.04	0.06	4.78E-01
rs2106480	IL27	0.42	rs10274646	KPNA7	0.12	0.97	0.05	4.86E-01
rs17855750	IL27	0.04	rs4729516	KPNA7	0.37	1.05	0.07	4.90E-01
rs79317729	IL27	0.04	rs4729516	KPNA7	0.37	1.04	0.06	4.92E-01
rs2034874	IL27	0.04	rs9297145	KPNA7	0.27	0.96	0.06	4.95E-01
rs151227	IL27	0.13	rs10953281	KPNA7	0.37	0.97	0.04	4.96E-01
rs2034874	IL27	0.04	rs10274646	KPNA7	0.12	1.05	0.08	4.98E-01

SNP-1			SNP-2			Interaction		
ID	Gene	MAF	ID	Gene	MAF	OR	SE	P
rs62034319	IL27	0.42	rs2395022	KPNA7	0.05	1.04	0.06	4.99E-01
rs4788085	IL27	0.42	rs2395022	KPNA7	0.05	1.04	0.06	5.02E-01
rs4788084	IL27	0.42	rs2395022	KPNA7	0.05	1.04	0.06	5.03E-01
rs151181	IL27	0.41	rs10274646	KPNA7	0.12	0.97	0.05	5.06E-01
rs1074629	IL27	0.04	rs10274646	KPNA7	0.12	1.05	0.08	5.08E-01
rs71389552	IL27	0.40	rs2395022	KPNA7	0.05	0.96	0.05	5.09E-01
rs28772958	IL27	0.42	rs2395022	KPNA7	0.05	1.04	0.06	5.09E-01
rs153106	IL27	0.42	rs2395022	KPNA7	0.05	1.04	0.06	5.10E-01
rs3785354	IL27	0.37	rs2395022	KPNA7	0.05	1.04	0.06	5.16E-01
rs62034323	IL27	0.42	rs2395022	KPNA7	0.05	1.04	0.06	5.18E-01
rs4788083	IL27	0.42	rs2395022	KPNA7	0.05	1.04	0.06	5.21E-01
rs4787458	IL27	0.38	rs2395022	KPNA7	0.05	1.04	0.06	5.25E-01
rs2106480	IL27	0.42	rs2395022	KPNA7	0.05	1.04	0.06	5.29E-01
rs240702	IL27	0.42	rs2395022	KPNA7	0.05	1.04	0.06	5.33E-01
rs72791888	IL27	0.02	rs9297145	KPNA7	0.27	1.05	0.08	5.34E-01
rs1074631	IL27	0.37	rs2395022	KPNA7	0.05	1.03	0.06	5.48E-01
rs151233	IL27	0.14	rs9297145	KPNA7	0.27	1.03	0.04	5.48E-01
rs76764777	IL27	0.04	rs10274646	KPNA7	0.12	1.05	0.08	5.52E-01
rs151301	IL27	0.41	rs10274646	KPNA7	0.12	0.97	0.05	5.56E-01
rs1074631	IL27	0.37	rs10274646	KPNA7	0.12	0.98	0.04	5.60E-01
rs76007774	IL27	0.04	rs10953281	KPNA7	0.37	1.04	0.06	5.63E-01
rs80344130	IL27	0.04	rs10274646	KPNA7	0.12	1.04	0.08	5.77E-01
rs3785354	IL27	0.37	rs10274646	KPNA7	0.12	0.98	0.04	5.78E-01
rs151234	IL27	0.14	rs9297145	KPNA7	0.27	1.02	0.04	5.84E-01
rs151179	IL27	0.41	rs2395022	KPNA7	0.05	1.03	0.06	5.93E-01
rs151227	IL27	0.13	rs2395022	KPNA7	0.05	1.03	0.06	6.08E-01
rs111720341	IL27	0.07	rs10274646	KPNA7	0.12	1.03	0.05	6.08E-01
rs153107	IL27	0.12	rs4729516	KPNA7	0.37	1.02	0.04	6.09E-01
rs76007774	IL27	0.04	rs4729516	KPNA7	0.37	1.03	0.06	6.10E-01
rs151179	IL27	0.41	rs10274646	KPNA7	0.12	0.98	0.05	6.14E-01
rs231974	IL27	0.13	rs10953281	KPNA7	0.37	0.98	0.04	6.18E-01
rs78510287	IL27	0.00	rs9297145	KPNA7	0.27	0.87	0.27	6.18E-01
rs62034325	IL27	0.38	rs2395022	KPNA7	0.05	1.03	0.06	6.24E-01
rs118134830	IL27	0.04	rs10953281	KPNA7	0.37	1.03	0.06	6.27E-01
rs72791888	IL27	0.02	rs10953281	KPNA7	0.37	1.05	0.10	6.27E-01
rs61738759	IL27	0.06	rs2395022	KPNA7	0.05	1.05	0.10	6.28E-01
rs75738775	IL27	0.04	rs4729516	KPNA7	0.37	1.03	0.06	6.45E-01
rs151234	IL27	0.14	rs2395022	KPNA7	0.05	0.97	0.06	6.59E-01
rs231974	IL27	0.13	rs2395022	KPNA7	0.05	1.03	0.06	6.62E-01
rs118134830	IL27	0.04	rs4729516	KPNA7	0.37	1.03	0.06	6.80E-01
rs181206	IL27	0.33	rs10274646	KPNA7	0.12	0.98	0.05	6.83E-01
rs56354901	IL27	0.34	rs10274646	KPNA7	0.12	0.98	0.04	7.03E-01
rs116954433	IL27	0.00	rs9297145	KPNA7	0.27	0.62	1.24	7.04E-01
rs231974	IL27	0.13	rs4729516	KPNA7	0.37	0.98	0.05	7.06E-01
rs34395441	IL27	0.05	rs10274646	KPNA7	0.12	0.97	0.07	7.11E-01
rs116954433	IL27	0.00	rs10953281	KPNA7	0.37	1.38	0.89	7.21E-01
rs116954433	IL27	0.00	rs4729516	KPNA7	0.37	1.56	1.28	7.26E-01
rs181207	IL27	0.34	rs10274646	KPNA7	0.12	0.98	0.05	7.30E-01

SNP-1			SNP-2			Interaction		
ID	Gene	MAF	ID	Gene	MAF	OR	SE	P
rs181209	IL27	0.34	rs10274646	KPNA7	0.12	0.98	0.05	7.38E-01
rs62034325	IL27	0.38	rs10274646	KPNA7	0.12	0.99	0.04	7.58E-01
rs231976	IL27	0.45	rs10274646	KPNA7	0.12	1.01	0.04	7.61E-01
rs72791888	IL27	0.02	rs4729516	KPNA7	0.37	1.03	0.11	7.64E-01
rs79077329	IL27	0.03	rs10274646	KPNA7	0.12	0.98	0.09	8.12E-01
rs40835	IL27	0.07	rs10953281	KPNA7	0.37	1.01	0.05	8.15E-01
rs34835	IL27	0.42	rs10274646	KPNA7	0.12	0.99	0.05	8.17E-01
rs231974	IL27	0.13	rs9297145	KPNA7	0.27	1.01	0.04	8.20E-01
rs151233	IL27	0.14	rs2395022	KPNA7	0.05	0.99	0.06	8.41E-01
rs4787458	IL27	0.38	rs10274646	KPNA7	0.12	0.99	0.04	8.41E-01
rs34833	IL27	0.08	rs10953281	KPNA7	0.37	1.01	0.05	8.56E-01
rs76662516	IL27	0.03	rs10274646	KPNA7	0.12	0.98	0.09	8.58E-01
rs34833	IL27	0.08	rs10274646	KPNA7	0.12	1.01	0.05	8.58E-01
rs151234	IL27	0.14	rs10953281	KPNA7	0.37	1.01	0.04	8.65E-01
rs75227850	IL27	0.04	rs4729516	KPNA7	0.37	1.01	0.08	8.77E-01
rs34833	IL27	0.08	rs4729516	KPNA7	0.37	0.99	0.05	8.80E-01
rs181206	IL27	0.33	rs2395022	KPNA7	0.05	0.99	0.06	9.07E-01
rs181209	IL27	0.34	rs2395022	KPNA7	0.05	0.99	0.06	9.11E-01
rs151234	IL27	0.14	rs4729516	KPNA7	0.37	1.00	0.05	9.25E-01
rs56354901	IL27	0.34	rs2395022	KPNA7	0.05	0.99	0.06	9.31E-01
rs151227	IL27	0.13	rs9297145	KPNA7	0.27	1.00	0.04	9.41E-01
rs117481765	IL27	0.03	rs10274646	KPNA7	0.12	1.01	0.08	9.46E-01
rs34833	IL27	0.08	rs9297145	KPNA7	0.27	1.00	0.04	9.47E-01
rs61738759	IL27	0.06	rs10274646	KPNA7	0.12	1.00	0.06	9.51E-01
rs151233	IL27	0.14	rs10953281	KPNA7	0.37	1.00	0.05	9.51E-01
rs75227850	IL27	0.04	rs10274646	KPNA7	0.12	1.00	0.08	9.54E-01
rs181207	IL27	0.34	rs2395022	KPNA7	0.05	1.00	0.06	9.61E-01
rs151233	IL27	0.14	rs4729516	KPNA7	0.37	1.00	0.05	9.76E-01
rs40835	IL27	0.07	rs4729516	KPNA7	0.37	1.00	0.05	9.81E-01
rs45542432	IL27	0.00	rs9297145	KPNA7	0.27	1.00	0.48	9.93E-01

Supplement 3a: G×G interactions between SNPs associated with CD on the same chromosome arm in case control design.

ID: SNP rs number; MAF: minor allele frequency; OR: odds ratio; SE: standard error of interaction odds ratio; P: nominal p value of meta-analysis Wald test; \$: allelic disease odds ratio of minor allele, according to Liu et al. (2015); &: interaction odds ratio.

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs26528	IL27	0.46	1.13	rs11641184	LITAF	0.49	1.08	1.31	0.08	1.54E-03
rs10781499	CARD9	0.44	1.18	rs4743820	LINC00484	0.29	0.95	0.80	0.07	1.96E-03
rs2382817	TMBIM1	0.41	1.07	rs1517352	STAT4	0.39	0.92	1.20	0.07	1.32E-02
rs212388	LOC105378083	0.41	1.11	rs2503322	RSPO3	0.45	0.94	1.21	0.08	1.60E-02
rs6556412	LOC285626	0.35	1.17	rs10051722		0.29	0.91	1.16	0.07	2.68E-02
rs10185424		0.47	1.07	rs925255	FOSL2	0.43	0.90	0.84	0.08	2.87E-02
rs1505992		0.30	0.82	rs1842076		0.27	0.90	1.16	0.07	2.99E-02
rs6716753	SP140	0.20	1.14	rs13407913	ADCY3	0.45	1.12	0.85	0.08	3.00E-02
rs9847710	SFMBT1	0.41	0.99	rs9868809	CELSR3	0.11	1.12	1.22	0.09	3.01E-02
rs7097656	TSPAN14	0.19	0.88	rs224090		0.43	1.14	1.18	0.08	3.17E-02
rs564349	ERGIC1	0.32	1.05	rs10065637	ANKRD55	0.21	0.91	1.16	0.07	3.25E-02
rs1292053	TUBD1	0.46	1.10	rs12946510		0.48	1.13	1.20	0.09	3.39E-02
rs9313808		0.16	0.86	rs254560	C5orf66	0.41	1.03	0.81	0.10	3.47E-02
rs1819333		0.45	0.89	rs7746082		0.30	1.14	0.84	0.09	3.50E-02
rs3740415	MFSD13A	0.46	0.95	rs224090		0.43	1.14	0.85	0.08	3.52E-02
rs2641348	ADAM30	0.10	0.89	rs3766606	PARK7	0.16	0.90	0.82	0.10	4.10E-02
rs2155219		0.47	1.19	rs559928		0.17	0.91	0.85	0.08	4.37E-02
rs3764147	LACC1	0.25	1.15	rs915286		0.44	0.94	1.16	0.07	4.57E-02
rs72810983	CPEB4	0.29	0.91	rs564349	ERGIC1	0.32	1.05	1.14	0.07	4.99E-02
rs3742130	GPR18	0.21	0.91	rs915286		0.44	0.94	1.16	0.08	5.00E-02
rs6716753	SP140	0.20	1.14	rs7608910	PUS10	0.41	1.13	1.16	0.08	5.25E-02
rs12994997	ATG16L1	0.44	0.80	rs11681525	TEX41	0.08	0.86	1.21	0.10	5.26E-02
rs17622378	C5orf56	0.45	1.21	rs10051722		0.29	0.91	0.87	0.08	5.83E-02
rs10865331		0.39	1.08	rs1260326	GCKR	0.44	1.12	1.21	0.10	5.92E-02
rs12718244	SPATA48	0.42	1.08	rs10486483	SKAP2	0.25	1.10	0.85	0.09	6.42E-02
rs4802307	PPP5C	0.29	0.91	rs17694108		0.29	1.08	0.88	0.07	7.07E-02
rs26528	IL27	0.46	1.13	rs7404095	PRKCB	0.42	0.96	0.87	0.08	7.08E-02

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs212388	LOC105378083	0.41	1.11	rs7758080	TAB2	0.29	1.08	1.14	0.07	7.12E-02
rs7517810		0.25	1.14	rs4845604	RORC	0.14	0.91	0.87	0.08	7.15E-02
rs6716753	SP140	0.20	1.14	rs6740462		0.25	0.91	0.88	0.07	7.20E-02
rs17119	LOC101928354	0.19	0.92	rs13204048	SLC22A23	0.37	0.93	0.88	0.07	7.27E-02
rs7554511	C1orf106	0.26	0.86	rs4845604	RORC	0.14	0.91	0.87	0.08	7.31E-02
rs224090		0.43	1.14	rs10761659		0.44	0.83	1.17	0.09	7.63E-02
rs11150589	ITGAL	0.47	1.02	rs11641184	LITAF	0.49	1.08	1.16	0.08	8.04E-02
rs1842076		0.27	0.90	rs395157	OSMR	0.50	1.10	1.14	0.08	8.36E-02
rs6588248	IL23R	0.45	0.88	rs12103	INTS11	0.19	1.08	1.15	0.08	8.37E-02
rs2382817	TMBIM1	0.41	1.07	rs13407913	ADCY3	0.45	1.12	1.23	0.12	8.48E-02
rs56167332		0.36	1.19	rs4703855		0.29	0.93	0.89	0.07	8.74E-02
rs6708413	IL18RAP	0.25	1.12	rs6740462		0.25	0.91	0.89	0.07	9.08E-02
rs1292053	TUBD1	0.46	1.10	rs12942547	STAT3	0.40	0.90	0.88	0.08	9.13E-02
rs2641348	ADAM30	0.10	0.89	rs6679677	PHTF1	0.09	0.83	0.81	0.12	9.17E-02
rs12627970		0.21	1.12	rs2413583		0.15	0.81	0.87	0.09	9.18E-02
rs8005161	GPR65	0.09	1.17	rs1569328		0.16	0.90	0.85	0.10	9.49E-02
rs12994997	ATG16L1	0.44	0.80	rs6740462		0.25	0.91	0.88	0.07	9.79E-02
rs9358372	CDKAL1	0.38	1.08	rs13204048	SLC22A23	0.37	0.93	0.85	0.10	9.83E-02
rs6679677	PHTF1	0.09	0.83	rs17391694		0.11	0.89	1.21	0.12	1.03E-01
rs7554511	C1orf106	0.26	0.86	rs670523	RIT1	0.34	1.08	0.89	0.07	1.04E-01
rs2503322	RSP03	0.45	0.94	rs1847472	BACH2	0.34	0.92	1.17	0.10	1.06E-01
rs7608910	PUS10	0.41	1.13	rs13407913	ADCY3	0.45	1.12	0.88	0.08	1.08E-01
rs516246	FUT2	0.48	1.12	rs17694108		0.29	1.08	0.86	0.09	1.09E-01
rs1819333		0.45	0.89	rs12199775		0.06	0.89	0.84	0.11	1.10E-01
rs13204742		0.13	1.15	rs1847472	BACH2	0.34	0.92	0.88	0.08	1.11E-01
rs6863411	NDFIP1	0.36	0.91	rs1363907	ERAP2	0.43	1.11	0.88	0.08	1.14E-01
rs2641348	ADAM30	0.10	0.89	rs1748195	DOCK7	0.33	1.07	0.87	0.09	1.14E-01
rs9313808		0.16	0.86	rs6556412	LOC285626	0.35	1.17	1.18	0.11	1.15E-01
rs6740462		0.25	0.91	rs13407913	ADCY3	0.45	1.12	0.89	0.07	1.16E-01
rs10185424		0.47	1.07	rs10865331		0.39	1.08	1.13	0.08	1.18E-01
rs653178	ATXN2	0.49	1.07	rs7134472	LOC107984526	0.39	1.06	1.13	0.08	1.21E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs254560	C5orf66	0.41	1.03	rs10061469		0.31	0.92	1.11	0.07	1.25E-01
rs7608910	PUS10	0.41	1.13	rs1260326	GCKR	0.44	1.12	1.13	0.08	1.25E-01
rs9319943		0.20	1.08	rs7240004		0.37	0.95	0.85	0.11	1.25E-01
rs4409764		0.49	1.19	rs12778642		0.44	1.04	0.88	0.08	1.26E-01
rs3742130	GPR18	0.21	0.91	rs941823	LINC00598	0.24	0.94	1.11	0.07	1.30E-01
rs17780256	SLC39A11	0.19	0.95	rs12946510		0.48	1.13	1.18	0.11	1.31E-01
rs1748195	DOCK7	0.33	1.07	rs3766606	PARK7	0.16	0.90	1.12	0.08	1.35E-01
rs12627970		0.21	1.12	rs2256609	UBE2L3	0.21	1.11	0.90	0.07	1.37E-01
rs2651244		0.40	1.05	rs12103	INTS11	0.19	1.08	0.89	0.07	1.37E-01
rs10781499	CARD9	0.44	1.18	rs11554257		0.14	1.16	1.13	0.08	1.37E-01
rs3749171	GPR35	0.18	1.08	rs6716753	SP140	0.20	1.14	0.89	0.08	1.40E-01
rs11742570		0.36	0.78	rs2930047	DAP	0.39	1.09	0.90	0.07	1.41E-01
rs12199775		0.06	0.89	rs13204742		0.13	1.15	0.84	0.12	1.43E-01
rs7657746	KIAA1109	0.24	0.92	rs13126505	BANK1	0.07	1.20	1.16	0.10	1.44E-01
rs254560	C5orf66	0.41	1.03	rs17622378	C5orf56	0.45	1.21	1.12	0.08	1.53E-01
rs3740415	MFSD13A	0.46	0.95	rs10761659		0.44	0.83	1.15	0.10	1.54E-01
rs564349	ERGIC1	0.32	1.05	rs254560	C5orf66	0.41	1.03	1.15	0.10	1.54E-01
rs12199775		0.06	0.89	rs3851228	TRAF3IP2-AS1	0.07	1.14	1.24	0.15	1.57E-01
rs1819333		0.45	0.89	rs13204742		0.13	1.15	1.13	0.09	1.58E-01
rs7236492	NFATC1	0.14	0.91	rs7240004		0.37	0.95	0.89	0.08	1.60E-01
rs10761659		0.44	0.83	rs1199103		0.21	0.90	1.14	0.09	1.62E-01
rs12942547	STAT3	0.40	0.90	rs3091315		0.26	0.87	0.91	0.07	1.64E-01
rs13204742		0.13	1.15	rs7746082		0.30	1.14	0.90	0.08	1.65E-01
rs1250546	ZMIZ1	0.40	0.90	rs224090		0.43	1.14	0.90	0.07	1.65E-01
rs3024505		0.17	1.18	rs7517810		0.25	1.14	1.11	0.07	1.66E-01
rs314313	EPHB4	0.31	1.07	rs9297145	KPNA7	0.27	1.08	0.86	0.11	1.66E-01
rs13300483		0.26	1.14	rs11554257		0.14	1.16	0.82	0.14	1.67E-01
rs11064881	CIT	0.08	1.10	rs4768236	LRRK2	0.35	1.12	1.14	0.10	1.67E-01
rs2503322	RSP03	0.45	0.94	rs3851228	TRAF3IP2-AS1	0.07	1.14	1.16	0.11	1.70E-01
rs12994997	ATG16L1	0.44	0.80	rs13407913	ADCY3	0.45	1.12	0.90	0.08	1.70E-01
rs7555082		0.12	1.13	rs4656958		0.30	0.94	0.86	0.11	1.70E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs7554511	C1orf106	0.26	0.86	rs7555082		0.12	1.13	0.83	0.14	1.73E-01
rs17391694		0.11	0.89	rs6588248	IL23R	0.45	0.88	0.88	0.10	1.75E-01
rs1517352	STAT4	0.39	0.92	rs1260326	GCKR	0.44	1.12	0.87	0.10	1.75E-01
rs6740462		0.25	0.91	rs1260326	GCKR	0.44	1.12	1.11	0.07	1.76E-01
rs254560	C5orf66	0.41	1.03	rs4703855		0.29	0.93	1.10	0.07	1.79E-01
rs3749171	GPR35	0.18	1.08	rs12994997	ATG16L1	0.44	0.80	1.14	0.09	1.79E-01
rs7517847	IL23R	0.38	0.71	rs1748195	DOCK7	0.33	1.07	0.91	0.07	1.79E-01
rs9313808		0.16	0.86	rs10065637	ANKRD55	0.21	0.91	0.90	0.08	1.79E-01
rs10865331		0.39	1.08	rs13407913	ADCY3	0.45	1.12	0.90	0.08	1.80E-01
rs564349	ERGIC1	0.32	1.05	rs10051722		0.29	0.91	1.13	0.09	1.82E-01
rs3740415	MFSD13A	0.46	0.95	rs12778642		0.44	1.04	0.87	0.10	1.82E-01
rs7097656	TSPAN14	0.19	0.88	rs1199103		0.21	0.90	0.87	0.11	1.84E-01
rs10185424		0.47	1.07	rs10495903	THADA	0.14	1.13	1.12	0.09	1.85E-01
rs3764147	LACC1	0.25	1.15	rs9525625	LINC02341	0.49	1.08	1.11	0.08	1.88E-01
rs6651252	LINC00824	0.12	0.86	rs7015630		0.25	0.92	0.86	0.11	1.89E-01
rs6708413	IL18RAP	0.25	1.12	rs1260326	GCKR	0.44	1.12	1.10	0.07	1.93E-01
rs564349	ERGIC1	0.32	1.05	rs11743851	CDC42SE2	0.39	1.15	0.90	0.08	1.94E-01
rs7554511	C1orf106	0.26	0.86	rs7517810		0.25	1.14	0.89	0.09	1.95E-01
rs17780256	SLC39A11	0.19	0.95	rs3091315		0.26	0.87	1.10	0.07	1.99E-01
rs7404095	PRKCB	0.42	0.96	rs11641184	LITAF	0.49	1.08	1.12	0.09	2.03E-01
rs13277237	CCDC26	0.44	1.05	rs6651252	LINC00824	0.12	0.86	0.88	0.10	2.06E-01
rs6679677	PHTF1	0.09	0.83	rs3766606	PARK7	0.16	0.90	1.14	0.10	2.07E-01
rs13300218	NOTCH1	0.10	0.85	rs11554257		0.14	1.16	1.13	0.10	2.07E-01
rs6740462		0.25	0.91	rs925255	FOSL2	0.43	0.90	0.91	0.07	2.13E-01
rs1748195	DOCK7	0.33	1.07	rs12103	INTS11	0.19	1.08	1.09	0.07	2.16E-01
rs7015630		0.25	0.92	rs7011507		0.12	0.93	1.10	0.08	2.22E-01
rs4692386		0.39	0.94	rs2073505	HGFAC	0.09	1.11	0.89	0.10	2.27E-01
rs2227551	PLAU	0.25	0.91	rs10761659		0.44	0.83	0.91	0.08	2.28E-01
rs2155219		0.47	1.19	rs11230563	CD6	0.34	0.92	0.91	0.08	2.32E-01
rs3749171	GPR35	0.18	1.08	rs10865331		0.39	1.08	0.91	0.08	2.35E-01
rs12994997	ATG16L1	0.44	0.80	rs7608910	PUS10	0.41	1.13	1.14	0.11	2.35E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs2155219		0.47	1.19	rs568617	FIBP	0.19	1.09	1.10	0.08	2.37E-01
rs72810983	CPEB4	0.29	0.91	rs10061469		0.31	0.92	1.08	0.07	2.37E-01
rs2361755		0.07	0.84	rs1728785	ZFP90	0.23	1.00	1.13	0.10	2.38E-01
rs566416		0.23	0.94	rs559928		0.17	0.91	1.09	0.07	2.49E-01
rs7438704	SLAIN2	0.34	0.92	rs6856616		0.07	1.10	0.89	0.10	2.49E-01
rs3853824	C17orf67	0.34	0.92	rs12946510		0.48	1.13	1.09	0.08	2.49E-01
rs11064881	CIT	0.08	1.10	rs653178	ATXN2	0.49	1.07	1.13	0.11	2.52E-01
rs17391694		0.11	0.89	rs2651244		0.40	1.05	0.90	0.09	2.54E-01
rs4409764		0.49	1.19	rs1250546	ZMIZ1	0.40	0.90	1.09	0.08	2.57E-01
rs3749171	GPR35	0.18	1.08	rs925255	FOSL2	0.43	0.90	1.10	0.09	2.57E-01
rs559928		0.17	0.91	rs174537	MYRF	0.34	1.09	1.09	0.08	2.58E-01
rs2382817	TMBIM1	0.41	1.07	rs10495903	THADA	0.14	1.13	0.91	0.08	2.59E-01
rs6679677	PHTF1	0.09	0.83	rs1748195	DOCK7	0.33	1.07	1.11	0.09	2.63E-01
rs2227551	PLAU	0.25	0.91	rs224090		0.43	1.14	1.08	0.07	2.69E-01
rs7097656	TSPAN14	0.19	0.88	rs10761659		0.44	0.83	1.09	0.08	2.69E-01
rs72810983	CPEB4	0.29	0.91	rs6556412	LOC285626	0.35	1.17	1.08	0.07	2.71E-01
rs6556412	LOC285626	0.35	1.17	rs1363907	ERAP2	0.43	1.11	0.91	0.08	2.72E-01
rs921720	LOC105375746	0.37	0.89	rs7015630		0.25	0.92	0.93	0.07	2.77E-01
rs10051722		0.29	0.91	rs1363907	ERAP2	0.43	1.11	0.91	0.09	2.79E-01
rs7555082		0.12	1.13	rs670523	RIT1	0.34	1.08	1.09	0.08	2.84E-01
rs6708413	IL18RAP	0.25	1.12	rs925255	FOSL2	0.43	0.90	0.93	0.07	2.91E-01
rs11554257		0.14	1.16	rs4246905	TNFSF15	0.27	0.87	1.10	0.09	2.93E-01
rs7758080	TAB2	0.29	1.08	rs12199775		0.06	0.89	1.12	0.11	2.96E-01
rs9847710	SFMBT1	0.41	0.99	rs3197999	MST1	0.30	1.17	0.93	0.07	2.96E-01
rs11681525	TEX41	0.08	0.86	rs13407913	ADCY3	0.45	1.12	0.89	0.12	2.99E-01
rs6856616		0.07	1.10	rs2073505	HGFAC	0.09	1.11	0.87	0.14	2.99E-01
rs10185424		0.47	1.07	rs7608910	PUS10	0.41	1.13	1.08	0.08	3.05E-01
rs12994997	ATG16L1	0.44	0.80	rs6708413	IL18RAP	0.25	1.12	1.10	0.09	3.06E-01
rs6708413	IL18RAP	0.25	1.12	rs13407913	ADCY3	0.45	1.12	0.93	0.07	3.08E-01
rs11741861	ZNF300	0.09	1.33	rs11743851	CDC42SE2	0.39	1.15	1.10	0.09	3.08E-01
rs6716753	SP140	0.20	1.14	rs6708413	IL18RAP	0.25	1.12	1.12	0.11	3.10E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs6679677	PHTF1	0.09	0.83	rs2651244		0.40	1.05	1.10	0.10	3.12E-01
rs10781499	CARD9	0.44	1.18	rs4246905	TNFSF15	0.27	0.87	1.08	0.07	3.15E-01
rs8005161	GPR65	0.09	1.17	rs194749		0.23	1.08	0.88	0.13	3.22E-01
rs2382817	TMBIM1	0.41	1.07	rs925255	FOSL2	0.43	0.90	0.93	0.08	3.24E-01
rs3749171	GPR35	0.18	1.08	rs1260326	GCKR	0.44	1.12	0.90	0.11	3.25E-01
rs3024505		0.17	1.18	rs10798069	PLA2G4A	0.48	0.93	0.90	0.11	3.25E-01
rs921720	LOC105375746	0.37	0.89	rs7011507		0.12	0.93	0.92	0.09	3.25E-01
rs6074022		0.27	1.10	rs4243971		0.44	0.95	0.92	0.09	3.27E-01
rs6908425	CDKAL1	0.20	0.90	rs17119	LOC101928354	0.19	0.92	0.92	0.08	3.27E-01
rs11741861	ZNF300	0.09	1.33	rs254560	C5orf66	0.41	1.03	0.87	0.14	3.28E-01
rs3740415	MFSD13A	0.46	0.95	rs1250546	ZMIZ1	0.40	0.90	1.08	0.08	3.31E-01
rs2836878		0.26	0.90	rs2284553	IFNGR2	0.39	0.90	0.93	0.07	3.32E-01
rs6863411	NDFIP1	0.36	0.91	rs10065637	ANKRD55	0.21	0.91	1.07	0.07	3.33E-01
rs9847710	SFMBT1	0.41	0.99	rs4256159	LOC105376976	0.15	1.13	0.92	0.08	3.35E-01
rs727563		0.22	1.10	rs2413583		0.15	0.81	1.08	0.08	3.36E-01
rs925255	FOSL2	0.43	0.90	rs13407913	ADCY3	0.45	1.12	1.10	0.10	3.39E-01
rs6651252	LINC00824	0.12	0.86	rs921720	LOC105375746	0.37	0.89	0.92	0.08	3.40E-01
rs913678		0.33	0.95	rs6074022		0.27	1.10	0.94	0.07	3.40E-01
rs3197999	MST1	0.30	1.17	rs4256159	LOC105376976	0.15	1.13	1.08	0.08	3.41E-01
rs1517352	STAT4	0.39	0.92	rs6708413	IL18RAP	0.25	1.12	1.09	0.09	3.43E-01
rs17622378	C5orf56	0.45	1.21	rs11743851	CDC42SE2	0.39	1.15	1.10	0.10	3.47E-01
rs6716753	SP140	0.20	1.14	rs10185424		0.47	1.07	0.93	0.08	3.49E-01
rs7517847	IL23R	0.38	0.71	rs3766606	PARK7	0.16	0.90	1.10	0.10	3.50E-01
rs3742130	GPR18	0.21	0.91	rs9525625	LINC02341	0.49	1.08	0.91	0.10	3.52E-01
rs7657746	KIAA1109	0.24	0.92	rs2189234	TET2	0.37	1.03	1.07	0.07	3.52E-01
rs568617	FIBP	0.19	1.09	rs11230563	CD6	0.34	0.92	0.93	0.08	3.53E-01
rs3749171	GPR35	0.18	1.08	rs7608910	PUS10	0.41	1.13	1.07	0.08	3.53E-01
rs564349	ERGIC1	0.32	1.05	rs4703855		0.29	0.93	1.06	0.07	3.54E-01
rs3749171	GPR35	0.18	1.08	rs2382817	TMBIM1	0.41	1.07	1.12	0.12	3.55E-01
rs10185424		0.47	1.07	rs1260326	GCKR	0.44	1.12	0.93	0.08	3.57E-01
rs1517352	STAT4	0.39	0.92	rs10185424		0.47	1.07	0.93	0.08	3.62E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs72810983	CPEB4	0.29	0.91	rs9313808		0.16	0.86	0.90	0.11	3.63E-01
rs72810983	CPEB4	0.29	0.91	rs17622378	C5orf56	0.45	1.21	0.94	0.07	3.69E-01
rs254560	C5orf66	0.41	1.03	rs10065637	ANKRD55	0.21	0.91	0.93	0.08	3.70E-01
rs12946510		0.48	1.13	rs3091315		0.26	0.87	1.07	0.08	3.72E-01
rs2641348	ADAM30	0.10	0.89	rs6588248	IL23R	0.45	0.88	1.10	0.11	3.75E-01
rs6716753	SP140	0.20	1.14	rs10495903	THADA	0.14	1.13	1.09	0.10	3.76E-01
rs2651244		0.40	1.05	rs3766606	PARK7	0.16	0.90	0.93	0.08	3.79E-01
rs516246	FUT2	0.48	1.12	rs4802307	PPP5C	0.29	0.91	1.07	0.08	3.84E-01
rs2651244		0.40	1.05	rs6588248	IL23R	0.45	0.88	0.94	0.08	3.84E-01
rs6556412	LOC285626	0.35	1.17	rs10065637	ANKRD55	0.21	0.91	0.94	0.07	3.85E-01
rs6716753	SP140	0.20	1.14	rs2382817	TMBIM1	0.41	1.07	0.91	0.11	3.88E-01
rs4409764		0.49	1.19	rs1199103		0.21	0.90	0.91	0.11	3.89E-01
rs11741861	ZNF300	0.09	1.33	rs17622378	C5orf56	0.45	1.21	1.09	0.10	3.92E-01
rs670523	RIT1	0.34	1.08	rs4845604	RORC	0.14	0.91	0.94	0.08	3.94E-01
rs4409764		0.49	1.19	rs2227551	PLAU	0.25	0.91	0.94	0.08	3.95E-01
rs3764147	LACC1	0.25	1.15	rs941823	LINC00598	0.24	0.94	0.94	0.07	3.99E-01
rs72810983	CPEB4	0.29	0.91	rs56167332		0.36	1.19	1.06	0.07	4.00E-01
rs9525625	LINC02341	0.49	1.08	rs941823	LINC00598	0.24	0.94	0.94	0.08	4.09E-01
rs10061469		0.31	0.92	rs10065637	ANKRD55	0.21	0.91	0.94	0.07	4.10E-01
rs11741861	ZNF300	0.09	1.33	rs1363907	ERAP2	0.43	1.11	1.09	0.10	4.14E-01
rs11681525	TEX41	0.08	0.86	rs6740462		0.25	0.91	1.08	0.09	4.14E-01
rs1517352	STAT4	0.39	0.92	rs925255	FOSL2	0.43	0.90	0.94	0.07	4.14E-01
rs3091315		0.26	0.87	rs2945412	KSR1	0.39	0.89	0.94	0.08	4.15E-01
rs4409764		0.49	1.19	rs7097656	TSPAN14	0.19	0.88	1.07	0.08	4.17E-01
rs7097656	TSPAN14	0.19	0.88	rs2227551	PLAU	0.25	0.91	0.93	0.09	4.18E-01
rs13204742		0.13	1.15	rs3851228	TRAF3IP2-AS1	0.07	1.14	0.91	0.12	4.19E-01
rs3197999	MST1	0.30	1.17	rs9868809	GELSR3	0.11	1.12	0.89	0.14	4.20E-01
rs6556412	LOC285626	0.35	1.17	rs17622378	C5orf56	0.45	1.21	0.94	0.07	4.20E-01
rs11150589	ITGAL	0.47	1.02	rs26528	IL27	0.46	1.13	1.07	0.08	4.21E-01
rs254560	C5orf66	0.41	1.03	rs11743851	CDC42SE2	0.39	1.15	1.08	0.09	4.24E-01
rs568617	FIBP	0.19	1.09	rs559928		0.17	0.91	1.06	0.08	4.24E-01

SNP-1				SNP-2				Interaction	
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	P
rs254560	C5orf66	0.41	1.03	rs10051722		0.29	0.91	0.93	0.09
rs56167332		0.36	1.19	rs10065637	ANKRD55	0.21	0.91	0.95	0.07
rs6716753	SP140	0.20	1.14	rs925255	FOSL2	0.43	0.90	1.08	0.10
rs7554511	C1orf106	0.26	0.86	rs4656958		0.30	0.94	0.95	0.07
rs4409764		0.49	1.19	rs10761659		0.44	0.83	0.94	0.08
rs564349	ERGIC1	0.32	1.05	rs56167332		0.36	1.19	0.95	0.07
rs7438704	SLAIN2	0.34	0.92	rs4692386		0.39	0.94	0.95	0.07
rs11743851	CDC42SE2	0.39	1.15	rs10065637	ANKRD55	0.21	0.91	0.95	0.07
rs4656958		0.30	0.94	rs670523	RIT1	0.34	1.08	0.95	0.07
rs12199775		0.06	0.89	rs7746082		0.30	1.14	1.08	0.11
rs4409764		0.49	1.19	rs2224090		0.43	1.14	0.94	0.08
rs17391694		0.11	0.89	rs12103	INTS11	0.19	1.08	1.07	0.09
rs12994997	ATG16L1	0.44	0.80	rs1517352	STAT4	0.39	0.92	0.93	0.09
rs9313808		0.16	0.86	rs1363907	ERAP2	0.43	1.11	1.07	0.09
rs17119	LOC101928354	0.19	0.92	rs7773324		0.38	0.92	0.95	0.07
rs1250546	ZMIZ1	0.40	0.90	rs1199103		0.21	0.90	0.95	0.07
rs12994997	ATG16L1	0.44	0.80	rs6716753	SP140	0.20	1.14	0.94	0.08
rs6708413	IL18RAP	0.25	1.12	rs10865331		0.39	1.08	1.06	0.08
rs10051722		0.29	0.91	rs10065637	ANKRD55	0.21	0.91	1.08	0.11
rs6556412	LOC285626	0.35	1.17	rs11743851	CDC42SE2	0.39	1.15	0.95	0.07
rs6716753	SP140	0.20	1.14	rs1260326	GCKR	0.44	1.12	0.93	0.09
rs12718244	SPATA48	0.42	1.08	rs864745	JAZF1	0.49	1.08	0.94	0.09
rs6908425	CDKAL1	0.20	0.90	rs13204048	SLC22A23	0.37	0.93	1.07	0.09
rs7554511	C1orf106	0.26	0.86	rs10798069	PLA2G4A	0.48	0.93	0.95	0.08
rs3749171	GPR35	0.18	1.08	rs10495903	THADA	0.14	1.13	0.93	0.10
rs13277237	CCDC26	0.44	1.05	rs921720	LOC105375746	0.37	0.89	0.95	0.07
rs3742130	GPR18	0.21	0.91	rs3764147	LACC1	0.25	1.15	0.94	0.09
rs7236492	NFATC1	0.14	0.91	rs727088		0.49	1.05	1.06	0.09
rs11681525	TEX41	0.08	0.86	rs7608910	PUS10	0.41	1.13	1.11	0.15
rs56167332		0.36	1.19	rs11743851	CDC42SE2	0.39	1.15	0.95	0.07
rs3740415	MFSD13A	0.46	0.95	rs2227551	PLAU	0.25	0.91	0.95	0.07

SNP-1			SNP-2			Interaction				
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs3740415	MFSD13A	0.46	0.95	rs4409764		0.49	1.19	1.06	0.08	4.90E-01
rs7165170	CRTC3	0.18	0.93	rs17293632	SMAD3	0.25	1.14	1.05	0.07	4.92E-01
rs564349	ERGIC1	0.32	1.05	rs11741861	ZNF300	0.09	1.33	1.07	0.09	4.93E-01
rs212388	LOC105378083	0.41	1.11	rs7746082		0.30	1.14	1.05	0.07	4.93E-01
rs11010067		0.37	1.14	rs12722515	IL2RA	0.15	0.88	0.95	0.08	4.95E-01
rs6708413	IL18RAP	0.25	1.12	rs10495903	THADA	0.14	1.13	1.05	0.08	4.99E-01
rs7758080	TAB2	0.29	1.08	rs7746082		0.30	1.14	1.06	0.09	5.00E-01

Supplement 3b: G×G interactions between SNPs associated with UC on the same chromosome arm in case control design.
 ID: SNP rs number; MAF: minor allele frequency; OR: odds ratio; SE: standard error of interaction odds ratio; P: nominal p value of meta-analysis Wald test; \$: allelic disease odds ratio of minor allele. according to Liu et al. (2015); &: interaction odds ratio.

ID	SNP-1			SNP-2			Interaction			
	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs1440088	PLCL1	0.19	0.92	rs10495903	THADA	0.14	1.07	0.75	0.09	1.75E-03
rs2651244		0.39	0.94	rs6667605	LOC100996583	0.48	0.92	0.77	0.09	2.46E-03
rs3116494	CD28	0.26	1.08	rs1517352	STAT4	0.39	0.93	0.80	0.08	3.56E-03
rs6062504		0.30	0.92	rs6088765	MMP24-AS1-EDEM2	0.45	1.06	0.80	0.08	5.23E-03
rs1517352	STAT4	0.39	0.93	rs13407913	ADCY3	0.44	1.07	0.80	0.08	6.36E-03
rs2111485		0.40	1.09	rs11681525	TEX41	0.09	0.94	1.29	0.10	1.03E-02
rs6708413	IL18RAP	0.24	1.05	rs925255	FOSL2	0.45	0.96	0.82	0.08	1.06E-02
rs6074022		0.26	1.05	rs4243971		0.44	0.96	0.81	0.09	1.33E-02
rs9313808		0.16	0.88	rs6556412	LOC285626	0.34	1.10	1.25	0.09	1.61E-02
rs6740462		0.25	0.94	rs925255	FOSL2	0.45	0.96	1.20	0.08	1.94E-02
rs3740415	MFSD13A	0.46	0.94	rs12778642		0.44	1.06	0.82	0.08	2.19E-02
rs4656958		0.30	0.92	rs4845604	RORC	0.14	0.85	0.83	0.08	2.46E-02
rs17229285	LOC105373831	0.49	1.10	rs10865331		0.38	0.98	0.82	0.09	2.71E-02
rs1405108		0.34	1.09	rs6708413	IL18RAP	0.24	1.05	0.85	0.08	3.97E-02
rs11681525	TEX41	0.09	0.94	rs10865331		0.38	0.98	1.23	0.10	4.10E-02
rs3806308	RNF186	0.35	0.84	rs3766606	PARK7	0.16	0.87	1.18	0.08	4.13E-02
rs3116494	CD28	0.26	1.08	rs1440088	PLCL1	0.19	0.92	1.17	0.08	4.30E-02
rs483905	MAML2	0.30	1.09	rs2155219		0.49	1.13	1.19	0.08	4.38E-02
rs4976646	RGS14	0.35	1.08	rs9313808		0.16	0.88	0.85	0.08	4.48E-02
rs38911		0.46	0.95	rs134313	EPHB4	0.31	1.06	0.86	0.08	4.49E-02
rs3740415	MFSD13A	0.46	0.94	rs10995235	ZNF365	0.18	1.12	0.84	0.08	4.50E-02
rs6426833		0.43	0.79	rs3766606	PARK7	0.16	0.87	1.22	0.10	4.70E-02
rs4812833		0.47	0.90	rs4243971		0.44	0.96	1.18	0.09	4.76E-02
rs17229285	LOC105373831	0.49	1.10	rs13407913	ADCY3	0.44	1.07	0.84	0.09	4.79E-02
rs2189234	TET2	0.38	1.08	rs3774937	NFKB1	0.34	1.10	1.16	0.07	4.87E-02
rs7517847	IL23R	0.41	0.86	rs12568930		0.16	0.88	1.22	0.10	5.04E-02
rs6724516		0.26	0.88	rs6740462		0.25	0.94	1.15	0.07	5.19E-02

SNP-1				SNP-2				Interaction			
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P	
rs3116494	CD28	0.26	1.08	rs2111485		0.40	1.09	0.86	0.08	5.21E-02	
rs1505992		0.32	0.95	rs395157	OSMR	0.50	1.09	1.17	0.08	5.38E-02	
rs3116494	CD28	0.26	1.08	rs1260326	GCKR	0.42	1.04	1.16	0.08	5.42E-02	
rs3749171	GPR35	0.19	1.15	rs3116494	CD28	0.26	1.08	0.83	0.09	5.51E-02	
rs4812833		0.47	0.90	rs6087990		0.41	1.06	0.85	0.08	5.74E-02	
rs2836878		0.25	0.80	rs2823286		0.28	0.91	1.15	0.07	5.75E-02	
rs11583043	DPH5	0.28	1.08	rs10799838		0.24	1.14	1.18	0.09	5.82E-02	
rs6856616		0.07	1.10	rs2073505	HGFAC	0.08	1.09	0.77	0.14	6.40E-02	
rs13300218	NOTCH1	0.10	0.87	rs11554257		0.14	1.13	1.21	0.11	6.72E-02	
rs11583043	DPH5	0.28	1.08	rs2651244		0.39	0.94	1.14	0.07	7.31E-02	
rs2111485		0.40	1.09	rs13407913	ADCY3	0.44	1.07	1.16	0.08	7.46E-02	
rs6708413	IL18RAP	0.24	1.05	rs10495903	THADA	0.14	1.07	1.15	0.08	9.92E-02	
rs1505992		0.32	0.95	rs2930047	DAP	0.38	1.07	0.83	0.11	1.01E-01	
rs3766606	PARK7	0.16	0.87	rs12103	INTS11	0.19	1.10	1.18	0.10	1.01E-01	
rs6667605	LOC100996583	0.48	0.92	rs12103	INTS11	0.19	1.10	0.87	0.09	1.02E-01	
rs12718244	SPATA48	0.42	1.07	rs1077773	KCCAT333	0.47	0.93	0.87	0.08	1.05E-01	
rs395157	OSMR	0.50	1.09	rs11739663		0.22	0.93	0.83	0.11	1.05E-01	
rs12942547	STAT3	0.41	0.92	rs12946510		0.48	1.14	0.87	0.08	1.07E-01	
rs9847710	SFMBT1	0.42	1.07	rs113010081	LOC105377068	0.11	1.14	1.17	0.10	1.08E-01	
rs13300218	NOTCH1	0.10	0.87	rs4246905	TNFSF15	0.28	0.89	1.21	0.12	1.08E-01	
rs4976646	RGS14	0.35	1.08	rs17622378	C5orf56	0.43	1.09	1.21	0.12	1.09E-01	
rs6740462		0.25	0.94	rs1260326	GCKR	0.42	1.04	0.88	0.08	1.11E-01	
rs2155219		0.49	1.13	rs11230563	CD6	0.34	0.93	0.88	0.08	1.11E-01	
rs259964	ZNF831	0.46	1.06	rs6088765	MMP24-AS1-EDEM2	0.45	1.06	0.87	0.09	1.12E-01	
rs1405108		0.34	1.09	rs925255	FOSL2	0.45	0.96	1.15	0.09	1.14E-01	
rs17771967		0.43	1.07	rs17694108		0.29	1.10	1.13	0.08	1.16E-01	
rs653178	ATXN2	0.50	1.05	rs11168249	HDAC7	0.47	1.06	1.15	0.09	1.18E-01	
rs1077773	KCCAT333	0.47	0.93	rs1182188	GNA12	0.29	0.90	1.13	0.08	1.18E-01	
rs4409764		0.50	1.17	rs224090		0.42	1.06	0.87	0.09	1.20E-01	
rs6908425	CDKAL1	0.21	0.93	rs17119	LOC101928354	0.19	0.93	0.88	0.08	1.20E-01	
rs6679677	PHTF1	0.10	1.06	rs3766606	PARK7	0.16	0.87	1.19	0.11	1.20E-01	

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs7517847	IL23R	0.41	0.86	rs10799838		0.24	1.14	1.12	0.08	1.21E-01
rs6740462		0.25	0.94	rs13407913	ADCY3	0.44	1.07	1.13	0.08	1.28E-01
rs10799838		0.24	1.14	rs3766606	PARK7	0.16	0.87	0.88	0.08	1.28E-01
rs2382817	TMBIM1	0.41	1.07	rs7608910	PUS10	0.41	1.14	0.88	0.08	1.30E-01
rs3806308	RNF186	0.35	0.84	rs10799838		0.24	1.14	0.87	0.09	1.31E-01
rs17780256	SLC39A11	0.18	0.89	rs12946510		0.48	1.14	1.15	0.09	1.35E-01
rs6679677	PHTF1	0.10	1.06	rs11583043	DPH5	0.28	1.08	0.87	0.09	1.35E-01
rs7517847	IL23R	0.41	0.86	rs6426833		0.43	0.79	1.13	0.08	1.37E-01
rs35320439	GAL3ST2	0.32	1.01	rs6740462		0.25	0.94	0.90	0.07	1.37E-01
rs7554511	C1orf106	0.26	0.85	rs4656958		0.30	0.92	0.90	0.07	1.40E-01
rs17229285	LOC105373831	0.49	1.10	rs1517352	STAT4	0.39	0.93	0.88	0.08	1.44E-01
rs1456896		0.31	0.95	rs1077773	KCCAT333	0.47	0.93	0.89	0.08	1.46E-01
rs12627970		0.21	1.11	rs2256609	UBE2L3	0.20	1.06	1.12	0.08	1.47E-01
rs564349	ERGIC1	0.32	1.06	rs56167332		0.35	1.15	0.90	0.07	1.48E-01
rs6708413	IL18RAP	0.24	1.05	rs13407913	ADCY3	0.44	1.07	0.87	0.09	1.49E-01
rs12718244	SPATA48	0.42	1.07	rs4722672		0.19	1.08	0.89	0.08	1.49E-01
rs6724516		0.26	0.88	rs1260326	GCKR	0.42	1.04	1.12	0.08	1.49E-01
rs17780256	SLC39A11	0.18	0.89	rs2945412	KSR1	0.41	1.00	0.87	0.10	1.53E-01
rs7608910	PUS10	0.41	1.14	rs13407913	ADCY3	0.44	1.07	0.88	0.09	1.54E-01
rs6740462		0.25	0.94	rs7608910	PUS10	0.41	1.14	1.14	0.10	1.60E-01
rs7657746	KIAA1109	0.24	0.91	rs2189234	TET2	0.38	1.08	1.11	0.08	1.62E-01
rs6724516		0.26	0.88	rs1517352	STAT4	0.39	0.93	0.89	0.08	1.62E-01
rs2111485		0.40	1.09	rs7608910	PUS10	0.41	1.14	1.12	0.08	1.65E-01
rs26528	IL27	0.46	1.06	rs7404095	PRKCB	0.42	0.93	0.89	0.08	1.68E-01
rs3116494	CD28	0.26	1.08	rs11681525	TEX41	0.09	0.94	0.87	0.10	1.69E-01
rs9847710	SFMBT1	0.42	1.07	rs9868809	CELSR3	0.11	1.16	1.14	0.10	1.69E-01
rs1440088	PLCL1	0.19	0.92	rs925255	FOSL2	0.45	0.96	0.89	0.08	1.69E-01
rs3853824	C17orf67	0.35	0.95	rs12946510		0.48	1.14	1.12	0.08	1.73E-01
rs6708413	IL18RAP	0.24	1.05	rs10865331		0.38	0.98	1.11	0.07	1.74E-01
rs3749171	GPR35	0.19	1.15	rs925255	FOSL2	0.45	0.96	1.12	0.08	1.74E-01
rs3749171	GPR35	0.19	1.15	rs10185424		0.47	1.10	1.12	0.08	1.76E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs564349	ERGIC1	0.32	1.06	rs254560	C5orf66	0.41	1.08	1.15	0.10	1.76E-01
rs2382817	TMBIM1	0.41	1.07	rs1440088	PLCL1	0.19	0.92	1.13	0.09	1.76E-01
rs17736589	CYTH1	0.22	1.09	rs12942547	STAT3	0.41	0.92	0.90	0.08	1.77E-01
rs6920220		0.22	1.16	rs7746082		0.29	1.08	1.11	0.07	1.78E-01
rs12199775		0.07	0.92	rs7746082		0.29	1.08	1.16	0.11	1.78E-01
rs2651244		0.39	0.94	rs12568930		0.16	0.88	1.14	0.10	1.80E-01
rs4409764		0.50	1.17	rs12778642		0.44	1.06	0.89	0.09	1.80E-01
rs224090		0.42	1.06	rs10761659		0.45	0.89	1.19	0.13	1.82E-01
rs11741861	ZNF300	0.08	1.13	rs254560	C5orf66	0.41	1.08	0.81	0.16	1.83E-01
rs4664304	PLA2R1	0.45	1.06	rs6740462		0.25	0.94	1.12	0.09	1.85E-01
rs561722	NXPE2	0.32	0.88	rs11229555		0.24	0.92	1.10	0.07	1.87E-01
rs4664304	PLA2R1	0.45	1.06	rs13407913	ADCY3	0.44	1.07	1.17	0.12	1.88E-01
rs12568930		0.16	0.88	rs6426833		0.43	0.79	0.90	0.08	1.93E-01
rs10781499	CARD9	0.44	1.14	rs4743820	LINC00484	0.29	0.92	0.90	0.08	1.94E-01
rs7097656	TSPAN14	0.20	0.94	rs10995235	ZNF365	0.18	1.12	0.90	0.08	1.95E-01
rs6062504		0.30	0.92	rs913678		0.32	0.93	0.88	0.10	1.95E-01
rs7404095	PRKCB	0.42	0.93	rs11641184	LITAF	0.49	1.08	0.89	0.09	1.97E-01
rs9847710	SFMBT1	0.42	1.07	rs3197999	MST1	0.30	1.19	0.91	0.08	1.98E-01
rs6863411	NDFIP1	0.37	0.94	rs254560	C5orf66	0.41	1.08	1.18	0.13	1.98E-01
rs7608910	PUS10	0.41	1.14	rs1260326	GCKR	0.42	1.04	1.14	0.10	1.98E-01
rs35320439	GAL3ST2	0.32	1.01	rs17229285	LOC105373831	0.49	1.10	0.89	0.09	2.04E-01
rs12946510		0.48	1.14	rs3091315		0.27	0.94	1.11	0.08	2.04E-01
rs17229285	LOC105373831	0.49	1.10	rs6740462		0.25	0.94	0.90	0.08	2.05E-01
rs9847710	SFMBT1	0.42	1.07	rs3172494	IP6K2	0.12	1.11	0.89	0.09	2.10E-01
rs3742130	GPR18	0.22	0.93	rs941823	LINC00598	0.24	0.90	1.10	0.07	2.13E-01
rs1405108		0.34	1.09	rs13407913	ADCY3	0.44	1.07	1.10	0.08	2.14E-01
rs1505992		0.32	0.95	rs11739663		0.22	0.93	0.89	0.09	2.17E-01
rs1505992		0.32	0.95	rs1842076		0.28	0.95	1.11	0.08	2.17E-01
rs1405108		0.34	1.09	rs1517352	STAT4	0.39	0.93	0.91	0.08	2.18E-01
rs3749171	GPR35	0.19	1.15	rs6740462		0.25	0.94	0.91	0.08	2.22E-01
rs6426833		0.43	0.79	rs121103	INTS11	0.19	1.10	1.10	0.08	2.24E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs4380874		0.42	1.14	rs297145	KPNA7	0.26	1.07	1.11	0.09	2.25E-01
rs1517352	STAT4	0.39	0.93	rs10865331		0.38	0.98	1.10	0.08	2.26E-01
rs35320439	GAL3ST2	0.32	1.01	rs10865331		0.38	0.98	1.09	0.07	2.27E-01
rs4246905	TNFSF15	0.28	0.89	rs4743820	LINC00484	0.29	0.92	1.11	0.09	2.27E-01
rs2413583		0.16	0.87	rs2256609	UBE2L3	0.20	1.06	0.91	0.08	2.29E-01
rs3749171	GPR35	0.19	1.15	rs11681525	TEX41	0.09	0.94	1.13	0.10	2.35E-01
rs11583043	DPH5	0.28	1.08	rs12568930		0.16	0.88	0.89	0.10	2.37E-01
rs9358372	CDKAL1	0.37	1.04	rs17119	LOC101928354	0.19	0.93	1.10	0.08	2.37E-01
rs10761659		0.45	0.89	rs10995235	ZNF365	0.18	1.12	0.91	0.08	2.39E-01
rs1819333		0.46	0.96	rs7746082		0.29	1.08	0.88	0.11	2.43E-01
rs941823	LINC00598	0.24	0.90	rs17085007		0.19	1.14	1.09	0.08	2.44E-01
rs10185424		0.47	1.10	rs1260326	GCKR	0.42	1.04	1.10	0.09	2.45E-01
rs17229285	LOC105373831	0.49	1.10	rs925255	FOSL2	0.45	0.96	0.90	0.09	2.46E-01
rs1440088	PLCL1	0.19	0.92	rs11681525	TEX41	0.09	0.94	0.89	0.10	2.46E-01
rs17229285	LOC105373831	0.49	1.10	rs7608910	PUS10	0.41	1.14	0.91	0.09	2.47E-01
rs7805114		0.42	0.89	rs314313	EPHB4	0.31	1.06	0.92	0.08	2.53E-01
rs564349	ERGIC1	0.32	1.06	rs9313808		0.16	0.88	0.91	0.08	2.57E-01
rs11554257		0.14	1.13	rs4246905	TNFSF15	0.28	0.89	0.91	0.09	2.58E-01
rs17736589	CYTH1	0.22	1.09	rs3091315		0.27	0.94	1.13	0.11	2.59E-01
rs1405108		0.34	1.09	rs1440088	PLCL1	0.19	0.92	1.09	0.08	2.65E-01
rs3024505		0.18	1.25	rs4845604	RORC	0.14	0.85	0.90	0.09	2.66E-01
rs12718244	SPATA48	0.42	1.07	rs1182188	GNA12	0.29	0.90	1.09	0.08	2.66E-01
rs17780256	SLC39A11	0.18	0.89	rs3091315		0.27	0.94	1.09	0.08	2.66E-01
rs4976646	RGS14	0.35	1.08	rs11741861	ZNF300	0.08	1.13	0.90	0.10	2.68E-01
rs6679677	PHTF1	0.10	1.06	rs6426833		0.43	0.79	1.12	0.10	2.70E-01
rs11681525	TEX41	0.09	0.94	rs6708413	IL18RAP	0.24	1.05	1.11	0.10	2.72E-01
rs3749171	GPR35	0.19	1.15	rs6708413	IL18RAP	0.24	1.05	1.09	0.08	2.72E-01
rs2382817	TMBIM1	0.41	1.07	rs3116494	CD28	0.26	1.08	0.92	0.08	2.76E-01
rs113010081	LOC105377068	0.11	1.14	rs4256159	LOC105376976	0.15	1.06	0.89	0.11	2.77E-01
rs3749171	GPR35	0.19	1.15	rs2111485		0.40	1.09	0.92	0.08	2.78E-01
rs1842076		0.28	0.95	rs2930047	DAP	0.38	1.07	0.91	0.09	2.78E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs12199775		0.07	0.92	rs6920220		0.22	1.16	0.86	0.14	2.80E-01
rs6724516		0.26	0.88	rs925255	FOSL2	0.45	0.96	1.09	0.08	2.80E-01
rs1819333		0.46	0.96	rs3851228	TRAF3IP2-AS1	0.07	1.20	1.13	0.12	2.81E-01
rs564349	ERGIC1	0.32	1.06	rs11741861	ZNF300	0.08	1.13	1.15	0.13	2.81E-01
rs17229285	LOC105373831	0.49	1.10	rs4664304	PLA2R1	0.45	1.06	1.10	0.09	2.86E-01
rs10865331		0.38	0.98	rs1260326	GCKR	0.42	1.04	1.12	0.10	2.90E-01
rs6556412	LOC285626	0.34	1.10	rs6863411	NDFIP1	0.37	0.94	0.93	0.07	2.96E-01
rs3116494	CD28	0.26	1.08	rs925255	FOSL2	0.45	0.96	1.08	0.08	2.96E-01
rs395157	OSMR	0.50	1.09	rs2930047	DAP	0.38	1.07	0.91	0.09	2.97E-01
rs3742130	GPR18	0.22	0.93	rs17085007		0.19	1.14	1.09	0.08	2.97E-01
rs6426833		0.43	0.79	rs10799838		0.24	1.14	1.12	0.11	2.98E-01
rs3749171	GPR35	0.19	1.15	rs2382817	TMBIM1	0.41	1.07	1.09	0.08	2.99E-01
rs2651244		0.39	0.94	rs3806308	RNF186	0.35	0.84	0.93	0.07	3.03E-01
rs6062504		0.30	0.92	rs259964	ZNF831	0.46	1.06	1.09	0.08	3.06E-01
rs4664304	PLA2R1	0.45	1.06	rs11681525	TEX41	0.09	0.94	0.90	0.11	3.06E-01
rs6724516		0.26	0.88	rs10185424		0.47	1.10	0.92	0.08	3.14E-01
rs1819333		0.46	0.96	rs1847472	BACH2	0.34	0.95	1.08	0.08	3.15E-01
rs1250546	ZMIZ1	0.41	0.96	rs10761659		0.45	0.89	0.92	0.08	3.22E-01
rs3116494	CD28	0.26	1.08	rs10495903	THADA	0.14	1.07	1.09	0.09	3.26E-01
rs913678		0.32	0.93	rs6074022		0.26	1.05	0.93	0.07	3.26E-01
rs11150589	ITGAL	0.48	1.08	rs26528	IL27	0.46	1.06	0.92	0.09	3.27E-01
rs2651244		0.39	0.94	rs7517847	IL23R	0.41	0.86	1.08	0.08	3.27E-01
rs3749171	GPR35	0.19	1.15	rs17229285	LOC105373831	0.49	1.10	1.09	0.09	3.30E-01
rs9313808		0.16	0.88	rs4703855		0.29	0.94	1.08	0.08	3.30E-01
rs1250546	ZMIZ1	0.41	0.96	rs224090		0.42	1.06	0.93	0.08	3.31E-01
rs4664304	PLA2R1	0.45	1.06	rs10495903	THADA	0.14	1.07	0.91	0.10	3.32E-01
rs3116494	CD28	0.26	1.08	rs13407913	ADCY3	0.44	1.07	0.90	0.11	3.33E-01
rs10185424		0.47	1.10	rs925255	FOSL2	0.45	0.96	0.92	0.09	3.34E-01
rs11741861	ZNF300	0.08	1.13	rs4703855		0.29	0.94	1.14	0.13	3.36E-01
rs2651244		0.39	0.94	rs6426833		0.43	0.79	1.08	0.08	3.37E-01
rs11741861	ZNF300	0.08	1.13	rs6863411	NDFIP1	0.37	0.94	0.91	0.10	3.39E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs6074022		0.26	1.05	rs6088765	MIMP24-AS1-EDEM2	0.45	1.06	0.91	0.10	3.40E-01
rs4664304	PLA2R1	0.45	1.06	rs10865331		0.38	0.98	1.08	0.08	3.40E-01
rs1405108		0.34	1.09	rs10185424		0.47	1.10	0.93	0.08	3.41E-01
rs3774937	NFKB1	0.34	1.10	rs2457996		0.11	0.90	1.09	0.09	3.41E-01
rs26528	IL27	0.46	1.06	rs11641184	LITAF	0.49	1.08	1.09	0.09	3.48E-01
rs10799838		0.24	1.14	rs12103	INTS11	0.19	1.10	1.08	0.08	3.51E-01
rs11150589	ITGAL	0.48	1.08	rs7404095	PRKCB	0.42	0.93	1.09	0.10	3.51E-01
rs11583043	DPH5	0.28	1.08	rs6426833		0.43	0.79	0.93	0.08	3.53E-01
rs10781499	CARD9	0.44	1.14	rs11554257		0.14	1.13	1.10	0.10	3.54E-01
rs4664304	PLA2R1	0.45	1.06	rs925255	FOSL2	0.45	0.96	1.08	0.09	3.54E-01
rs6724516		0.26	0.88	rs1405108		0.34	1.09	0.93	0.07	3.56E-01
rs7282490		0.40	1.11	rs2823286		0.28	0.91	0.91	0.10	3.57E-01
rs6074022		0.26	1.05	rs4812833		0.47	0.90	0.93	0.08	3.57E-01
rs1405108		0.34	1.09	rs4664304	PLA2R1	0.45	1.06	0.92	0.09	3.63E-01
rs3853824	C17orf67	0.35	0.95	rs2945412	KSR1	0.41	1.00	1.09	0.09	3.64E-01
rs11583043	DPH5	0.28	1.08	rs3766606	PARK7	0.16	0.87	0.93	0.08	3.66E-01
rs2382817	TMBIM1	0.41	1.07	rs17229285	LOC105373831	0.49	1.10	1.08	0.09	3.66E-01
rs1517352	STAT4	0.39	0.93	rs925255	FOSL2	0.45	0.96	0.93	0.08	3.67E-01
rs11681525	TEX41	0.09	0.94	rs7608910	PUS10	0.41	1.14	1.10	0.10	3.69E-01
rs7608910	PUS10	0.41	1.14	rs10495903	THADA	0.14	1.07	0.92	0.09	3.69E-01
rs35320439	GAL3ST2	0.32	1.01	rs6708413	IL18RAP	0.24	1.05	1.07	0.07	3.70E-01
rs1819333		0.46	0.96	rs6920220		0.22	1.16	0.91	0.10	3.71E-01
rs6724516		0.26	0.88	rs3749171	GPR35	0.19	1.15	0.92	0.09	3.73E-01
rs2816958	NR5A2	0.11	0.83	rs4656958		0.30	0.92	0.90	0.12	3.77E-01
rs2382817	TMBIM1	0.41	1.07	rs4664304	PLA2R1	0.45	1.06	1.08	0.08	3.80E-01
rs6724516		0.26	0.88	rs10495903	THADA	0.14	1.07	1.10	0.11	3.80E-01
rs6062504		0.30	0.92	rs6087990		0.41	1.06	1.07	0.08	3.81E-01
rs6679677	PHTF1	0.10	1.06	rs2651244		0.39	0.94	1.09	0.10	3.82E-01
rs11681525	TEX41	0.09	0.94	rs925255	FOSL2	0.45	0.96	0.91	0.10	3.84E-01
rs17780256	SLC39A11	0.18	0.89	rs3853824	C17orf67	0.35	0.95	0.93	0.08	3.84E-01
rs7282490		0.40	1.11	rs2836878		0.25	0.80	1.07	0.08	3.85E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs35320439	GAL3ST2	0.32	1.01	rs1405108		0.34	1.09	1.07	0.08	3.85E-01
rs4976646	RGS14	0.35	1.08	rs6863411	NDFIP1	0.37	0.94	1.09	0.10	3.86E-01
rs6863411	NDFIP1	0.37	0.94	rs17622378	C5orf56	0.43	1.09	1.07	0.08	3.86E-01
rs9868809	CELSR3	0.11	1.16	rs113010081	LOC105377068	0.11	1.14	0.90	0.12	3.88E-01
rs12627970		0.21	1.11	rs2413583		0.16	0.87	0.92	0.10	3.90E-01
rs6679677	PHTF1	0.10	1.06	rs12568930		0.16	0.88	0.92	0.10	3.90E-01
rs3749171	GPR35	0.19	1.15	rs1440088	PLCL1	0.19	0.92	0.93	0.08	3.93E-01
rs6679677	PHTF1	0.10	1.06	rs3806308	RNF186	0.35	0.84	1.08	0.09	3.96E-01
rs3740415	MFSD13A	0.46	0.94	rs10761659		0.45	0.89	1.08	0.09	3.96E-01
rs6724516		0.26	0.88	rs10865331		0.38	0.98	0.94	0.07	3.97E-01
rs17229285	LOC105373831	0.49	1.10	rs11681525	TEX41	0.09	0.94	0.89	0.14	3.97E-01
rs11742570		0.38	0.92	rs395157	OSMR	0.50	1.09	1.07	0.08	4.05E-01
rs17229285	LOC105373831	0.49	1.10	rs2111485		0.40	1.09	0.92	0.10	4.05E-01
rs11681525	TEX41	0.09	0.94	rs10185424		0.47	1.10	0.92	0.11	4.05E-01
rs7236492	NFATC1	0.15	0.95	rs7240004		0.37	0.92	0.93	0.09	4.09E-01
rs1456896		0.31	0.95	rs4722672		0.19	1.08	1.07	0.08	4.12E-01
rs6062504		0.30	0.92	rs4812833		0.47	0.90	0.93	0.09	4.14E-01
rs11083840	PTGIR	0.41	1.07	rs17694108		0.29	1.10	0.94	0.08	4.14E-01
rs1517352	STAT4	0.39	0.93	rs6708413	IL18RAP	0.24	1.05	1.08	0.10	4.16E-01
rs6556412	LOC285626	0.34	1.10	rs254560	C5orf66	0.41	1.08	0.94	0.08	4.16E-01
rs925255	FOSL2	0.45	0.96	rs13407913	ADCY3	0.44	1.07	1.07	0.08	4.19E-01
rs11741861	ZNF300	0.08	1.13	rs17622378	C5orf56	0.43	1.09	0.92	0.10	4.23E-01
rs11583043	DPH5	0.28	1.08	rs6667605	LOC100996583	0.48	0.92	1.07	0.08	4.27E-01
rs2382817	TMBIM1	0.41	1.07	rs13407913	ADCY3	0.44	1.07	0.94	0.08	4.29E-01
rs7746082		0.29	1.08	rs1847472	BACH2	0.34	0.95	1.06	0.07	4.31E-01
rs3197999	MST1	0.30	1.19	rs9868809	CELSR3	0.11	1.16	0.88	0.16	4.34E-01
rs6724516		0.26	0.88	rs2111485		0.40	1.09	1.06	0.08	4.37E-01
rs913678		0.32	0.93	rs6088765	MMP24-AS1-EDEM2	0.45	1.06	1.06	0.08	4.43E-01
rs6088765	MMP24-AS1-EDEM2	0.45	1.06	rs6087990		0.41	1.06	1.12	0.15	4.45E-01
rs3024505		0.18	1.25	rs4656958		0.30	0.92	1.07	0.09	4.46E-01
rs1819333		0.46	0.96	rs12199775		0.07	0.92	0.90	0.14	4.48E-01

SNP-1				SNP-2				Interaction		
ID	Gene	MAF	OR\$	ID	Gene	MAF	OR\$	OR&	SE	P
rs254560	C5orf66	0.41	1.08	rs17622378	C5orf56	0.43	1.09	1.06	0.08	4.50E-01
rs1505992		0.32	0.95	rs11742570		0.38	0.92	0.95	0.07	4.53E-01
rs11554257		0.14	1.13	rs4743820	LINC00484	0.29	0.92	0.93	0.10	4.54E-01
rs11010067		0.36	1.08	rs12722515	IL2RA	0.15	0.93	0.94	0.08	4.55E-01
rs6863411	NDFIP1	0.37	0.94	rs4703855		0.29	0.94	1.09	0.11	4.55E-01
rs17780256	SLC39A11	0.18	0.89	rs12942547	STAT3	0.41	0.92	0.94	0.08	4.56E-01
rs35320439	GAL3ST2	0.32	1.01	rs1260326	GCKR	0.42	1.04	1.07	0.09	4.59E-01
rs56167332		0.35	1.15	rs6863411	NDFIP1	0.37	0.94	0.95	0.07	4.60E-01
rs3116494	CD28	0.26	1.08	rs10865331		0.38	0.98	0.94	0.08	4.63E-01
rs10495903	THADA	0.14	1.07	rs925255	FOSL2	0.45	0.96	0.94	0.09	4.63E-01
rs1729285	LOC105373831	0.49	1.10	rs1260326	GCKR	0.42	1.04	0.94	0.09	4.64E-01
rs9313808		0.16	0.88	rs17622378	C5orf56	0.43	1.09	0.94	0.08	4.67E-01
rs1405108		0.34	1.09	rs2111485		0.40	1.09	0.93	0.10	4.68E-01
rs1517352	STAT4	0.39	0.93	rs1260326	GCKR	0.42	1.04	0.94	0.08	4.69E-01
rs2111485		0.40	1.09	rs4664304	PLA2R1	0.45	1.06	0.94	0.08	4.72E-01
rs4409764		0.50	1.17	rs10995235	ZNF365	0.18	1.12	1.07	0.10	4.73E-01
rs6556412	LOC285626	0.34	1.10	rs11741861	ZNF300	0.08	1.13	0.93	0.10	4.79E-01
rs6708413	IL18RAP	0.24	1.05	rs1260326	GCKR	0.42	1.04	1.06	0.08	4.80E-01
rs11583043	DPH5	0.28	1.08	rs3806308	RNF186	0.35	0.84	0.95	0.07	4.81E-01
rs925255	FOSL2	0.45	0.96	rs1260326	GCKR	0.42	1.04	0.94	0.08	4.84E-01
rs17293632	SMAD3	0.24	1.08	rs28374715		0.25	0.92	0.95	0.07	4.84E-01
rs1405108		0.34	1.09	rs10865331		0.38	0.98	1.05	0.08	4.85E-01
rs4976646	RGS14	0.35	1.08	rs254560	C5orf66	0.41	1.08	0.95	0.08	4.87E-01
rs38911		0.46	0.95	rs7805114		0.42	0.89	1.06	0.08	4.88E-01
rs1729285	LOC105373831	0.49	1.10	rs10495903	THADA	0.14	1.07	0.94	0.09	4.88E-01
rs10865331		0.38	0.98	rs13407913	ADCY3	0.44	1.07	0.93	0.11	4.89E-01
rs38911		0.46	0.95	rs4380874		0.42	1.14	1.06	0.08	4.95E-01
rs4409764		0.50	1.17	rs7097656	TSPAN14	0.20	0.94	1.06	0.09	4.98E-01
rs10865331		0.38	0.98	rs10495903	THADA	0.14	1.07	0.93	0.11	5.00E-01

2.3 Case-Only Analysis of Gene-Gene Interactions in Parkinson Disease

2.3.1 Summary

The extent to which SNP associations can explain the heritability of the complex human diseases is limited (Maher 2008) and the neurodegenerative PD is no exception. While the exact cause of PD is still unknown, there are two forms – monogenic and sporadic and there is evidence that genetic variability plays a role in both. Moreover, Blauwendraat et al. have found two genome-wide significant associations between PD and age at onset (Blauwendraat et al. 2019). Complex diseases are believed to result from G×G and G×E interactions, genetic heterogeneity, very rare variants and potentially even more reasons, which are yet unknown (Manolio et al. 2009). In this paper I focus on G×G interactions between SNPs in PD using the CO study design.

The data used in this paper originated from the IPDGC and consisted of 36 362 cases from 16 different centers across Europe and North America and 7.8 million SNPs were available after imputation. The analysis was carried out on all cases and two independent subsets were considered: patients with early onset PD, where the age of the patient at diagnosis was earlier than 50 years of age ($n = 6\,962$), and not early onset of patients of over 50 years of age at diagnosis of PD ($n = 29\,400$). The analysis method was very similar as described in Chapter 2.2: SNPs of interest were narrowed down to 90 SNPs with known ME (Nalls et al. 2019), only SNP pairs on different chromosome arms were considered, a center-wise logistic regression with PCs was carried out and the results were meta analysed using random effects and a Wald test. I expanded the method in this paper to incorporate a genome-wide G×G interaction search by making sure that one SNP in the interaction pair had a known ME in order to reduce the computational burden. The significance threshold was thus set at the genome-wide significance of 5×10^{-8} divided by the number of genome-wide searches conducted (90), equalling to 5.56×10^{-10} .

Even though the genome-wide G×G interaction analysis was confined to one of the SNPs in the interaction pair having a ME, the analysis was still computationally demanding. Output files consisting of over 11 billion logistic regression results were generated due to the fact that the logistic regression was applied for each of the 16 centers separately. Therefore, in order to manage the output data, a two-step screening process was applied. First, a meta-analysis was conducted on only those centers, where the specific SNP pair center-level p-values were less than 0.05. All SNP pairs that had a combined p-value from the meta-analysis of less than 5×10^{-5} were selected for further analysis. Subsequently, a meta-analysis was done for all selected SNPs including all centers. While some statistical power was lost during this process, it made it possible to handle the data and by no means could produce false positive results.

The genome-wide search yielded 337 significant G×G interactions, consisting of ten unique SNP pair combinations in which the two SNPs were located in unique genes or gene regions. Out of these significant interactions, 136 were found when considering all available cases and 201 in the subset of cases with not early onset PD. No statistically significant interactions,

given the significance level of 5.56×10^{-10} were found when considering the early onset PD data subset separately.

The most interesting region found included rs76904798 in the *AC079630.4* antisense gene with a known main effect, which overlaps with the *LRRK2* gene and the area in close proximity to the promoter region of *SYT10*. Many, 227 to be exact, of the significant G×G interactions were found in this area ranging from 33 Mb to 33.75 Mb on chromosome 12. The SNP×SNP combination pair with the lowest p-value equal to 2.67×10^{-43} in this region was with rs1007709 and the OR was 1.80 (CI = [1.65, 1.95]). The molecular function of *LRRK2* is directly associated with neural plasticity and the gene is strongly associated with PD. *SYT10*, on the other hand, contributes to the exocytosis of secretory vesicles in neurons, therefore a biological interaction may be plausible between these two genes.

In conclusion, using one of the largest available datasets of PD cases from the IPDGC, the investigation of G×G interactions using the statistically powerful CO study design found ten unique and statistically significant ($p < 5.56 \times 10^{-10}$) G×G interaction pairs. While these findings give ground for further research on the underlying gene combinations, it must also be taken into account, that statistical interaction does not necessarily imply biological interaction. Moreover, there may even be more potential G×G interaction pairs that could not be discovered in this large dataset due to rare SNPs. Some SNPs relevant for PD, for example in the *LRRK2* gene region, are very rare and have a MAF of less than 0.02. Thus, especially in combination with other SNPs with low MAF, there may not be enough information on rare allele carriers of both SNPs, resulting in not converging logistic regressions or meta-analysis, very wide confidence intervals and unrealistically high ORs. In conclusion, this study used a computationally efficient method of focusing on SNPs with ME and using a screening process and showed that there are statistically significant interactions when examining PD cases.

2.3.2 Publication

Case-Only Analysis of Gene-Gene Interactions in Parkinson Disease

Milda Aleknyonytė-Resch¹, Hampton Leonard, Cornelis Blauwendraat, Arunabh Sharma¹, Pegah Rahmati¹, Sandra Freitag-Wolf¹, Andrew B. Singleton for the International Parkinson Disease Genomics Consortium, Christine Klein², Michael Krawczak¹, Astrid Dempfle^{1*}

1. Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany
2. Institute of Neurogenetics, Lübeck University, Germany

*corresponding author, contact details:

Prof. Dr. Astrid Dempfle, dempfle@medinfo.uni-kiel.de

Introduction

Parkinson disease (PD) is a complex disease; the most common movement disorder and second most common neurodegenerative disease after Alzheimer's. Genome-wide association studies (GWAS) have been used extensively in the past to unravel the 'genetic architecture' of complex human diseases, i.e. to assess the population-wide level of statistical association between a disease of interest and the genotypes of single nucleotide polymorphisms (SNPs). While GWAS have successfully identified potentially causal links between genes (or gene regions) and numerous diseases, including PD (Buniello et al. 2019), the extent to which SNP associations alone can explain the heritability of the complex human diseases is limited (Maher 2008) and PD is no exception. There are known risk genes such as *LRRK2* and *VPS35* (Trinh et al. 2018) that are low penetrance and small effect sizes that could possibly be modified by other genetic or environmental factors. Complex diseases are believed to result from several to many genes and environmental factors and their interactions, genetic heterogeneity and potentially even more reasons, which are yet unknown (Manolio et al. 2009). In this study, we will focus on G×G interactions between SNPs in PD.

PD leads to loss of motor skills and can lead to mental, behavioural, talking difficulties, insomnia, depression, among other issues (Reich and Savitt 2019). PD is a highly age-related disease, while there are cases of early-onset PD, it is rare before the fifth decade (de Lau and Breteler 2006) and usually the onset is at the age of 65-70. Naturally, the prevalence increases with age. In the developed countries, the prevalence of PD is estimated to be 0.3%, roughly 1-2% in the population older than 60 (Nussbaum and Ellis 2003; Capriotti and Terzakis 2016). The male-female ratio in PD increases with age and men are roughly 1.5 times more susceptible to PD (Moisan et al. 2016).

While the exact cause of PD is still unknown, there are two forms – monogenic and sporadic – that are analysed and there is evidence that genetic variability plays a role in both. Around 1-5% of all PD cases are considered to be monogenic and several genes appear to be causal (Singleton and Hardy 2016). However, some of the genes that cause monogenic PD, such as *SNCA*, *LRRK2* and *VPS13C*, among others, seem to play a role in the sporadic disease as well (Trinh et al. 2018). A recent study by Nalls et al. identified 90 genome-wide significant

associations across 78 genomic regions that explained 16-36% of the heritability of PD (depending on the prevalence) (Nalls et al. 2019). Moreover, Blauwendraat et al. have found two genome-wide significant associations between PD with age-at-onset: one in the *SNCA* gene and another in *TMEM175* (Blauwendraat et al. 2019).

G×G interactions, also known as epistasis, can be understood as functional/biological or statistical interaction (Phillips 2008). Biological interaction focuses on the molecular interactions that proteins as well as other genetic elements have with one another and how biological pathways are affected (Boone, Bussey, and Andrews 2007). In this study, we refer to G×G interactions in the statistical sense, meaning effect modification and the departure from additivity in a linear model. It must be noted that statistical interaction does not necessarily imply biological interaction.

In practice, the analysis of G×G interactions poses a computational problem due to the number of possible pairwise SNP combinations, which is a quadratic function of the number (n) of SNPs under study:

$$\frac{n \times (n - 1)}{2}$$

Given say, the 7.8 million SNPs investigated for association with PD in the Nalls et al (2019) paper, this would equal over 30 trillion possible combinations. Since statistical interaction is equivalent to effect modification, it may be more likely that SNPs with proven main effects are involved in G×G interactions. It is therefore advisable to concentrate on G×G interactions among those SNP pairs in which at least one has a proven main effect. Even though G×G without main effects is theoretically possible, to the best of our knowledge, it has never been demonstrated in reality in the context of a complex human disease.

The case-only (CO) design is a statistically powerful approach when it comes to detecting G×G or G×E interactions. In comparison to the case-control (CC) design, it has main advantages in that it obviates the need for proper controls and achieves greater statistical power given the same number of cases (W. J. Gauderman 2002). However, these advantages come at the price of requiring the validity of two assumptions, namely that the disease of interest is sufficiently rare (i.e. has prevalence $\leq 5\%$) and that the two risk factors under study are uncorrelated in the general population (Piegorsch, Weinberg, and Taylor 1994). Linkage disequilibrium, population structure and cryptic relatedness all potentially induce pairwise genotype associations at the population level, which renders the practical utility of CO for G×G studies less straightforward than G×E interactions. In addition, technical artefacts such as genotyping batch effects may also create spurious SNP-SNP associations among cases. Addressing these associations by only pairing SNPs on different chromosome arms, including principal components on the center level and meta-analysing the center-wise results have been shown to retain the validity of the CO design as a means of G×G analysis across potentially heterogeneous centers (Aleknonytė-Resch et al. 2020).

Previous findings of G×G interactions in PD have been mainly focused on a few specific genome regions. The studies either concluded that no interactions in their specific chosen regions were present (Singh et al. 2014; Wider et al. 2011) or focus upon biological interaction,

for example between *PINK1* and *Parkin* in terms of their common pathway in regulating mitochondrial function, or rather, causing mitochondrial dysfunction (Clark et al. 2006; Narendra et al. 2010), which is known to be an important trigger for Parkinson disease (Moore et al. 2005). Multi-loci interactions of three and four SNPs in PD associated genes have also been analysed (Fernández-Santiago et al. 2019). All of these studies used a CC design with less than 2000 individuals each.

In the present study, we examine G×G for PD using the CO design. Our study draws upon one of the largest currently available PD datasets from the International Parkinson Disease Genomics Consortium (IPDGC). As was detailed above, G×G is not very likely in the absence of main effects of both SNPs present, therefore it is plausible to focus the searches for G×G interactions for PD on SNPs with a proven main effect on PD. Thus, our study will focus on the SNPs provided by Nalls et al. (2019), as it is the most up-to-date summary of genetic PD associations available.

Methods

Data

The data used in this study originated from the IPDGC and consisted of 36 362 cases from 16 different centers across Europe and North America. We carried out our analysis on all cases as well as considering two mutually exclusive subsets: patients with early onset PD, where the age of the patient at diagnosis was less than 50 years ($n = 6\,962$), and not early onset of patients of over 50 years of age at diagnosis of PD ($n = 29\,400$). An overview of the number of cases per center can be found in Supplement 1.

Based on the center, genotyping was either conducted with the 650Y, Human660-Quad, Human610K, HumanCNV370 version1_C, HumanHap300 or Infinium BeadChips (Illumina). SNP genotype data from all 16 centers were subject to the same quality control measures as described elsewhere (International Parkinson Disease Genomics Consortium (IPDGC) et al. 2014). Depending on the genotyping chip used, the number of genotyped SNPs ranged from 240 000 to 600 000. After imputation, using a hard-call threshold of 0.8, the number of available SNPs for further analysis equalled to 7.8 million.

For reasons alluded in the introduction, we focused upon SNPs with a known main effect instead of undertaking a full genome-wide search for G×G interactions. Therefore, our study encompassed 90 SNPs with a known main effect from the largest available GWAS study to date (Nalls et al. 2019). We examined all possible pairwise G×G interactions with one SNP confined to one of the 90 SNPs with a main effect and the other with one from the 7.8 million available SNPs. This led to 702 million potential G×G interaction pairs.

Statistical Analysis

All statistical analyses were performed using either R (v. 3.6.2) or PLINK2 (Chang et al. 2015), as appropriate. Two genetic models were examined for encoding genotypes of the 90 SNPs with a known main effect: dominant and recessive. Meanwhile, three genetic models were considered for encoding genotypes of the genome-wide 7.8 million SNPs: dominant, recessive and additive. In the dominant (recessive) model, genotypes (G) were encoded assuming a dominant effect of the minor (major) allele, i.e. $G=1$ for homozygous or heterozygous carriers of the minor (major) allele, $G=0$ for homozygous carriers of the major (minor) allele. The

additive model encoded the predictor variable with 0, 1 or 2 depending on the number of minor alleles and considered the response variable to follow the dominant genetic model.

In the CO design, the genotype of the first SNP (G_1), is treated as a predictor variable, with respective regression coefficient δ_1 representing the interaction effect, while the other genotype (G_2), in our case always a SNP with a known main effect, is treated as the response variable, i.e.

$$\text{logit}\{P(G_2 = 1)\} = \delta_0 + \delta_1 G_1$$

Following Piergosch et al. (Piegorsch, Weinberg, and Taylor 1994), no classical confounders such as age or sex were included in the CO model, because their main effects cannot sensibly be modelled when using a CO design.

However, spurious pair-wise correlation may arise between SNP genotypes when the population under study comprises distinct subpopulations with different allele frequencies (Cardon and Palmer 2003). Genotyping batch effects can also create similar artefacts. To address this problem, we carried out principal component analysis (PCA) of cases at the IPDGC study center level, as the top principal components (PCs) of SNP genotypes have been shown to adjust well for population stratification in previous genetic studies (Price et al. 2006). Thus, the top 10 PCs obtained from all SNPs that passed quality control were included in the logistic regression. For the CO design to be valid, G_1 and G_2 have to be uncorrelated in the general population (Piegorsch, Weinberg, and Taylor 1994). In order to fulfil this requirement, only SNPs on different chromosome arms were included in the analysis thus ensuring that no pairs in strong linkage disequilibrium are tested together. A previous study has shown that such a precaution is necessary to ensure the validity of the CO design (Aleknonytė-Resch et al. 2020).

To allow for heterogeneity of the IPDGC data, the logistic regression was applied for each of the 16 centers separately, yielding center-specific G×G interaction estimates of regression coefficient δ_1 . Then, a meta-analysis fitting a random effects model with inverse variance weights including all centers with reliable OR estimates was carried out using the R package metafor (Viechtbauer 2010) or PLINK2. In scenarios where one or both SNPs in the interaction pair of interest are rare, zero counts in the contingency table of genotype combinations can occur, resulting in unreliable OR estimates which does not allow the meta-analysis to converge. We successively excluded ORs with large confidence intervals until stable results were obtained. Therefore, we included ORs with confidence intervals widths of less than 35. The meta-analysis was followed by a Wald test to assess whether the average δ_1 , taken over centers, was significantly different from zero. Due to the possibility of very rare to zero counts in the contingency table of genotype combinations, a sensitivity analysis was conducted using Simmonds and Higgins model with random effects (Jackson et al. 2018) and a binomial-normal hierarchical model (Günhan, Röver, and Friede 2020) with the latter specifically designed for the inclusion of studies involving rare events.

The analysis was carried out in a joint effort of two research groups. The genome-wide analysis of 702 million potential G×G interaction pairs was computationally and logistically demanding. Since the logistic regression was applied for each of the 16 centers separately, result files consisting of over 11 billion logistic regression outputs were calculated. Thus, in order to manage the output data, a two-step screening process was implemented. First, a meta-

analysis was conducted on only those centers for which the specific SNP pair center-level p-values were less than 0.05. All SNP pairs that had a combined p-value from the meta-analysis of less than 5×10^{-5} were selected for further analysis. A meta-analysis was done for all selected SNPs including all centers. In order to control for the family-wise error rate, the widely accepted genome-wide p-value threshold of 5×10^{-8} was corrected according to Bonferroni by the number of genome-wide G×G interaction searches conducted (90), setting the significance threshold to 5.56×10^{-10} . Given such a low significance threshold of the final analysis, our screening process may make false negative results more likely, however, it should not have produced any false positive results.

The statistical power to detect G×G with a CO design was calculated using the Quanto software (W. James Gauderman 2002) covering a wide spectrum of possible realistic scenarios. These scenarios were made up of a SNP pair in which both SNPs had very low MAFs of 0.05 and another in which both SNPs had a moderate MAF of 0.2. Main effects of one of the SNPs in the pair were set to resemble those from the Nalls et al. (2020) paper. The mean main effect of SNPs with a MAF of less than 0.06, equal to 1.75, was used in the former scenario and the mean main effect over all SNPs, equal to 1.25, in the latter. The MAF combinations were considered given either an optimistic case number including all possible cases ($n = 36\,362$) or a pessimistic assuming only 2 000 cases remained. The type one error rate was set to 5.56×10^{-10} . A comparison with the standard CC study design given the same number of cases and equal to the number of controls was also performed.

Another aspect that had to be taken into consideration when examining genome-wide G×G interactions is extracting unique pairs. Since the genome-wide search included all possible SNPs regardless of how strong the linkage disequilibrium between them was, there were numerous statistically significant findings in the same areas. Thus, locus zoom plots were generated in chromosome regions where the G×G interaction was significant and both SNPs in the interaction pair were within a known gene.

Results

The statistical power analysis highlights difference in statistical power under the various scenarios based on SNP MAF and possible number of cases. In the pessimistic scenario, where information on only 2 000 cases is left available and the CO study design, a G×G interaction OR of 1.9 is needed to achieve $\geq 80\%$ power if the SNP MAF is equal to 0.2. The same power can only be achieved with an interaction OR of 3 if both SNPs have a MAF of 0.05 (Figure 1a). In the optimistic scenario given 36 362 cases, $\geq 80\%$ power is achieved in the CO study design with an interaction OR of 1.12 (1.35) if the SNP MAF is equal to 0.2 (0.05) (Figure 1b). Given the same number of cases, the CO design had greater statistical power to detect G×G interactions than the CC design under all parameter settings considered.

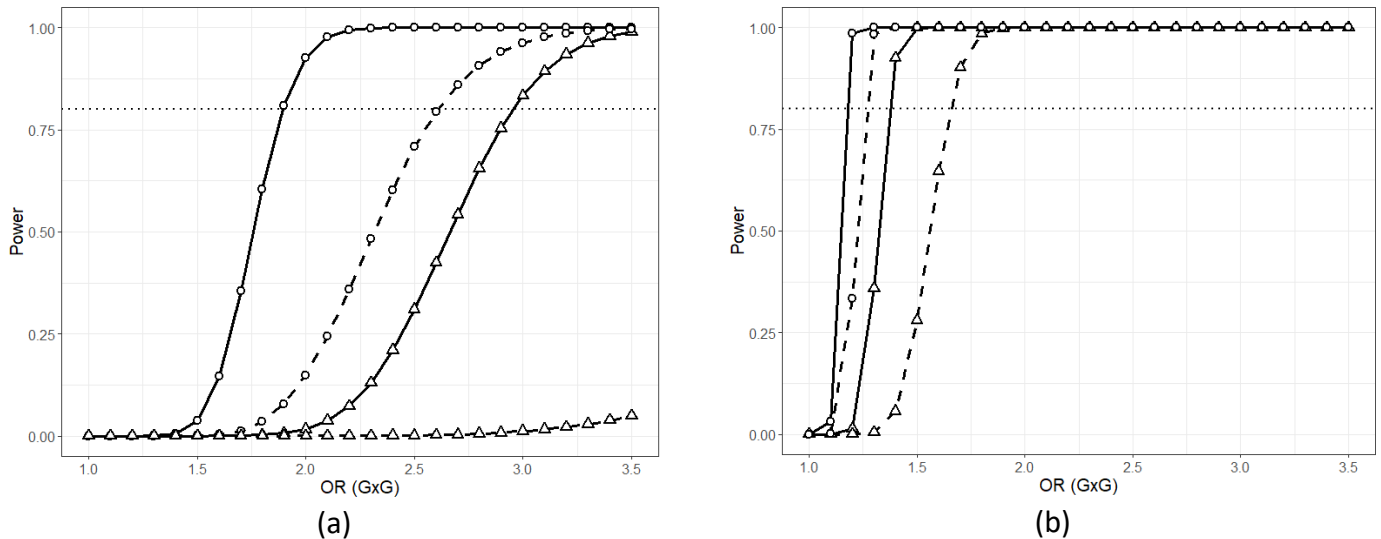
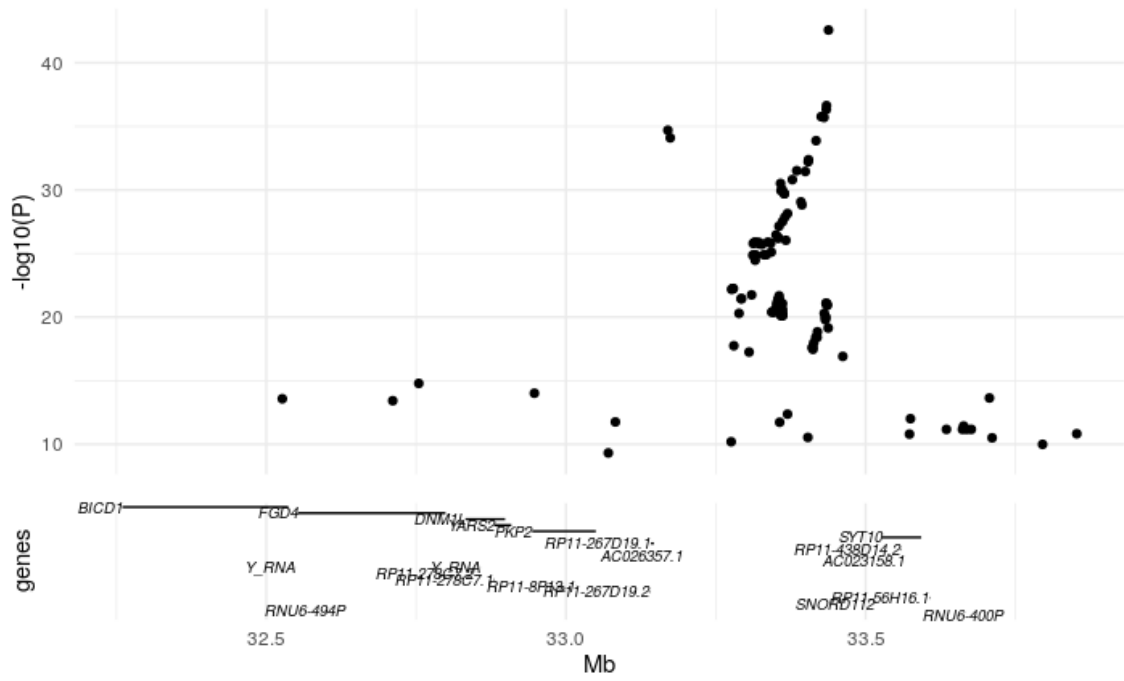


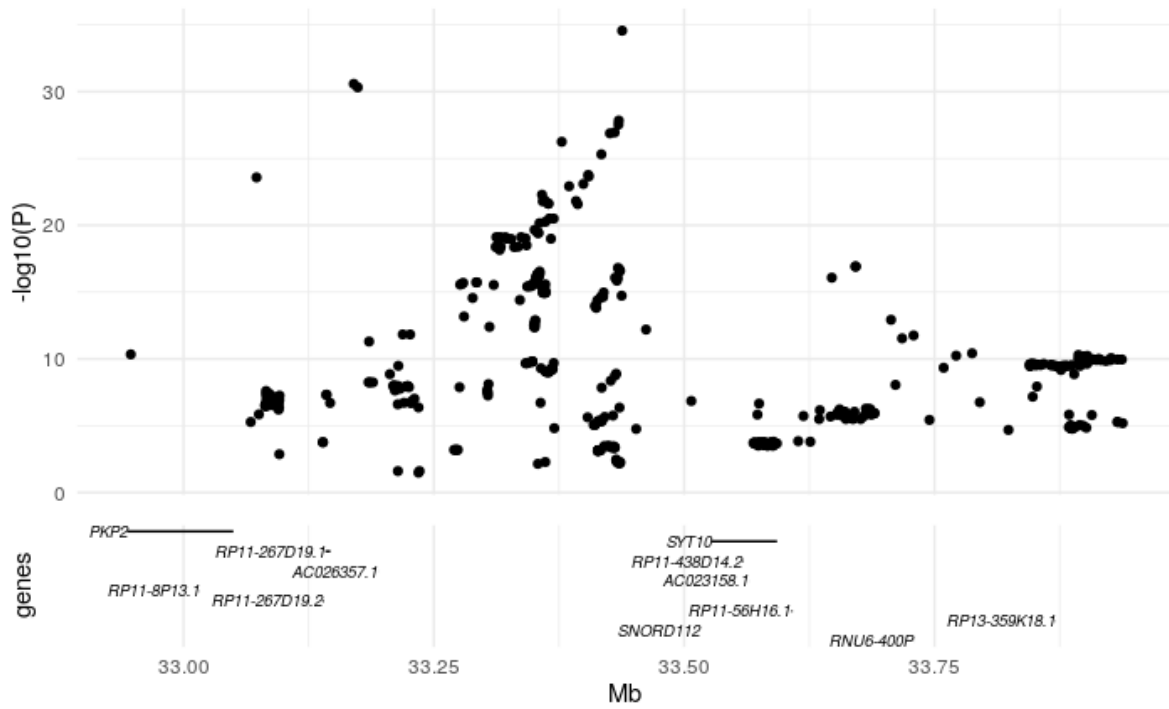
Figure 1: Power of the CO (solid line) and CC (dashed line) designs of G×G analysis (calculated using the Quanto software with parameter settings apt for PD; Bonferroni-corrected significance level: 5.56×10^{-10}) given 2 000 (a) and 36 362 (b) cases. Pairs of SNP allele frequencies are marked by symbols: 0.05, (triangle); 0.2, (circle). Dotted horizontal line marks 80% statistical power.

Out of all the significant SNP×SNP interaction pairs, 10 pairs of unique gene (or gene region) combinations could be identified. The genome-wide search yielded 337 significant SNP×SNP interactions ($p \leq 5.56 \times 10^{-10}$) listed in Supplement 2. Of these significant interactions, 136 were found when considering all available cases and 201 in the subset of cases with not early onset PD. No statistically significant interactions, given the significance level of 5.56×10^{-10} were found when considering the early onset PD data subset separately. The sensitivity analysis showed effects of similar magnitude and, where available, similar p-values (Supplement 3). The forest plots of the SNP×SNP interactions with the lowest p-value for the specific gene pair can be seen in Figures 4-12.

When examining pairs including rs76904798 with a ME (OR = 1.15, $p = 1.52 \times 10^{-28}$, CI = [1.14, 1.17]) on chromosome 12, many significant pairs were found also located on chromosome 12. The SNP rs76904798 is located in the *AC079630.4* antisense gene overlapping with *LRRK2*, according to Ensembl database (Yates et al. 2019). Locus zoom plots of the other SNPs on chromosome 12 in combination with the rs76904798 SNP revealed a unique signal in close proximity to the *SYT10* gene (Figures 2a and 2b). A forest plot depicting the meta-analysis of the G×G interaction of the SNP pair with the lowest p-value of 2.67×10^{-43} and an OR of 1.80 with rs1007709 (CI = [1.65, 1.95]) is seen in Figure 3.



(a)



(b)

Figure 2: Locus zoom plots of chromosome 12. Nalls SNP rs76904798 in *AC079630.4* on chromosome 12.

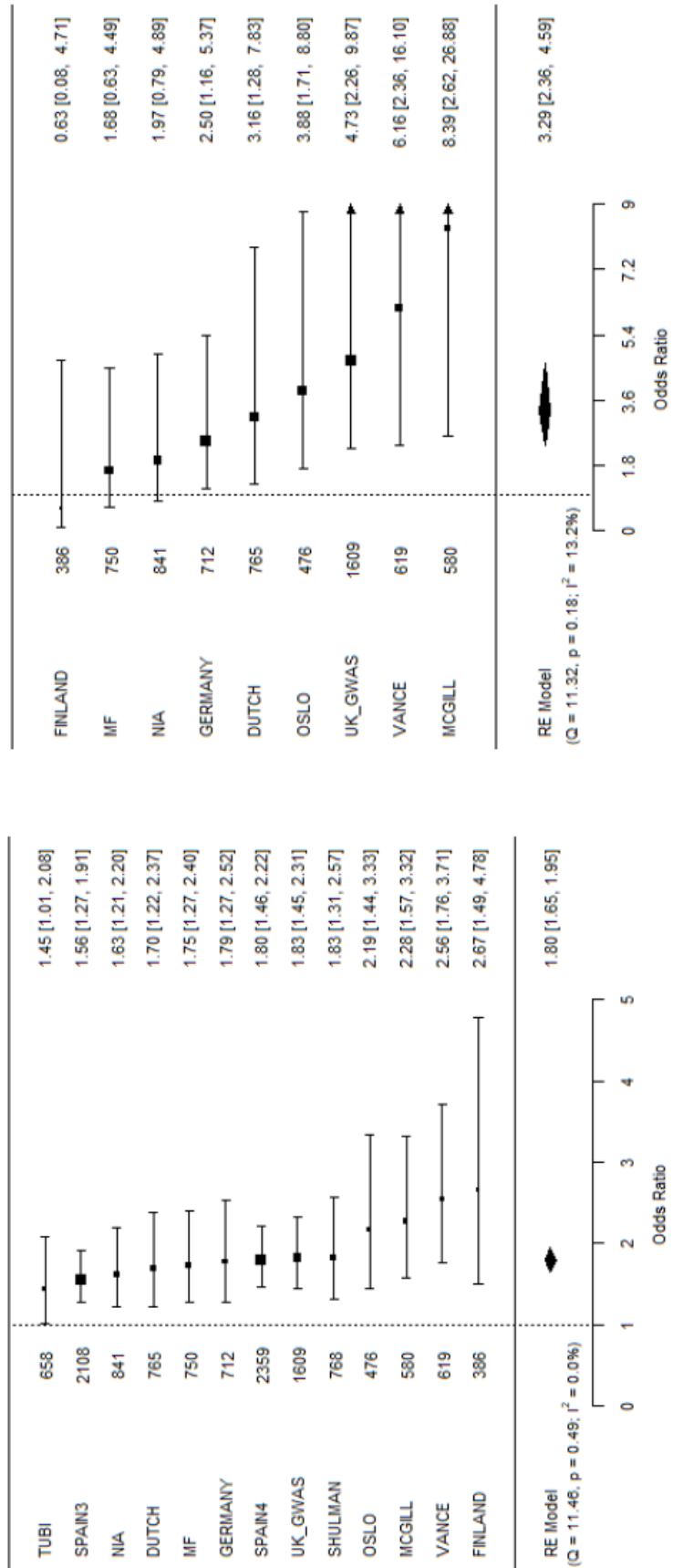


Figure 3: Interaction between rs76904798 in AC079630.4 on chromosome 12 and rs1007709 near STY10 on chromosome 12, all cases, genetic models: dominant Nalls SNP and additive other SNP. $p = 2.67 \times 10^{-43}$

Figure 4: Interaction between rs76904798 in AC079630.4 on chromosome 12 and rs187879258 in BICD1 on chromosome 12, all cases, genetic models: dominant Nalls SNP and additive other SNP. $p = 2.61 \times 10^{-12}$

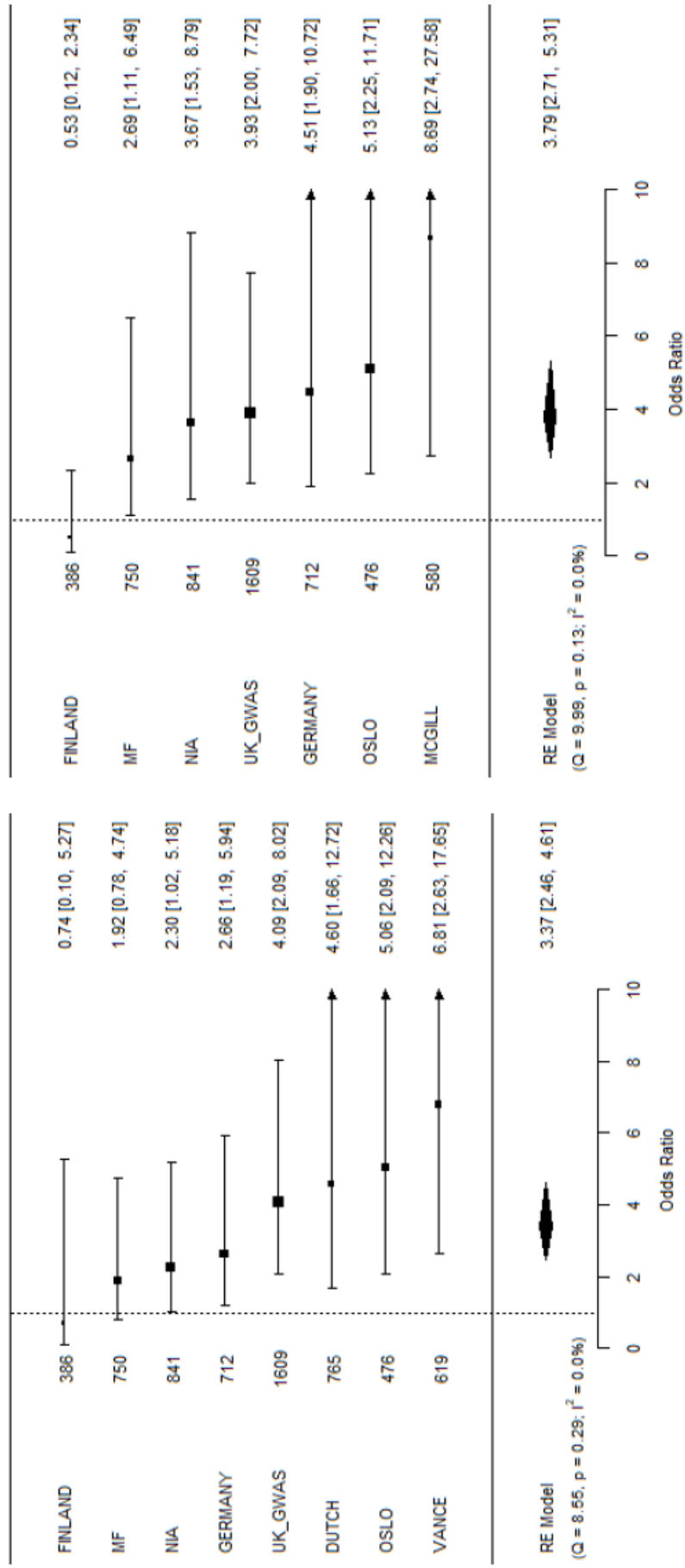


Figure 5: Interaction between rs76904798 in AC079630.4 on chromosome 12 and rs117835488 in FGD4 on chromosome 12, all cases, genetic models: dominant Nalls SNP and additive other SNP. p = 3.79x10⁻¹⁴

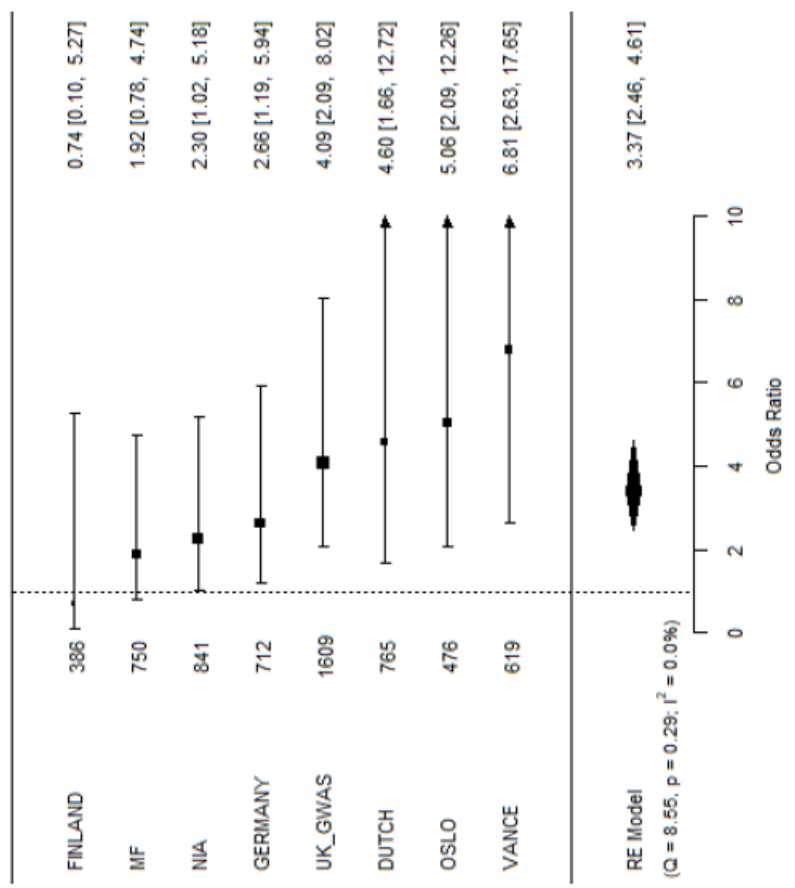


Figure 6: Interaction between rs76904798 in AC079630.4 on chromosome 12 and rs139007869 in PKP2 on chromosome 12, all cases, genetic models: dominant Nalls SNP and additive other SNP. p = 9.66x10⁻¹⁵

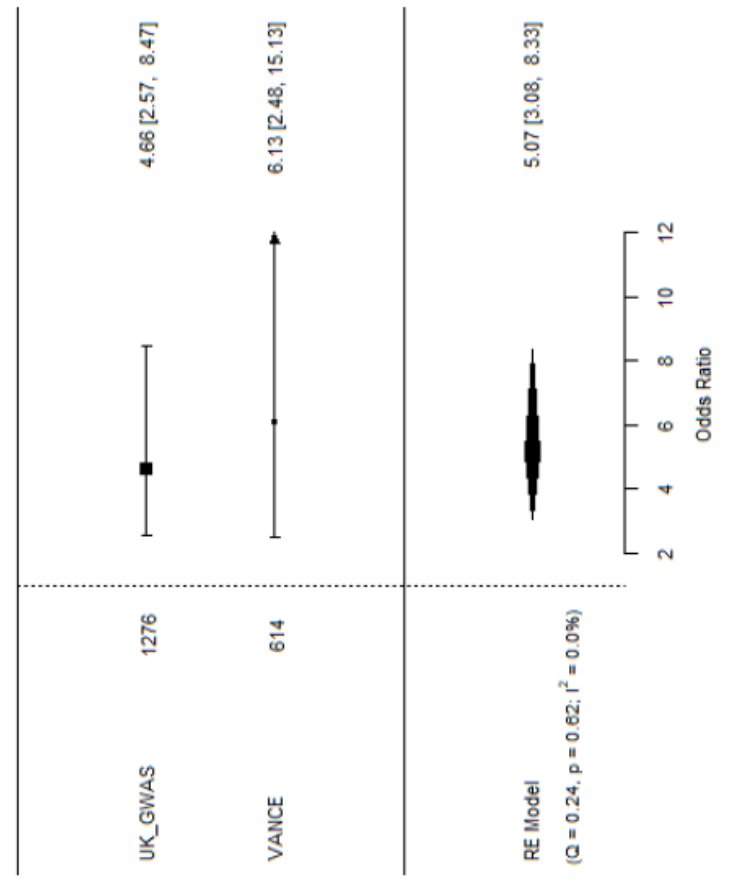


Figure 7: Interaction between rs76904798 in AC079630.4 on chromosome 12 and rs150084348 on chromosome 12, not early onset PD subset, genetic models: dominant Nalls SNP and additive other SNP. p = 8.10x10⁻¹²

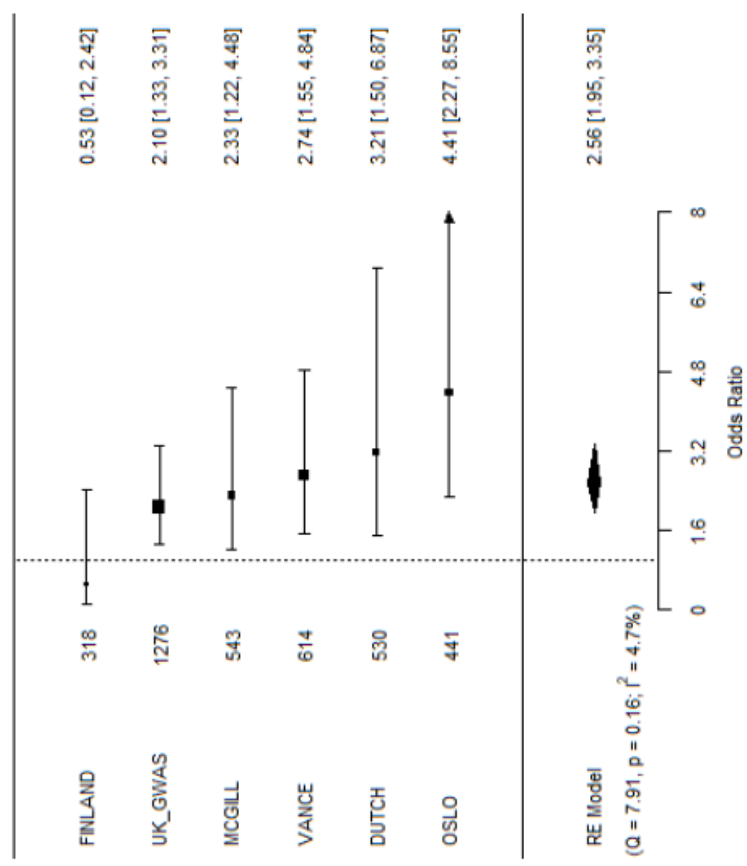


Figure 8: Interaction between rs76904798 in AC079630.4 on chromosome 12 and rs141945110 on chromosome 12, not early onset PD subset, genetic models: dominant Nalls SNP and additive other SNP. p = 1.64x10⁻¹⁰

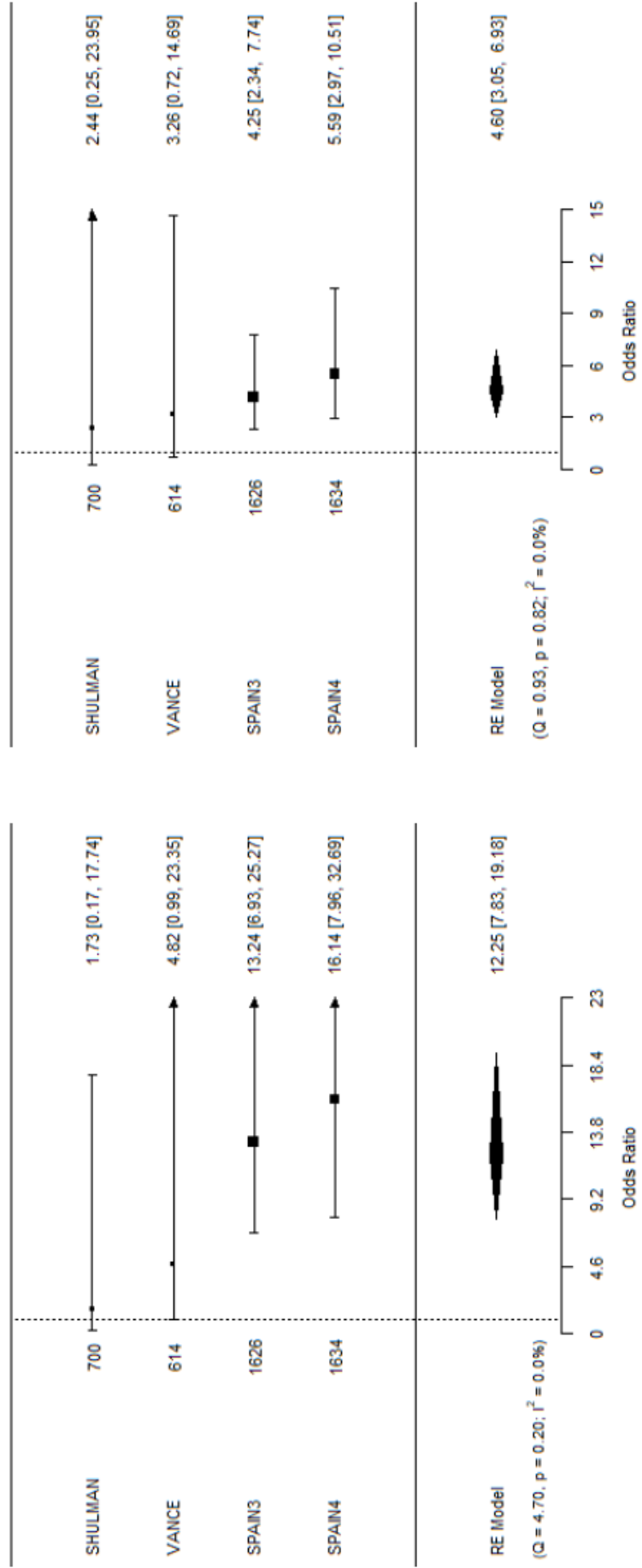


Figure 9: Interaction between rs34637584 in *LRRK2* on chromosome 12 and rs151094822 in *FGD4* on chromosome 12, not early onset PD subset, genetic models: dominant Nalls SNP and dominant other SNP. $p = 5.81 \times 10^{-28}$

Figure 10: Interaction between rs34637584 in *LRRK2* on chromosome 12 and rs80291054 in *PKP2* on chromosome 12, not early onset PD subset, genetic models: dominant Nalls SNP and dominant other SNP. $p = 3.26 \times 10^{-13}$

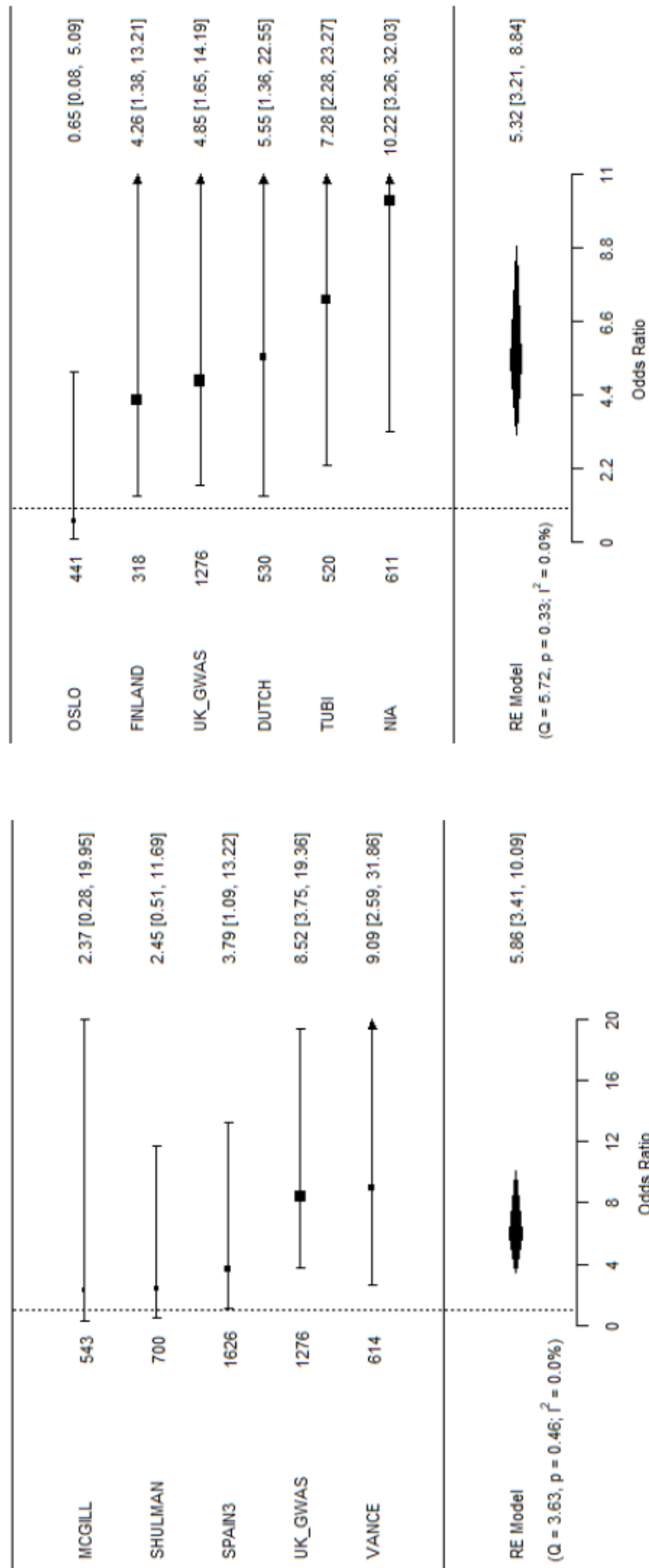


Figure 11: Interaction between rs112485576 on chromosome 6 and rs7856915 in KANK1 on chromosome 9, not early onset PD subset, genetic models: recessive Nalls SNP and recessive other SNP. $p = 1.73 \times 10^{-10}$

Figure 12: Interaction between rs26431 in PAM on chromosome 5 and rs139186308 in TTC6 on chromosome 14, not early onset PD subset, genetic models: recessive Nalls SNP and dominant other SNP. $p = 9.75 \times 10^{-11}$

Five of the nine significant unique gene-gene combinations include rs76904798, which is located on chromosome 12 in the *AC079630.4* gene, an antisense gene according to the Ensembl database (Yates et al. 2019). The nearest protein coding gene defined by Nalls et al. (2019) for rs76904798 is *LRRK2*. This SNP has a known main effect (OR = 1.15, $p = 1.52 \times 10^{-28}$, CI = [1.14, 1.17]) for PD according to Nalls et al. (2019). Significant association with other SNPs in protein coding genes also on chromosome 12, all at least 6 Mb away and showing no LD in the 1000 Genomes reference base, include rs187879258 in the *BICD1* gene (Figure 4, interaction OR = 3.29, CI = [2.36-4.59], $p = 2.61 \times 10^{-12}$), rs117835488 in the *FGD4* gene (Figure 5, interaction OR = 3.37, CI = [2.46-4.61], $p = 3.79 \times 10^{-14}$), rs139007869 in the *PKP2* gene (Figure 6, interaction OR = 3.79, CI = [2.71-5.31], $p = 9.66 \times 10^{-15}$). Significant G×G interaction effects were also observed with rs150084348 on chromosome 12, which is in *RNU6-400P*, a snRNA (Figure 7, interaction OR = 3.79, CI = [2.71-5.31], $p = 8.10 \times 10^{-12}$) and rs141945110, which is in *RP11-313F23.4.*, a lincRNA (Figure 8, interaction OR = 5.07, CI = [3.98-8.33], $p = 1.64 \times 10^{-10}$).

The rs34637584 SNP, also located on chromosome 12 in the *LRRK2* gene, 0.1 Mb away from rs76904798, was included in two unique significant G×G interactions with SNPs located in other protein coding genes. This SNP has a known main effect (OR = 11.35, $p = 3.61 \times 10^{-148}$, CI = [10.33, 12.46]) for PD according to Nalls et al. (2019). In our dataset it had a very low MAF of 0.0021. Significant association with other SNPs in protein coding genes also on chromosome 12 include: rs151094822 in the *FGD4* gene (Figure 9, interaction OR = 12.25, CI = [7.83-19.18], $p = 5.81 \times 10^{-28}$) and rs80291054 in the *PKP2* gene (Figure 10, interaction OR = 4.60, CI = [3.05-6.93], $p = 3.23 \times 10^{-13}$).

Finally, two SNP pairs on different chromosomes showed statistically significant G×G interactions. One of the pairs consisted of rs112485576 on chromosome 6 with the nearest gene being *HLA-DRB5*. This SNP has a known main effect (OR = 0.85, $p = 6.96 \times 10^{-28}$, CI = [0.83, 0.86]) for PD according to Nalls et al. (2019). The other SNP in the pair was rs7856915 on chromosome 9, located in the *KANK1* gene. The G×G interactions significance was 1.73×10^{-10} (Figure 11, OR = 5.83, CI = [3.38, 10.03]). The other SNP pair consisted of rs26431 and rs139186308 (Figure 12, interaction OR = 5.32, CI = [3.21-8.84], $p = 9.75 \times 10^{-11}$). rs26431 is located in the *PAM* gene and has a known main effect (OR = 0.85, $p = 6.96 \times 10^{-28}$, CI = [0.83, 0.86]) for PD according to Nalls et al. (2019), while rs139186308 is found in the *TTC6* gene.

Discussion

Using one of the largest available datasets of PD cases from the IPDGC, our investigation of G×G interactions using the statistically powerful CO study design found ten statistically significant ($p < 5.56 \times 10^{-10}$) unique signals across two subsets (early and not early onset PD) and all cases together as one dataset. A logistic regression was used to analyse G×G interactions with ten PCs included within the regression for each center separately and the center-wise results were meta-analysed fitting a random effects model.

G×G interactions pose problems due to the high number of possible pair combinations that is only increased if various centers of heterogeneous origin are included in the analysis, as these interactions need to be computed for each center separately. This becomes a computational, time, data storage and transfer problem. Moreover, the family-wise error rate needs to be controlled as the widely accepted genome-wide significance level of 5×10^{-8} needs to be

adjusted by the number of times a genome-wide search was executed. Therefore, we proposed to ensure that one of the SNPs in the possible interaction pair is a SNP with a known main effect for the disease phenotype. This reduces the number of times a genome-wide search needs to be done and reduces the computational burden. Furthermore, a screening of low center-wise p values and including only those pairs with at least one low p value in any center considerably reduces the data size. In our case, the reduction was from 702 to one million SNP pairs. Such a screening method would not cause any false positives. While it may exclude some true positives, we conclude that the probability is small, if the initial screening p value is high (say, the nominal 0.05) and the final threshold of statistical significance is very low (say, 1×10^{-10}).

Another issue in G×G interaction analysis is the role and results of SNPs with low MAFs (MAF<0.05). On the one hand, SNPs with MEs in rare diseases often have low MAFs. In the Nalls et al (2019) paper, 13 out of 90 SNPs with MEs had a MAF of less than 0.05. On the other hand, in G×G interaction analysis, if both SNPs are rare, a situation may occur in which there are not enough minor allele carriers for the logistic regression to converge, resulting in missing information from whole centers. Or, the CI becomes so wide that it is neither plausible, nor allows the meta-analysis to converge. Low SNP MAFs lead to an unbalanced design and have low statistical power in general. This problem can increase depending on the genetic model (dominant, additive or recessive) assumed, as even fewer homozygous minor allele carriers would be present in a recessive model. Sensitivity analyses using different meta-analysis methods (Jackson et al. 2018) especially those designed for studies involving rare events (Günhan, Röver, and Friede 2020) aid in evaluating the robustness of the results. Furthermore, SNPs with low MAFs may also become problematic when imputed, especially in areas with MEs (as discussed in Chapter 2.1.). While hard-calls with a higher threshold can be used, it increases the number of missing values and does not solve the issue of incorrectly imputed SNPs with low MAFs. Thus, it may be that G×G interactions with low MAFs remain undetected.

Nevertheless, our study found ten unique G×G interactions. Six of these included rs76904798 in the *AC079630.4* antisense gene with a known main effect, which overlaps with the *LRRK2* gene. The pair with the highest density of significant SNP pairs was with rs1007709, located in close proximity to the promoter region of *SYT10*. *LRRK2* is strongly associated with PD and the molecular function of this gene is directly associated with neural plasticity (Matikainen-Ankney et al. 2018). Meanwhile, *SYT10* contributes to the exocytosis of secretory vesicles in neurons (Cao, Yang, and Sudhof 2013), therefore a biological interaction may be plausible between these genes. Two other genes, namely *FGD4* and *PKP2* stand out. SNPs located in their regions formed statistically significant pairs with SNPs from within *AC079630.4* and directly *LRRK2*. Phenotypes associated with *FGD4* include systolic blood pressure (Stelzer et al. 2016), while high blood pressure is a predictor for motor decline in PD (Kotagal et al. 2014). Phenotypes for *PKP2* include heart rate response to exercise (Stelzer et al. 2016), which is an issue for PD patients (Speelman et al. 2012). These interactions have biological plausibility, laying ground for further specific research of these pairs in their combined role in PD.

Supplement 1: Number of cases per center.

Center	Early Onset	Not Early Onset	All
Dutch	1083	1664	2747
Finland	82	787	869
Germany	743	833	1576
HBS	39	955	994
McGill	303	1182	1485
NEUROX_DBGAP	2107	9102	11209
NIA	466	3376	3842
Oslo	117	821	938
PDBP	74	715	789
PPMI	49	476	525
Shulman	72	880	952
Spain3	498	2704	3202
Spain4	957	2690	3647
Tubi	150	1042	1192
UK_GWAS	202	1276	1478
Vance	20	897	917
Sum	6962	29400	36362

Supplement 2: Significant SNP×SNP interactions ($p \leq 5.56 \times 10^{-10}$). * - genetic model with regard to the minor allele is coded as REC for recessive, DOM for dominant and ADD for additive. † - dataset is coded as NEO for not early onset PD and All for all PD cases.

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset†	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs114412656	4		REC	rs13117519	1		DOM	NEO	7	4523	1.64	0.25	1.20E-10	1.14	2.14
rs139186308	5	PAM	REC	rs26431	14	TTC6	DOM	NEO	6	3696	1.67	0.26	9.75E-11	1.17	2.18
rs6477081	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.76	0.28	1.97E-10	1.23	2.31
rs7865421	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.76	0.28	1.93E-10	1.22	2.31
rs7856915	6		REC	rs112485576	9	KANK1	REC	NEO	5	4759	1.77	0.28	1.73E-10	1.22	2.31
rs9407331	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.09E-10	1.20	2.28
rs9407332	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.09E-10	1.20	2.28
rs9407333	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.09E-10	1.20	2.28
rs9407334	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.09E-10	1.20	2.28
rs2361097	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.09E-10	1.20	2.28
rs2361098	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.09E-10	1.19	2.28
rs7044387	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.19	2.28
rs7851435	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.20	2.28
rs7863364	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.20	2.28
rs7863471	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.20	2.28
rs7875094	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.20	2.28
rs7851766	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.20	2.28
rs74307570	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.11E-10	1.19	2.28
rs4259462	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.11E-10	1.19	2.28
rs4463482	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.11E-10	1.19	2.28
rs4259463	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.11E-10	1.19	2.28
rs9408648	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.02E-10	1.20	2.28
rs7851799	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.19	2.28
rs7851886	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.19	2.28

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset*	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs7863869	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.19	2.28
rs72691349	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.73	0.28	3.71E-10	1.19	2.28
rs16921959	6		REC	rs504594	9	KANK1	REC	NEO	5	4759	1.74	0.28	3.51E-10	1.19	2.28
rs74688167	7		DOM	rs76949143	7		ADD	All	13	12631	1.19	0.11	1.03E-25	0.97	1.41
rs74688167	7		DOM	rs76949143	7		DOM	All	13	12631	1.22	0.11	1.77E-26	1.00	1.45
rs74354038	7		DOM	rs76949143	7		ADD	All	13	12631	1.19	0.11	1.03E-25	0.97	1.41
rs74354038	7		DOM	rs76949143	7		DOM	All	13	12631	1.22	0.11	1.77E-26	1.00	1.45
rs77298580	7		DOM	rs76949143	7		ADD	All	13	12631	1.19	0.11	1.03E-25	0.97	1.41
rs77298580	7		DOM	rs76949143	7		DOM	All	13	12631	1.22	0.11	1.77E-26	1.00	1.45
rs77469578	7		DOM	rs76949143	7		ADD	All	13	12631	1.19	0.11	1.03E-25	0.97	1.41
rs77469578	7		DOM	rs76949143	7		DOM	All	13	12631	1.22	0.11	1.77E-26	1.00	1.45
rs78879853	7		DOM	rs76949143	7		ADD	All	13	12631	1.19	0.11	1.03E-25	0.97	1.41
rs78879853	7		DOM	rs76949143	7		DOM	All	13	12631	1.22	0.11	1.77E-26	1.00	1.45
rs12671726	7		DOM	rs76949143	7		ADD	All	12	11973	1.19	0.13	7.08E-19	0.93	1.46
rs12671726	7		DOM	rs76949143	7		DOM	All	12	11973	1.22	0.14	1.74E-18	0.95	1.50
rs60458555	7		DOM	rs76949143	7		ADD	All	12	11973	1.19	0.13	7.08E-19	0.93	1.46
rs60458555	7		DOM	rs76949143	7		DOM	All	12	11973	1.22	0.14	1.74E-18	0.95	1.50
rs60122908	7		DOM	rs76949143	7		ADD	All	12	11973	1.19	0.13	7.08E-19	0.93	1.46
rs60122908	7		DOM	rs76949143	7		DOM	All	12	11973	1.22	0.14	1.74E-18	0.95	1.50
rs74807171	7		DOM	rs76949143	7		ADD	All	12	11973	1.15	0.15	9.94E-15	0.86	1.45
rs74807171	7		DOM	rs76949143	7		DOM	All	12	11973	1.19	0.15	7.64E-15	0.89	1.48
rs79394524	7		DOM	rs76949143	7		ADD	All	12	11973	1.15	0.15	9.94E-15	0.86	1.45
rs79394524	7		DOM	rs76949143	7		DOM	All	12	11973	1.19	0.15	7.64E-15	0.89	1.48
rs78678506	7		DOM	rs76949143	7		ADD	All	12	11973	1.15	0.15	9.94E-15	0.86	1.45
rs78678506	7		DOM	rs76949143	7		DOM	All	12	11973	1.19	0.15	7.64E-15	0.89	1.48
rs79693828	7		DOM	rs76949143	7		ADD	All	12	11973	1.15	0.15	9.94E-15	0.86	1.45
rs79693828	7		DOM	rs76949143	7		DOM	All	12	11973	1.19	0.15	7.64E-15	0.89	1.48

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset [†]	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs56201751	7		DOM	rs76949143	7		ADD	All	10	7506	1.22	0.19	1.59E-10	0.85	1.59
rs12667883	7		DOM	rs76949143	7		ADD	All	12	11973	1.16	0.15	6.36E-15	0.87	1.45
rs12667883	7		DOM	rs76949143	7		DOM	All	12	11973	1.19	0.15	5.04E-15	0.89	1.48
rs77223318	7		DOM	rs76949143	7		ADD	All	12	11973	1.16	0.15	1.14E-14	0.87	1.46
rs77223318	7		DOM	rs76949143	7		DOM	All	12	11973	1.19	0.15	8.10E-15	0.89	1.49
rs117870187	7		DOM	rs76949143	7		ADD	All	12	11973	1.16	0.15	7.81E-15	0.87	1.46
rs117870187	7		DOM	rs76949143	7		DOM	All	12	11973	1.19	0.15	5.67E-15	0.89	1.49
rs77092740	7		DOM	rs76949143	7		ADD	All	12	11973	1.16	0.15	7.81E-15	0.87	1.46
rs77092740	7		DOM	rs76949143	7		DOM	All	12	11973	1.19	0.15	5.67E-15	0.89	1.49
rs187879258	12	AC079630.4	DOM	rs76904798	12	BICD1	ADD	All	9	6738	1.19	0.17	2.61E-12	0.86	1.52
rs117835488	12	AC079630.4	DOM	rs76904798	12	FGD4	ADD	All	8	6158	1.21	0.16	3.79E-14	0.90	1.53
rs117835488	12	AC079630.4	DOM	rs76904798	12	FGD4	ADD	NEO	6	3679	1.29	0.21	3.07E-10	0.89	1.70
rs41276676	12	AC079630.4	DOM	rs76904798	12	FGD4	ADD	All	8	6158	1.29	0.20	2.86E-10	0.89	1.69
rs139007869	12	AC079630.4	DOM	rs76904798	12	PKP2	ADD	All	7	5354	1.33	0.17	9.66E-15	1.00	1.67
rs112664776	12	AC079630.4	DOM	rs76904798	12		DOM	All	13	12631	0.63	0.06	1.56E-24	0.51	0.75
rs112664776	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.70	0.07	2.62E-24	0.56	0.83
rs112664776	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.72	0.07	3.10E-24	0.58	0.86
rs55969207	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.21	0.03	1.35E-10	0.15	0.27
rs117561021	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	1.11	0.09	2.02E-35	0.93	1.28
rs117561021	12	AC079630.4	DOM	rs76904798	12		DOM	All	13	12631	1.12	0.09	4.59E-35	0.94	1.30
rs117561021	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	1.20	0.10	2.65E-31	1.00	1.40
rs117561021	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	1.22	0.10	5.51E-31	1.01	1.42
rs140305553	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	1.10	0.09	8.06E-35	0.92	1.27
rs140305553	12	AC079630.4	DOM	rs76904798	12		DOM	All	13	12631	1.11	0.09	2.69E-34	0.93	1.28
rs140305553	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	1.19	0.10	4.79E-31	0.99	1.40
rs140305553	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	1.21	0.11	1.23E-30	1.00	1.42
rs61927365	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.41	0.06	3.25E-11	0.29	0.53

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset†	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs17555138	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.41	0.06	3.37E-11	0.29	0.53
rs4931058	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.41	0.06	3.46E-11	0.29	0.53
rs1445310	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	-0.25	0.03	1.84E-18	-0.31	-0.20
rs1445310	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	-0.25	0.03	6.74E-14	-0.32	-0.19
rs12423471	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.37	0.06	2.44E-10	0.26	0.49
rs55801348	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.39	0.06	8.79E-11	0.27	0.51
rs55801348	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.37	0.06	2.18E-10	0.26	0.49
rs12427282	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.39	0.06	8.79E-11	0.27	0.51
rs12427282	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.37	0.06	2.18E-10	0.26	0.49
rs1160292	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.38	0.06	1.12E-11	0.27	0.49
rs1160292	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.36	0.06	6.61E-11	0.25	0.47
rs79186412	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	4.93E-12	0.29	0.52
rs79186412	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	6.43E-12	0.30	0.54
rs12304318	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	2.11E-13	0.30	0.52
rs12304318	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	4.86E-13	0.31	0.53
rs12304318	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.46	0.06	9.22E-13	0.33	0.58
rs12425568	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	4.93E-12	0.29	0.52
rs12425568	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	6.43E-12	0.30	0.54
rs17631655	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	4.93E-12	0.29	0.52
rs17631655	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	6.43E-12	0.30	0.54
rs7967103	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	2.11E-13	0.30	0.52
rs7967103	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	3.94E-13	0.31	0.53
rs7967103	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.46	0.06	9.22E-13	0.33	0.58
rs55638410	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	4.93E-12	0.29	0.52
rs55638410	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	6.43E-12	0.30	0.54
rs10506105	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	1.10E-11	0.29	0.52
rs10506105	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	1.45E-11	0.30	0.54

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset†	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs12317335	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	2.11E-13	0.30	0.52
rs12317335	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	3.94E-13	0.31	0.53
rs12317335	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.46	0.06	9.22E-13	0.33	0.58
rs61927377	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	4.93E-12	0.29	0.52
rs61927377	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	6.43E-12	0.30	0.54
rs4931694	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	2.11E-13	0.30	0.52
rs4931694	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	3.94E-13	0.31	0.53
rs4931694	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.46	0.06	9.22E-13	0.33	0.58
rs12313340	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	2.59E-13	0.30	0.52
rs12313340	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	4.91E-13	0.31	0.53
rs12313340	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.46	0.06	1.31E-12	0.33	0.58
rs4931696	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	1.99E-13	0.30	0.52
rs4931696	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	3.78E-13	0.31	0.53
rs4931696	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.45	0.06	8.86E-13	0.33	0.58
rs12423297	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	1.07E-11	0.29	0.53
rs12423297	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	1.82E-11	0.30	0.54
rs1445313	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	1.34E-11	0.29	0.53
rs1445313	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	2.07E-11	0.30	0.54
rs7967146	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.27	0.04	1.60E-12	0.20	0.35
rs9989005	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	1.29E-12	0.30	0.53
rs9989005	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	2.24E-12	0.30	0.54
rs9989005	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.46	0.07	7.11E-12	0.33	0.59
rs11052506	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	1.21E-12	0.30	0.52
rs11052506	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	2.17E-12	0.30	0.54
rs11052506	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.46	0.07	6.23E-12	0.33	0.59
rs4931700	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.06	1.33E-11	0.29	0.53
rs4931700	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	1.46E-11	0.30	0.55

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset*	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs7970713	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	9.76E-13	0.24	0.43
rs7970713	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	7.13E-12	0.24	0.43
rs4931701	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	1.52E-12	0.24	0.43
rs4931701	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.07E-11	0.24	0.43
rs4325386	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	1.56E-12	0.24	0.43
rs4325386	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.07E-11	0.24	0.43
rs11561220	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	1.02E-12	0.24	0.43
rs11561220	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	6.39E-12	0.24	0.43
rs11052511	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	9.53E-13	0.25	0.43
rs11052511	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	5.28E-12	0.24	0.44
rs9943817	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.24	0.03	3.15E-13	0.18	0.31
rs11052513	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.24	0.03	2.61E-13	0.18	0.31
rs11559813	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.24	0.03	2.61E-13	0.18	0.31
rs11537416	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.24	0.03	4.22E-13	0.18	0.31
rs11536034	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.24	0.03	4.22E-13	0.18	0.31
rs61927401	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.42	0.06	6.74E-13	0.30	0.53
rs61927401	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.43	0.06	1.34E-12	0.31	0.55
rs61927401	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.47	0.07	5.16E-12	0.33	0.60
rs11532228	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.24	0.03	2.61E-13	0.18	0.31
rs10844494	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.25	0.03	1.33E-13	0.18	0.31
rs10844495	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.25	0.03	1.33E-13	0.18	0.31
rs11052514	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	4.72E-13	0.25	0.43
rs11052514	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	2.31E-12	0.25	0.44
rs11052515	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	6.20E-13	0.25	0.43
rs11052515	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	2.66E-12	0.25	0.44
rs12813315	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	4.01E-13	0.25	0.43
rs12813315	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.73E-12	0.25	0.44

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset+	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs11052518	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	3.52E-13	0.25	0.43
rs11052518	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.35	0.05	1.38E-12	0.25	0.44
rs1822885	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	4.19E-13	0.25	0.43
rs1822885	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.49E-12	0.25	0.44
rs1822886	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	2.54E-13	0.25	0.43
rs1822886	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.35	0.05	9.21E-13	0.25	0.44
rs4931702	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.42	0.06	3.46E-12	0.30	0.54
rs4931702	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.43	0.06	5.76E-12	0.31	0.55
rs4931065	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	3.39E-13	0.25	0.44
rs4931065	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.49E-12	0.25	0.44
rs4931066	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	2.01E-13	0.25	0.44
rs4931066	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.35	0.05	8.36E-13	0.25	0.44
rs4492898	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	3.39E-13	0.25	0.44
rs4492898	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.49E-12	0.25	0.44
rs4285963	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	3.39E-13	0.25	0.44
rs4285963	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.49E-12	0.25	0.44
rs112233518	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.42	0.06	3.46E-12	0.30	0.54
rs112233518	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.43	0.06	5.76E-12	0.31	0.55
rs7969274	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	3.72E-13	0.25	0.43
rs7969274	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.49E-12	0.25	0.44
rs7969303	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	2.26E-13	0.25	0.43
rs7969303	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.35	0.05	9.21E-13	0.25	0.44
rs1902777	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.42	0.06	1.38E-13	0.31	0.53
rs1902777	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.43	0.06	3.48E-13	0.31	0.55
rs1902777	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.47	0.07	6.00E-13	0.34	0.60
rs1902776	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	2.89E-13	0.25	0.44
rs1902776	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.35	0.05	1.23E-12	0.25	0.44

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset+	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs1902775	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	4.56E-14	0.25	0.43
rs1902775	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.35	0.05	2.22E-13	0.25	0.44
rs7973462	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	3.09E-13	0.25	0.43
rs7973462	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	8.17E-12	0.24	0.44
rs4931067	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.49	0.06	2.25E-18	0.38	0.60
rs4931067	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.43	0.05	3.95E-18	0.33	0.53
rs4931067	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.44	0.05	3.22E-17	0.34	0.55
rs1545643	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	6.03E-13	0.25	0.43
rs1545643	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.10E-11	0.24	0.43
rs1545642	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	6.03E-13	0.25	0.43
rs1545642	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.10E-11	0.24	0.43
rs61927403	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.48	0.06	2.03E-18	0.38	0.59
rs61927403	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.42	0.05	3.41E-18	0.33	0.52
rs61927403	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.44	0.05	1.59E-17	0.34	0.54
rs11052525	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.33	0.05	5.45E-13	0.24	0.42
rs11052525	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.33	0.05	1.09E-11	0.24	0.43
rs1457682	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.33	0.05	5.45E-13	0.24	0.42
rs1457682	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.33	0.05	1.09E-11	0.24	0.43
rs10844499	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.33	0.05	5.26E-13	0.24	0.42
rs10844499	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.33	0.05	9.84E-12	0.24	0.43
rs7136789	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	6.03E-13	0.25	0.43
rs7136789	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.10E-11	0.24	0.43
rs1545641	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	6.03E-13	0.25	0.43
rs1545641	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.10E-11	0.24	0.43
rs1545640	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	6.03E-13	0.25	0.43
rs1545640	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	1.10E-11	0.24	0.43
rs7961592	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.33	0.05	5.45E-13	0.24	0.42

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset*	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs7961592	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.33	0.05	1.09E-11	0.24	0.43
rs7961717	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.33	0.05	5.50E-13	0.24	0.42
rs7961717	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.33	0.05	1.10E-11	0.24	0.43
rs7961931	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.49	0.06	2.12E-18	0.38	0.59
rs7961931	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.42	0.05	3.54E-18	0.33	0.52
rs7961931	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.44	0.05	1.67E-17	0.34	0.54
rs7962018	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.33	0.05	5.45E-13	0.24	0.42
rs7962018	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.33	0.05	1.09E-11	0.24	0.43
rs7307554	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	6.17E-13	0.25	0.44
rs7307554	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.34	0.05	2.09E-11	0.24	0.44
rs56169515	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.42	0.06	7.77E-14	0.31	0.53
rs56169515	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.44	0.06	5.16E-13	0.32	0.55
rs7307442	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.34	0.05	5.23E-13	0.24	0.43
rs7307442	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.33	0.05	1.19E-11	0.24	0.43
rs7307622	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.33	0.05	5.45E-13	0.24	0.42
rs7307622	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.33	0.05	1.09E-11	0.24	0.43
rs56905231	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.48	0.05	1.68E-18	0.37	0.59
rs56905231	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.42	0.05	4.14E-18	0.33	0.52
rs56905231	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.44	0.05	1.83E-17	0.34	0.54
rs61370554	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.48	0.05	1.68E-18	0.37	0.59
rs61370554	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.42	0.05	4.14E-18	0.33	0.52
rs61370554	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.44	0.05	1.83E-17	0.34	0.54
rs74348260	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.43	0.06	1.61E-14	0.32	0.53
rs74348260	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.44	0.06	3.78E-14	0.33	0.55
rs73302078	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.41	0.05	9.99E-14	0.30	0.51
rs73302078	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	5.20E-13	0.31	0.54
rs4931709	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.43	0.06	2.09E-14	0.32	0.54

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset*	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs4931709	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.48	0.07	2.95E-13	0.35	0.61
rs4931709	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.45	0.06	3.16E-13	0.33	0.57
rs79966895	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.73	0.08	2.97E-22	0.58	0.88
rs79966895	12	AC079630.4	DOM	rs76904798	12		DOM	All	13	12631	0.74	0.08	7.46E-22	0.59	0.89
rs79966895	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.78	0.09	8.26E-19	0.61	0.96
rs79966895	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.79	0.09	2.12E-17	0.61	0.97
rs55788009	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.46	0.06	5.57E-14	0.34	0.59
rs55788009	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.51	0.07	1.89E-13	0.37	0.65
rs55788009	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.48	0.07	3.87E-13	0.35	0.61
rs12425140	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.46	0.07	7.74E-12	0.33	0.59
rs12425140	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.47	0.07	3.27E-11	0.33	0.61
rs12425140	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.50	0.08	3.41E-11	0.35	0.65
rs11835508	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.46	0.07	6.90E-12	0.33	0.59
rs11835508	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.50	0.08	3.24E-11	0.35	0.65
rs11835508	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.47	0.07	3.36E-11	0.33	0.60
rs61927448	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.47	0.06	3.82E-14	0.35	0.59
rs61927448	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.51	0.07	1.62E-13	0.38	0.65
rs61927448	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.48	0.07	3.85E-13	0.35	0.61
rs55892073	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.48	0.06	4.88E-14	0.35	0.60
rs55892073	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.52	0.07	1.25E-13	0.38	0.66
rs55892073	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.49	0.07	2.73E-13	0.36	0.62
rs11832405	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.48	0.06	4.26E-14	0.35	0.60
rs11832405	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.52	0.07	1.02E-13	0.38	0.66
rs11832405	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.49	0.07	1.79E-13	0.36	0.62
rs76534616	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.50	0.07	4.38E-14	0.37	0.63
rs76534616	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.51	0.07	1.23E-13	0.38	0.65
rs76534616	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.55	0.08	3.69E-13	0.40	0.70

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset*	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs17633701	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.51	0.06	8.79E-16	0.39	0.64
rs17633701	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.52	0.07	1.72E-14	0.39	0.66
rs17633701	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.56	0.07	2.54E-14	0.42	0.70
rs11830774	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.51	0.06	2.38E-15	0.39	0.64
rs11830774	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.52	0.07	3.28E-14	0.39	0.66
rs11830774	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.56	0.07	7.50E-14	0.41	0.71
rs1840973	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.52	0.07	4.81E-15	0.39	0.65
rs1840973	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.53	0.07	2.01E-14	0.39	0.66
rs1840973	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.57	0.08	1.76E-13	0.42	0.72
rs61929560	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.52	0.07	1.58E-15	0.39	0.65
rs61929560	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.53	0.07	1.39E-14	0.39	0.66
rs61929560	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.57	0.08	7.90E-14	0.42	0.72
rs1007709	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.59	0.04	2.67E-43	0.50	0.67
rs1007709	12	AC079630.4	DOM	rs76904798	12		DOM	All	13	12631	0.63	0.05	2.34E-41	0.53	0.72
rs1007709	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	0.63	0.06	1.78E-26	0.52	0.75
rs1007709	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	0.67	0.07	6.25E-23	0.54	0.80
rs61929593	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	0.42	0.06	1.32E-13	0.31	0.53
rs138094630	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	9	5452	0.91	0.11	8.34E-17	0.69	1.12
rs117177190	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	9	5452	0.94	0.11	1.22E-17	0.73	1.16
rs17575456	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	9	5452	0.94	0.11	1.22E-17	0.73	1.16
rs118147351	12	AC079630.4	DOM	rs76904798	12		ADD	All	6	3538	1.23	0.16	2.29E-14	0.91	1.54
rs118147351	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	6	2865	1.32	0.18	1.18E-13	0.97	1.67
rs117426805	12	AC079630.4	DOM	rs76904798	12		ADD	All	3	2993	1.58	0.24	3.14E-11	1.12	2.05
rs117178442	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	7	4141	0.88	0.13	4.09E-11	0.62	1.14
rs150084348	12	AC079630.4	DOM	rs76904798	12	RNU6-400P	ADD	NEO	6	3722	0.94	0.14	8.10E-12	0.67	1.21
rs117304585	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	5	2446	1.01	0.16	4.61E-10	0.69	1.33
rs140582511	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	4	2128	1.09	0.17	5.75E-11	0.76	1.41

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset*	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs116909529	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	4	2128	1.10	0.17	3.78E-11	0.77	1.43
rs117629066	12	AC079630.4	DOM	rs76904798	12		ADD	All	3	2993	1.53	0.23	1.49E-11	1.08	1.97
rs141945110	12	AC079630.4	DOM	rs76904798	12	RP11-313F23.4	ADD	NEO	2	1890	1.62	0.25	1.64E-10	1.13	2.12
rs7134491	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	-0.31	0.03	1.30E-19	-0.38	-0.24
rs7134491	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	-0.33	0.04	4.78E-17	-0.41	-0.26
rs7134491	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	-0.38	0.05	2.07E-12	-0.49	-0.27
rs9739724	12	AC079630.4	DOM	rs76904798	12		ADD	All	13	12631	-0.31	0.04	1.69E-18	-0.38	-0.24
rs9739724	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	12	9232	-0.34	0.04	1.11E-15	-0.42	-0.25
rs9739724	12	AC079630.4	DOM	rs76904798	12		DOM	NEO	12	9232	-0.37	0.06	1.25E-10	-0.49	-0.26
rs7296974	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	2	1890	1.60	0.25	1.43E-10	1.11	2.09
rs6488226	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	2	1890	1.60	0.25	1.43E-10	1.11	2.09
rs34207584	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	9	7884	-0.27	0.04	3.49E-12	-0.35	-0.20
rs9705853	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	9	7884	-0.28	0.04	6.53E-13	-0.36	-0.21
rs9705553	12	AC079630.4	DOM	rs76904798	12		ADD	NEO	9	7884	-0.28	0.04	9.13E-13	-0.36	-0.20
rs1995303	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	2	3260	1.55	0.22	2.80E-12	1.12	1.99
rs142523062	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3874	1.81	0.27	2.53E-11	1.28	2.34
rs11052071	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	4	4574	1.29	0.21	3.58E-10	0.89	1.70
rs11052072	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	4	4574	1.29	0.21	3.76E-10	0.89	1.69
rs10844245	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	4	4574	1.30	0.21	3.12E-10	0.89	1.70
rs11052076	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	4	4574	1.29	0.21	3.49E-10	0.89	1.70
rs11052078	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3960	1.74	0.24	2.29E-13	1.27	2.20
rs10844246	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	2	3260	1.85	0.24	3.69E-14	1.37	2.32
rs2102043	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3960	1.74	0.24	2.39E-13	1.27	2.20
rs112530044	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	2	3260	1.84	0.24	3.84E-14	1.37	2.32
rs6488066	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3960	1.69	0.23	5.19E-13	1.23	2.15
rs11052083	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	2	3260	1.84	0.24	3.84E-14	1.37	2.32

SNP with ME				SNP without ME				Meta-Analysis Results							
ID	Chr.	Gene Name	Genetic Model*	ID	Chr.	Gene Name	Genetic Model*	Dataset*	Center Count	Cases Count	Beta	Standard Error	p-Value	CI Lower Bound	CI Upper Bound
rs11052084	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3960	1.74	0.24	2.46E-13	1.27	2.20
rs73100149	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	2	3260	1.84	0.24	4.59E-14	1.36	2.32
rs16920058	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3960	1.72	0.23	1.89E-13	1.26	2.18
rs1909509	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3960	1.72	0.23	1.89E-13	1.26	2.18
rs10844248	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	2	3260	1.84	0.24	3.97E-14	1.37	2.32
rs7313539	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3960	1.69	0.23	5.33E-13	1.23	2.15
rs73081434	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	2	3260	1.84	0.24	3.97E-14	1.37	2.32
rs73081435	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	2	3260	1.84	0.24	3.97E-14	1.37	2.32
rs35861269	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	2	3260	1.85	0.24	3.53E-14	1.37	2.32
rs7961549	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3960	1.69	0.23	5.24E-13	1.23	2.15
rs4931028	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	4	4574	1.29	0.21	4.68E-10	0.88	1.69
rs78346880	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	3	3874	1.85	0.27	5.16E-12	1.32	2.37
rs151094822	12	LRRK2	DOM	rs34637584	12	FGD4	DOM	NEO	4	4574	2.51	0.23	5.81E-28	2.06	2.95
rs80291054	12	LRRK2	DOM	rs34637584	12	PKP2	DOM	NEO	4	4574	1.53	0.21	3.23E-13	1.12	1.94
rs7966490	12	LRRK2	DOM	rs34637584	12	PKP2	DOM	NEO	4	4574	1.41	0.21	2.21E-11	1.00	1.83
rs74568164	12	LRRK2	DOM	rs34637584	12	PKP2	DOM	NEO	4	4574	1.42	0.21	1.79E-11	1.01	1.83
rs75650004	12	LRRK2	DOM	rs34637584	12	PKP2	DOM	NEO	4	4574	1.41	0.21	2.46E-11	1.00	1.82

Supplement 3: Sensitivity analysis of meta-analysis results for SNP pairs with available 2x2 contingency tables of genotype counts.

SNP with ME ID	SNP without ME ID	Main Meta-Analysis GxG Interaction			Simmonds and Higgins Meta-Analysis GxG Interaction			OR BNHM*		
		OR	p-Value	Center Count	Total Cases	OR	p-Value		Center Count	Total Cases
rs76904798	rs187879258	3.29	2.61E-12	9	6738	3.2	7.5E-15	9	6738	3.16
rs76904798	rs117835488	3.37	3.79E-14	8	6158	3.5	1.86E-15	9	6738	3.56
rs76904798	rs139007869	3.79	9.66E-15	7	5354	4.11	2.59E-10	9	6738	4.3
rs76904798	rs1007709	1.8	2.67E-43	13	12631	1.86	5.77E-42	13	12631	1.88
rs76904798	rs141945110	5.07	1.64E-10	2	1890	5.19	9.85E-11	2	1890	5.37
rs26431	rs139186308	5.32	9.75E-11	5	3696			9	5429	5.42
rs112485576	rs7856915	5.86	1.73E-10	6	4759	3.86	4.08E-07	8	6250	3.22

*Binomial-normal hierarchical model for meta-analysis calculation, had same center and cases count as Simmonds and Higgins method.

Chapter 3

3 Discussion

This thesis was largely motivated by the interest in the CO study design for studying G×E and G×G interactions. A further goal was to be able to better understand the origin of complex diseases through these possible interactions. The first publication (i) explored the effect of genotype imputation on the validity and power of statistical tests for G×E interactions in CO studies. The second publication (ii) focused on G×G interactions and their practical implementation in the CO study design using data on IBD. The third and final publication (iii) built and expanded on the findings from the second publication on how to execute G×G interaction analyses in the CO study design for a different complex disease, namely PD. While the key findings of these publications have been discussed in their respective discussion sections, in this chapter they are reflected upon together with some additional combined insights.

3.1 Imputation

Issues regarding imputed data are found in all three publications. While imputation bias was in focus in publication (i), issues regarding imputed data were also present in the G×G interaction analyses as imputed data was used. In the first publication (i) regarding imputed data in a CO study design, we expected discrepancies between the true and imputed genotypes due to the fact that the reference base used for imputation consisted of healthy controls while our dataset of cases included only patients with a certain disease phenotype. Similarly, population stratification can cause difficulties in the precision of the imputation due to different ethnical backgrounds of the reference base and data imputed. The large size and high quality of the reference base would still pose problems if, say the dataset for imputation consisted of African ancestry individuals and the reference base of only Caucasian ancestry. Since we can only expect the prevalence rate of the disease in question to be present in a reference base, this can make the imputation of rare disease cases problematic. Indeed, our results from publication (i) showed that the MAF of the imputed dataset's SNPs of interest in areas with known main resembled the MAF of the reference dataset as the SNPs of interest maximum LD to the nearest SNP decreased. This discrepancy later introduced a bias in the analysis of the G×E interaction simulation that could be shown by the decreasing statistical power as the LD to the nearest SNP decreased. Furthermore, in certain scenarios, for example, given very low MAFs of less than 0.05, the imputation accuracy score for SNPs of interest with ME is still high, even if the agreement between the true and imputed genotypes (Cohen's

kappa) is low. This calls for caution when imputing SNPs with small MAFs in areas with known MEs. Furthermore, when applying the CO study design, controls are not available, and it is not possible to estimate MEs without controls. However, it is possible to identify SNPs in gene regions known to have MEs with the help of publicly available databases such as dbSNP (Sherry 2001).

By no means do the conclusions from publication (i) suggest that G×E and or G×G interactions should exclude imputed SNPs and/or SNPs with low MAF. No false positive results could be detected, but a loss in statistical power in certain cases, meaning that some interaction effects remain hidden. Large datasets comprising of different centers where different genotyping chips were used depend on imputation to make the data comparable and available for analysis, as seen in publications (ii) and (iii). Since both publications either fully (publication (ii)) or partially (publication (iii)) focused on SNPs with MEs and included imputed data, it is probable that imputation bias was introduced. While the effect of genotype imputation on the validity and power of statistical tests for G×G interactions in CO studies remain to be analysed, similar conclusions to the G×E interaction study are probable, meaning loss in statistical power when examining the interaction effect. This may be one explanation for the lack of significant statistical interactions in the CO analysis of G×G interaction in IBD. In publication (iii), over five different genotyping chips were used for 16 different centers. It is obvious that imputation bias is a small price to pay for making all available datasets comparable. However, the results showed that in some cases the number of observations left after meta-analysis were reduced to 10 000 or even 4 000 total observations, only 27.5 and 11 percent, respectively, from the total available data. This was partially due to the fact that observations, where the hard-call threshold value of the probability of a certain genotype after imputation of 0.8 was implemented, were excluded. This finding also indirectly shows that imputation can be problematic in areas with MEs.

3.2 Rare SNPs

A common challenge in all three publications were rare SNPs, i.e. SNPs with very low MAF (less than 0.05). In the first publication (i), as previously mentioned, the imputation accuracy score for SNPs of interest with MEs was high, even if the imputed MAFs differed greatly from the CO sample's MAF. I hoped to simulate G×E interaction effects of a certain magnitude in order to achieve 80% statistical power and then see how the power changes based on the imputed data of varying maximum LD. This, however, was difficult with rare SNPs and my sample size of 719 cases. The mean statistical power for SNPs with known main effects and a MAF of less than 0.05 was less than 25%. This is retraceable to the fact that if, given, say a MAF of 0.01 and a sample size of 719, the number of minor allele carriers would equal to 7. In a simulation, regardless of the probability of exposure for the minor allele carriers, this low number of minor allele carriers make it difficult to simulate constant odds ratios with high statistical power.

Issues with rare SNPs also occurred in publications (ii) and (iii) analysing G×G interactions. While the overall sample size was much higher than the sample size in publication (i), due to the heterogeneity of the dataset, a center-wise analysis followed by meta-analysis over all results was needed. Thus, smaller center sizes posed a problem when analysing rare SNPs.

Moreover, the aspect of G×G interactions also meant that pairs where both SNPs were rare were difficult to consider. For example, given a large center of, say, 1 000 cases, and two rare SNPs with MAFs of 0.05, the number of minor allele carriers for either SNP would equal to 50. In a contingency table, the distribution of the allele carriers could easily lead to low/zero values in one of the cells. These rare occurrences lead to an unbalanced design with unequal group sizes that further leads to low statistical power in general. Moreover, in a logistic regression, this leads to convergence problems. The smaller the center, the more frequent this scenario becomes and the issue more relevant.

These issues with rare SNPs lead to the exclusion of center-wise results or the interaction pair altogether. This, of course, does not mean that there were no interactions, but rather, that it was not possible to identify and estimate their effect in our analysis with the data obtained. As many papers discuss the fact that the missing heritability may lie in very rare SNPs (Manolio et al. 2009; Génin 2020), even if rare SNPs had to be excluded from the analysis, one must keep in mind that true interactions could still be present.

3.3 An Approach for Investigating Gene-Gene Interactions

A large portion of this thesis focused on G×G interaction analysis and the challenges that need to be overcome in order to be able to execute it. Contrary to G×E interactions, G×G are less straight forward due to the number of possible interaction pairs and linkage disequilibrium issues. The main challenges in the analysis of G×G interactions discussed in this thesis include statistical power and multiple testing, fulfilment of the CO design assumptions, computation and interpretation of results.

Before examining G×G interactions, it was already obvious that the number of overall possible interaction pairs would lead to issues in the statistical analysis. Therefore, the CO study design was examined and chosen due the higher statistical power given the same number of cases in comparison to the standard CC study design. Furthermore, we examined possibilities to reduce the number of possible interaction pairs by focusing on SNPs with known main effects due to the assumption that these SNPs are more likely to exhibit G×G interactions in the first place. In the two datasets that I examined, IBD and PD, no statistically significant G×G interaction pairs were found in pairs where both SNPs exhibited proven main effects. However, if the search was expanded by limiting only one SNP in the pair to have a known main effect and the other genome-wide, some interesting, statistically significant pairs were found.

Another closely related issue is the family-wise error rate. This also has to do with the vast number of possible interaction pairs when examining G×G interactions. This high number results from the quadratic relationship between the number of SNPs included in the study and the accruing number of interaction tests. Thus, Bonferroni correction seems overly conservative at first sight. I examined other options of controlling the family-wise error rate, for example, by simulating the null hypothesis. However, once spurious pair-wise associations between SNPs were excluded, the family-wise error rate could not be controlled in any less conservative way than with Bonferroni correction. Thus, the solution to first concentrate on

SNP pairs where at least one SNP in the pair has a known main effect is effective in order to reduce the number of possibly redundant statistical tests performed. The genome-wide significance rate of 5×10^{-8} , Bonferroni corrected for the number of genome-wide searches (equal to the number of SNPs with known main effects) was found to be adequate for this approach of analysing G×G interactions.

The gain in statistical power from CO study design comes at the price of having to fulfil the two assumptions. The first one, stating that the disease of interest has to be sufficiently rare in the general population is unproblematic in both the G×G and G×E interaction studies. The prevalence rates of most diseases can be easily researched and identified. The second assumption, namely that the two risk factors under study are uncorrelated in the general population, is less straight forward. Generally, in G×E interactions this is rarely the case and can be analytically determined. Thus, it was not an issue in publication (i), which dealt with a simulation of the G×E interaction. However, this assumption is more problematic in the analysis of G×G interactions, considered in publications (ii) and (iii). Issues such as LD, population stratification and cryptic relatedness pose difficulties in the analysis of G×G interactions as the SNP pairs exhibiting these issues and the datasets analysed in this thesis were no different.

Thus, I examined different ways to determine whether a SNP pair fulfils the independence assumption, which included assuming that either the controls within the study or from a publicly available dataset represent the general population and tested for independence in the controls. The former method had too many drawbacks, because it inferred that any CO study would also need to have adequate controls, undermining the whole concept of a CO study design. Meanwhile, the results using a publicly available dataset were not satisfying, as they produced false positive results. This could be traced back to the fact that a publicly available dataset, while being from the same ethnic group, may still exhibit population stratification between cases and controls. Furthermore, batch effects, cryptic relatedness and other dataset specific artefacts will not be present in the publicly available dataset. Therefore, a simple and effective method of making sure the SNPs in the potential interaction pair fulfil the independence assumption is to only test pairs of SNPs that are on different chromosome arms. This ensures that there is no linkage disequilibrium between SNPs resulting into false positive associations. Analysing each center separately, including principal components on the center level and merging the results via meta-analysis takes care of systematically varying allele frequencies within and among the different centers.

A further issue that was addressed in this thesis is the computational difficulties associated with the many possible SNP combination pairs in G×G interaction analysis. In publication (ii), the computational burden was reduced by focusing only on SNPs with known main effects. However, for publication (iii), another approach was taken as genome-wide interactions were also the focus of this study. First, the number of possible tests was reduced by focusing the genome-wide search on pairs where at least one of the SNPs in the interaction pair had a known main effect. Second, a two-step screening process for the meta-analysis was performed, focusing on interaction pairs with the lowest p-values. In the first screening step, a meta-analysis was conducted on only those centers and SNP pairs for which the specific SNP pair's center level p-value was less than 0.05. All SNP pairs that had a combined p-value from

the meta-analysis of less than 5×10^{-5} were selected for the second step. This vastly reduced the number of potential interaction pairs. In the second step, a meta-analysis was executed on all data from all centers for the selected SNP pairs. Given such a high significance threshold in the first screening step (0.05) and low significance threshold level in the final analysis (genome-wide 5×10^{-8} , in publication (iii)'s case of 90 genome-wide searches 5.56×10^{-10}), the screening process should have made false negative results only slightly more likely and could not have produced any false positive results.

The final aspect of G×G interaction analyses considered in this thesis is the interpretation of the statistically significant results. Unlike G×E interactions, where further experiments in laboratories could be designed with exposed and unexposed scenarios, G×G interactions are not as straight forward. It is possible that many statistical interactions are not biological. For example, in publication (iii), some statistically significant interactions included SNPs in antisense genes. Databases such as Ensembl (Yates et al. 2019) could be used to determine the biological effects of the SNPs under study. Thus, it is always useful to consult experts from medicine and biology when interpreting the results.

To sum up, my proposed method for the analysis of G×G interactions is to apply the CO study design using a logistic regression and primarily focus on SNPs with main effects, meaning that at least one SNP in the potential interaction pair should be one with a known main effect. Furthermore, only SNP pairs on different chromosome arms should be paired in order to fulfil the independence assumption of the CO study design. If more than one center is present in the dataset, the logistic regression should be carried out on the center level, always including ten principal components and the results should be summarised by means of a meta-analysis. A screening process by first meta-analysing only those SNP pairs with nominally significant p-values and then gathering all data from all possible centers of the screened and meta-analysed pairs with a very low p-values (e.g. 5×10^{-5}) can be done in order to reduce the computational and data transfer burden. Bonferroni correction adequately controls the family-wise error rate, thus the genome-wide significance level of 5×10^{-8} should be adjusted by the number of times the whole genome was tested, i.e., the number of SNPs with main gene effects. Finally, when interpreting the results, pathway databases and experts in the field of the disease under study should be consulted in order to make a sound conclusion of the G×G interaction.

3.4 Strengths and Limitations

In the first part of this thesis, publication (i), the advantages of a simulation and a real dataset were combined by simulating the exposure status for G×E interactions while using genotypes from real data. Not only could a realistic LD structure be obtained by using a real dataset, but the dataset included diagnosed CD patients, thus, the areas with potential more imprecise areas with ME were also realistic. It was a convenient and inexpensive way to answer our research questions. Moreover, simulations allow to systematically vary parameters, such as LD or exposure probabilities, thus I could thoroughly examine many scenarios. Access to whole genome data allowed to select SNPs with varying MAFs and ME OR sizes in order to replicate realistic scenarios. Since 10 000 replicates of the simulation for publication (i) were performed, it would be safe to say that our estimates were precise.

In all three publications the CO study design was implemented. On the one hand, one of the goals of this thesis was to examine the CO study design under different conditions. On the other hand, it must be pointed out that for the purpose of investigating interactions, G×E and G×G alike, the CO study design is a very statistically powerful approach. Thus, the pure choice of the specific design is a strength in all studies performed. However, a few aspects must be kept in mind when considering the CO study design. Due to the nature of the CO study design, it is not possible to calculate main effects. Thus, it is only highly useful if the focus of the study are interactions, not main effects. Combined with the CO study design, the very large sample sizes under investigation in publications (ii) and (iii) add substantial weight to the significance of their results. For both diseases, IBD and PD, the currently largest available datasets were used. As the sample size increases, so does the statistical power. Therefore, by applying the CO study design to the largest available datasets, a very sound and statistically powerful methodology was implemented.

However, while the sample sizes were very beneficial in the studies on G×G interactions, a much smaller sample size of 719 was used in the simulation study (publication (i)). Due to the smaller sample size, it was only possible to achieve 80% statistical power with more common SNPs. For rarer SNPs, even though the exposure probabilities were high, the statistical power was low. Unfortunately, it is a very likely scenario that a rare SNP may exhibit a true G×E interaction effect. As with all simulations, they capture very specific scenarios, but the reality is often very much more complex. In some of the simulation scenarios, relatively large G×E interaction effects had to be simulated (e.g. OR = 6.80). For some rare SNPs, the exposure probabilities had to be set to higher than 80%. Such scenarios are not extremely likely in reality. What is more, other setting could also be fixed and not necessarily reflect reality, for example, only the dominant genetic model was examined in publications (i) and (ii).

Since the independence assumption needs to be fulfilled for the CO study design to be valid, the approach used was to exclude SNP pairs on the same chromosome arm in order to fulfil this assumption. Hence, it was not possible to examine all possible G×G interaction pairs. However, it would be plausible that biological, functional interactions occur on SNPs in closer proximity to each other, e.g. on the same chromosome arm. Even though the sample sizes in the G×G interaction studies were vast, very rare SNPs, especially combinations of two rare SNPs would lead to the logistic regression not converging. This also meant that some rare potential interaction pairs had to be disregarded. Finally, the technical limitations pressed me to implement the previously mentioned screening procedure. While the applied screening method does not produce any false positive results, it could exclude some true positive G×G interactions.

3.5 Conclusion and Outlook

One of the two primary goals of this thesis was to examine complex diseases more in depth from the aspect of interactions. Biological and statistical interaction are not on a two-way street. While some may question why study statistical interaction at all, if there are no underlying biological or functional interactions, it is important to note that SNPs interacting with other SNPs or environmental factors can serve as proxies. Moreover, assessing the true

magnitude of G×G interaction is also an important step in the direction of explaining some of the “missing heritability” of these complex diseases. (Maher 2008) Past estimates of total heritability probably have been significantly inflated by unaccounted G×G interactions (Zuk et al. 2012). Understanding the genetic variation and statistical interactions may contribute to better prevention, diagnosis and treatment of complex diseases. For example, based on our findings from publication (iii) on significant G×G interaction pairs, the presence of the specific genotypes might be used as an advantage to identify PD patients more quickly and effectively.

The second primary goal of this thesis was to explore the utility of the CO study design. As shown in publications (i) and (ii), the CO study design is viable and practically applicable when examining G×E and G×G interactions, if certain precautions are taken and aspects considered. When analysing imputed data and considering G×E interactions, caution is advised with rare SNPs (MAF < 0.05), because in areas of ME, the imputation quality score may indicate accurate imputation, yet the true MAF may differ greatly. This further results in bias when examining G×E interactions. Moreover, in imputed areas with MEs and lower LD to the nearest not imputed SNPs, a bias in the G×E interaction may also be introduced. As defined in publication (ii) and further discussed in Section 3.3., the CO study design is slightly less straight forward when analysing G×G interactions. Yet following the approach developed in this thesis, it is safe to state that the CO study design is a viable, statistically powerful and highly recommended method to analyse G×G interactions.

While some aspects of the CO study design were thoroughly discussed in this thesis, a few issues were not covered and doors opened for further research questions. After the genome-wide analysis of G×G interactions and significant findings in PD, it is a plausible interest to return to the IBD dataset and perform a genome-wide G×G interaction analysis following the same screening approach as used in the PD analysis. Furthermore, my thesis focused on a binary definition of environmental exposure, quantitative variation in phenotypes are possible and probable scenarios that could be further examined in the CO study design. In publication (iii), subsets were defined by early and normal age at onset of PD and some G×G interactions were significant in only one of the subsets. As studies have shown some genetic variability when it comes to age at onset of PD (Blauwendraat et al. 2019), it naturally comes to mind that there could be potential to further consider G×G interactions in smaller subgroups by age at onset or look into G×E interactions while considering age at onset as a continuous variable.

Two further areas for potential research in the future could be concluded from this thesis. First, the possibility of multi-way interactions. However, since even genome-wide G×G interactions are computationally difficult to master at the moment, either the CO study design could be implemented with other screening methods or more efficient computation should be employed. Second, the missing heritability could lie in the rare and very rare SNPs and/or their interactions. In order to further examine this area, either even larger datasets on specific diseases should be gathered, or new methodologies developed in order to decipher the effect of (very) rare SNPs.

In conclusion, this thesis addressed the practicability of implementing the statistically powerful CO study design in G×G interactions and G×E interactions using imputed data. A thorough, yet focused analysis of G×G interaction makes it possible to provide deeper insights into the genetic architecture of complex diseases. The work in this thesis implies, that in

further G×G and G×E interaction research the CO design may safely be adopted. It also showed that datasets with large numbers of cases are needed to investigate interactions, even if the statistically powerful CO study design is used. Thus, it is my hope that in the future, the CO study design will become common practice when investigating interactions and datasets on various diseases will be combined and available for extensive interaction analyses in order to further understand the etiology of complex human diseases.

Bibliography

- Alatab, Sudabeh, Sadaf G Sepanlou, Kevin Ikuta, Homayoon Vahedi, Catherine Bisignano, Saeid Safiri, Anahita Sadeghi, et al. 2020. "The Global, Regional, and National Burden of Inflammatory Bowel Disease in 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017." *The Lancet Gastroenterology & Hepatology* 5 (1): 17–30. [https://doi.org/10.1016/S2468-1253\(19\)30333-4](https://doi.org/10.1016/S2468-1253(19)30333-4).
- Albert, P. S. 2001. "Limitations of the Case-Only Design for Identifying Gene-Environment Interactions." *American Journal of Epidemiology* 154 (8): 687–93. <https://doi.org/10.1093/aje/154.8.687>.
- Aleknonytė-Resch, Milda, Sandra Freitag-Wolf, The International Inflammatory Bowel Disease Genetics Consortium, Stefan Schreiber, Michael Krawczak, and Astrid Dempfle. 2020. "Case-Only Analysis of Gene–Gene Interactions in Inflammatory Bowel Disease." *Scandinavian Journal of Gastroenterology*, July, 1–10. <https://doi.org/10.1080/00365521.2020.1790646>.
- Amre, Devendra K, Savio D'Souza, Kenneth Morgan, Gillian Seidman, Philippe Lambrette, Guy Grimard, David Israel, et al. 2007. "Imbalances in Dietary Consumption of Fatty Acids, Vegetables, and Fruits Are Associated With Risk for Crohn's Disease in Children." *The American Journal of Gastroenterology* 102 (9): 2016–25. <https://doi.org/10.1111/j.1572-0241.2007.01411.x>.
- Anderson, Carl A., Fredrik H. Pettersson, Jeffrey C. Barrett, Joanna J. Zhuang, Jiannis Ragoussis, Lon R. Cardon, and Andrew P. Morris. 2008. "Evaluating the Effects of Imputation on the Power, Coverage, and Cost Efficiency of Genome-Wide SNP Platforms." *The American Journal of Human Genetics* 83 (1): 112–19. <https://doi.org/10.1016/j.ajhg.2008.06.008>.
- Antonarakis, Stylianos E., and Jacques S. Beckmann. 2006. "Mendelian Disorders Deserve More Attention." *Nature Reviews Genetics* 7 (4): 277–82. <https://doi.org/10.1038/nrg1826>.
- Aschard, Hugues, Donna Spiegelman, Vincent Laville, Pete Kraft, and Molin Wang. 2018. "A Test for Gene-Environment Interaction in the Presence of Measurement Error in the Environmental Variable." *Genetic Epidemiology* 42 (3): 250–64. <https://doi.org/10.1002/gepi.22113>.
- Bjorklund, Geir, Vera Stejskal, Mauricio A. Urbina, Maryam Dadar, Salvatore Chirumbolo, and Joachim Mutter. 2018. "Metals and Parkinson's Disease: Mechanisms and Biochemical Processes." *Current Medicinal Chemistry* 25 (19): 2198–2214. <https://doi.org/10.2174/0929867325666171129124616>.
- Blauwendraat, Cornelis, Karl Heilbron, Costanza L. Vallerga, Sara Bandres-Ciga, Rainer von Coelln, Lasse Pihlstrøm, Javier Simón-Sánchez, et al. 2019. "Parkinson's Disease Age at Onset Genome-wide Association Study: Defining Heritability, Genetic Loci, and A-synuclein Mechanisms." *Movement Disorders* 34 (6): 866–75. <https://doi.org/10.1002/mds.27659>.
- Boone, Charles, Howard Bussey, and Brenda J. Andrews. 2007. "Exploring Genetic Interactions and Networks with Yeast." *Nature Reviews Genetics* 8 (6): 437–49. <https://doi.org/10.1038/nrg2085>.
- Browning, Brian L., Ying Zhou, and Sharon R. Browning. 2018. "A One-Penny Imputed Genome from Next-Generation Reference Panels." *The American Journal of Human Genetics* 103 (3): 338–48. <https://doi.org/10.1016/j.ajhg.2018.07.015>.
- Buniello, Annalisa, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." *Nucleic Acids Research* 47 (D1): D1005–12. <https://doi.org/10.1093/nar/gky1120>.
- Bush, William S., and Jason H. Moore. 2012. "Chapter 11: Genome-Wide Association Studies." Edited by Fran Lewitter and Maricel Kann. *PLoS Computational Biology* 8 (12): e1002822. <https://doi.org/10.1371/journal.pcbi.1002822>.
- CAAPA, Rasika Ann Mathias, Margaret A. Taub, Christopher R. Gignoux, Wenqing Fu, Shaila Musharoff, Timothy D. O'Connor, et al. 2016. "A Continuum of Admixture in the Western

- Hemisphere Revealed by the African Diaspora Genome." *Nature Communications* 7 (1): 12522. <https://doi.org/10.1038/ncomms12522>.
- Cao, P., X. Yang, and T. C. Sudhof. 2013. "Complexin Activates Exocytosis of Distinct Secretory Vesicles Controlled by Different Synaptotagmins." *Journal of Neuroscience* 33 (4): 1714–27. <https://doi.org/10.1523/JNEUROSCI.4087-12.2013>.
- Capriotti, Teri, and Kristina Terzakis. 2016. "Parkinson Disease:" *Home Healthcare Now* 34 (6): 300–307. <https://doi.org/10.1097/NHH.0000000000000398>.
- Cardon, Lon R, and Lyle J Palmer. 2003. "Population Stratification and Spurious Allelic Association." *The Lancet* 361 (9357): 598–604. [https://doi.org/10.1016/S0140-6736\(03\)12520-2](https://doi.org/10.1016/S0140-6736(03)12520-2).
- Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (1): 7. <https://doi.org/10.1186/s13742-015-0047-8>.
- Chen, G.-B., S. H. Lee, M.-J. A. Brion, G. W. Montgomery, N. R. Wray, G. L. Radford-Smith, P. M. Visscher, and the International IBD Genetics Consortium. 2014. "Estimation and Partitioning of (Co)Heritability of Inflammatory Bowel Disease from GWAS and Immunochip Data." *Human Molecular Genetics* 23 (17): 4710–20. <https://doi.org/10.1093/hmg/ddu174>.
- Chen, Junfang, Dietmar Lippold, Josef Frank, William Rayner, Andreas Meyer-Lindenberg, and Emanuel Schwarz. 2019. "Gimpute: An Efficient Genetic Data Imputation Pipeline." Edited by Oliver Stegle. *Bioinformatics* 35 (8): 1433–35. <https://doi.org/10.1093/bioinformatics/bty814>.
- Cheng, K.F., and W.J. Lin. 2009. "The Effects of Misclassification in Studies of Gene-Environment Interactions." *Human Heredity* 67 (2): 77–87. <https://doi.org/10.1159/000179556>.
- Clark, Ira E., Mark W. Dodson, Changan Jiang, Joseph H. Cao, Jun R. Huh, Jae Hong Seol, Soon Ji Yoo, Bruce A. Hay, and Ming Guo. 2006. "Drosophila Pink1 Is Required for Mitochondrial Function and Interacts Genetically with Parkin." *Nature* 441 (7097): 1162–66. <https://doi.org/10.1038/nature04779>.
- Clarke, Geraldine M, Fredrik H Pettersson, and Andrew P Morris. 2009. "A Comparison of Case-Only Designs for Detecting Gene × Gene Interaction in Rheumatoid Arthritis Using Genome-Wide Case-Control Data in Genetic Analysis Workshop 16." *BMC Proceedings* 3 (S7): S73. <https://doi.org/10.1186/1753-6561-3-S7-S73>.
- Clayton, David, and Paul M McKeigue. 2001. "Epidemiological Methods for Studying Genes and Environmental Factors in Complex Diseases." *The Lancet* 358 (9290): 1356–60. [https://doi.org/10.1016/S0140-6736\(01\)06418-2](https://doi.org/10.1016/S0140-6736(01)06418-2).
- Cleynen, Isabelle, Emilie Vazeille, Marta Artieda, Hein W Verspaget, Magdalena Szczypiorska, Marie-Agnès Bringer, Peter L Lakatos, et al. 2014. "Genetic and Microbial Factors Modulating the Ubiquitin Proteasome System in Inflammatory Bowel Disease." *Gut* 63 (8): 1265–74. <https://doi.org/10.1136/gutjnl-2012-303205>.
- Cordell, H. J. 2002. "Epistasis: What It Means, What It Doesn't Mean, and Statistical Methods to Detect It in Humans." *Human Molecular Genetics* 11 (20): 2463–68. <https://doi.org/10.1093/hmg/11.20.2463>.
- Cordell, Heather J. 2009. "Detecting Gene–Gene Interactions That Underlie Human Diseases." *Nature Reviews Genetics* 10 (6): 392–404. <https://doi.org/10.1038/nrg2579>.
- Cowman, Tyler, and Mehmet Koyutürk. 2017. "Prioritizing Tests of Epistasis through Hierarchical Representation of Genomic Redundancies." *Nucleic Acids Research* 45 (14): e131–e131. <https://doi.org/10.1093/nar/gkx505>.
- Das, Sayantan, Gonçalo R. Abecasis, and Brian L. Browning. 2018. "Genotype Imputation from Large Reference Panels." *Annual Review of Genomics and Human Genetics* 19 (1): 73–96. <https://doi.org/10.1146/annurev-genom-083117-021602>.
- Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, et al. 2016. "Next-Generation Genotype Imputation Service and Methods." *Nature Genetics* 48 (10): 1284–87. <https://doi.org/10.1038/ng.3656>.
- Dempfle, Astrid, André Scherag, Rebecca Hein, Lars Beckmann, Jenny Chang-Claude, and Helmut Schäfer. 2008. "Gene–Environment Interactions for Complex Traits: Definitions,

- Methodological Requirements and Challenges.” *European Journal of Human Genetics* 16 (10): 1164–72. <https://doi.org/10.1038/ejhg.2008.106>.
- Evangelou, Evangelos, Thomas A. Trikalinos, Georgia Salanti, and John P. A. Ioannidis. 2006. “Family-Based versus Unrelated Case-Control Designs for Genetic Associations.” *PLoS Genetics* 2 (8): e123. <https://doi.org/10.1371/journal.pgen.0020123>.
- Fernández-Santiago, Rubén, Núria Martín-Flores, Francesca Antonelli, Catalina Cerquera, Verónica Moreno, Sara Bandres-Ciga, Elisabetta Manduchi, et al. 2019. “SNCA and MTOR Pathway Single Nucleotide Polymorphisms Interact to Modulate the Age at Onset of Parkinson’s Disease.” *Movement Disorders* 34 (9): 1333–44. <https://doi.org/10.1002/mds.27770>.
- Fernández-Torres, Javier, Gabriela Angélica Martínez-Nava, Yessica Zamudio-Cuevas, Carlos Lozada, Daniela Garrido-Rodríguez, and Karina Martínez-Flores. 2020. “Epistasis of Polymorphisms Related to the Articular Cartilage Extracellular Matrix in Knee Osteoarthritis: Analysis-Based Multifactor Dimensionality Reduction.” *Genetics and Molecular Biology* 43 (2): e20180349. <https://doi.org/10.1590/1678-4685-gmb-2018-0349>.
- García-Closas, M., N. Rothman, and J. Lubin. 1999. “Misclassification in Case-Control Studies of Gene-Environment Interactions: Assessment of Bias and Sample Size.” *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 8 (12): 1043–50.
- Gardiner, Sharon J., and Evan J. Begg. 2005. “Pharmacogenetic Testing for Drug Metabolizing Enzymes: Is It Happening in Practice?” *Pharmacogenetics and Genomics* 15 (5): 365–69. <https://doi.org/10.1097/01213011-200505000-00013>.
- Gauderman, W. J. 2002. “Sample Size Requirements for Association Studies of Gene-Gene Interaction.” *American Journal of Epidemiology* 155 (5): 478–84. <https://doi.org/10.1093/aje/155.5.478>.
- Gauderman, W. James. 2002. “Sample Size Requirements for Matched Case-Control Studies of Gene-Environment Interaction.” *Statistics in Medicine* 21 (1): 35–50. <https://doi.org/10.1002/sim.973>.
- Génin, Emmanuelle. 2020. “Missing Heritability of Complex Diseases: Case Solved?” *Human Genetics* 139 (1): 103–13. <https://doi.org/10.1007/s00439-019-02034-4>.
- Glas, Jürgen, Julia Seiderer, Johanna Wagner, Torsten Olszak, Christoph Fries, Cornelia Tillack, Matthias Friedrich, et al. 2012. “Analysis of IL12B Gene Variants in Inflammatory Bowel Disease.” Edited by Giovambattista Pani. *PLoS ONE* 7 (3): e34349. <https://doi.org/10.1371/journal.pone.0034349>.
- Glas, Jürgen, Johanna Wagner, Julia Seiderer, Torsten Olszak, Martin Wetzke, Florian Beigel, Cornelia Tillack, et al. 2012. “PTPN2 Gene Variants Are Associated with Susceptibility to Both Crohn’s Disease and Ulcerative Colitis Supporting a Common Genetic Disease Background.” Edited by Dominik Hartl. *PLoS ONE* 7 (3): e33682. <https://doi.org/10.1371/journal.pone.0033682>.
- Goldman, David, Gabor Orozsi, and Francesca Ducci. 2005. “The Genetics of Addictions: Uncovering the Genes.” *Nature Reviews Genetics* 6 (7): 521–32. <https://doi.org/10.1038/nrg1635>.
- Guan, Weihua, Michael Boehnke, Anna Pluzhnikov, Nancy J. Cox, and Laura J. Scott. 2012. “Identifying Plausible Genetic Models Based on Association and Linkage Results: Application to Type 2 Diabetes: Identifying Genetic Models Based on Association and Linkage Results.” *Genetic Epidemiology*, August, n/a-n/a. <https://doi.org/10.1002/gepi.21668>.
- Günhan, Burak Kürsad, Christian Röver, and Tim Friede. 2020. “Random-effects Meta-analysis of Few Studies Involving Rare Events.” *Research Synthesis Methods* 11 (1): 74–90. <https://doi.org/10.1002/jrsm.1370>.
- Hardy, John, and Andrew Singleton. 2009. “Genomewide Association Studies and Human Disease.” *New England Journal of Medicine* 360 (17): 1759–68. <https://doi.org/10.1056/NEJMra0808700>.
- Howie, Bryan, Jonathan Marchini, and Matthew Stephens. 2011. “Genotype Imputation with Thousands of Genomes.” *G3 & Genes/Genomes/Genetics* 1 (6): 457–70. <https://doi.org/10.1534/g3.111.001198>.

- International Parkinson's Disease Genomics Consortium (IPDGC), Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI), 23andMe, GenePD, NeuroGenetics Research Consortium (NGRC), Hussman Institute of Human Genomics (HIHG), The Ashkenazi Jewish Dataset Investigator, et al. 2014. "Large-Scale Meta-Analysis of Genome-Wide Association Data Identifies Six New Risk Loci for Parkinson's Disease." *Nature Genetics* 46 (9): 989–93. <https://doi.org/10.1038/ng.3043>.
- Ioannidis, John P.A., Evangelia E. Ntzani, Thomas A. Trikalinos, and Despina G. Contopoulos-Ioannidis. 2001. "Replication Validity of Genetic Association Studies." *Nature Genetics* 29 (3): 306–9. <https://doi.org/10.1038/ng749>.
- Jackson, Dan, Martin Law, Theo Stijnen, Wolfgang Viechtbauer, and Ian R. White. 2018. "A Comparison of Seven Random-Effects Models for Meta-Analyses That Estimate the Summary Odds Ratio." *Statistics in Medicine* 37 (7): 1059–85. <https://doi.org/10.1002/sim.7588>.
- Kalia, Lorraine V, and Anthony E Lang. 2015. "Parkinson's Disease." *The Lancet* 386 (9996): 896–912. [https://doi.org/10.1016/S0140-6736\(14\)61393-3](https://doi.org/10.1016/S0140-6736(14)61393-3).
- Khoury, M. J., and W. D. Flanders. 1996. "Nontraditional Epidemiologic Approaches in the Analysis of Gene Environment Interaction: Case-Control Studies with No Controls!" *American Journal of Epidemiology* 144 (3): 207–13. <https://doi.org/10.1093/oxfordjournals.aje.a008915>.
- Kotagal, Vikas, Roger L. Albin, Martijn L. T. M. Müller, Robert A. Koeppe, Kirk A. Frey, and Nicolaas I. Bohnen. 2014. "Modifiable Cardiovascular Risk Factors and Axial Motor Impairments in Parkinson Disease." *Neurology* 82 (17): 1514–20. <https://doi.org/10.1212/WNL.0000000000000356>.
- Kraft, Peter, Yu-Chun Yen, Daniel O. Stram, John Morrison, and W. James Gauderman. 2007. "Exploiting Gene-Environment Interaction to Detect Genetic Associations." *Human Heredity* 63 (2): 111–19. <https://doi.org/10.1159/000099183>.
- Krawczak, Michael, Susanna Nikolaus, Huberta von Eberstein, Peter J.P. Croucher, Nour Eddine El Mokhtari, and Stefan Schreiber. 2006. "PopGen: Population-Based Recruitment of Patients and Controls for the Analysis of Complex Genotype-Phenotype Relationships." *Public Health Genomics* 9 (1): 55–61. <https://doi.org/10.1159/000090694>.
- Kulle, Bettina, Markus Schirmer, Mohammad R. Toliat, Anita Suk, Christian Becker, Mladen Vassilev Tzvetkov, Jürgen Brockmöller, et al. 2005. "Application of Genomewide SNP Arrays for Detection of Simulated Susceptibility Loci." *Human Mutation* 25 (6): 557–65. <https://doi.org/10.1002/humu.20174>.
- Lau, Lonneke ML de, and Monique MB Breteler. 2006. "Epidemiology of Parkinson's Disease." *The Lancet Neurology* 5 (6): 525–35. [https://doi.org/10.1016/S1474-4422\(06\)70471-9](https://doi.org/10.1016/S1474-4422(06)70471-9).
- Li, D., and D. V. Conti. 2008. "Detecting Gene-Environment Interactions Using a Combined Case-Only and Case-Control Approach." *American Journal of Epidemiology* 169 (4): 497–504. <https://doi.org/10.1093/aje/kwn339>.
- Li, Yun, Cristen J. Willer, Jun Ding, Paul Scheet, and Gonçalo R. Abecasis. 2010. "MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes." *Genetic Epidemiology* 34 (8): 816–34. <https://doi.org/10.1002/gepi.20533>.
- Liou, H. H., M. C. Tsai, C. J. Chen, J. S. Jeng, Y. C. Chang, S. Y. Chen, and R. C. Chen. 1997. "Environmental Risk Factors and Parkinson's Disease: A Case-Control Study in Taiwan." *Neurology* 48 (6): 1583–88. <https://doi.org/10.1212/WNL.48.6.1583>.
- Liu, Jimmy Z, Suzanne van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, et al. 2015. "Association Analyses Identify 38 Susceptibility Loci for Inflammatory Bowel Disease and Highlight Shared Genetic Risk across Populations." *Nature Genetics* 47 (July): 979.
- Mackay, Trudy F. C. 2014. "Epistasis and Quantitative Traits: Using Model Organisms to Study Gene–Gene Interactions." *Nature Reviews Genetics* 15 (1): 22–33. <https://doi.org/10.1038/nrg3627>.
- Maher, Brendan. 2008. "Personal Genomes: The Case of the Missing Heritability." *Nature* 456 (7218): 18–21. <https://doi.org/10.1038/456018a>.

- Mann, C J. 2003. "Observational Research Methods. Research Design II: Cohort, Cross Sectional, and Case-Control Studies." *Emergency Medicine Journal* 20 (1): 54–60. <https://doi.org/10.1136/emj.20.1.54>.
- Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–53. <https://doi.org/10.1038/nature08494>.
- Manta-Vogli, Penelope D., and Kleopatra H. Schulpis. 2018. "Phenylketonuria Dietary Management and an Emerging Development." *Journal of the Academy of Nutrition and Dietetics* 118 (8): 1361–63. <https://doi.org/10.1016/j.jand.2017.05.020>.
- Marchini, Jonathan, and Bryan Howie. 2010. "Genotype Imputation for Genome-Wide Association Studies." *Nature Reviews Genetics* 11 (7): 499–511. <https://doi.org/10.1038/nrg2796>.
- Maroille, Tatiana, and Maja Tarailo-Graovac. 2019. "Uncovering Missing Heritability in Rare Diseases." *Genes* 10 (4): 275. <https://doi.org/10.3390/genes10040275>.
- Martínez, A., C. Núñez, M. C. Martín, J. L. Mendoza, C. Taxonera, M. Díaz-Rubio, E. G. de la Concha, and E. Urcelay. 2007. "Epistatic Interaction between FCRL3 and MHC in Spanish Patients with IBD." *Tissue Antigens* 69 (4): 313–17. <https://doi.org/10.1111/j.1399-0039.2007.00816.x>.
- Matikainen-Ankney, Bridget A., Nebojsa Kezunovic, Caroline Menard, Meghan E. Flanigan, Yue Zhong, Scott J. Russo, Deanna L. Benson, and George W. Huntley. 2018. "Parkinson's Disease-Linked LRRK2-G2019S Mutation Alters Synaptic Plasticity and Promotes Resilience to Chronic Social Stress in Young Adulthood." *The Journal of Neuroscience* 38 (45): 9700–9711. <https://doi.org/10.1523/JNEUROSCI.1457-18.2018>.
- Meschia, James F., Michael Nalls, Mar Matarin, Thomas G. Brott, Robert D. Brown, John Hardy, Brett Kissela, et al. 2011. "Siblings With Ischemic Stroke Study: Results of a Genome-Wide Scan for Stroke Loci." *Stroke* 42 (10): 2726–32. <https://doi.org/10.1161/STROKEAHA.111.620484>.
- Moisan, Frédéric, Sofiane Kab, Fatima Mohamed, Marianne Canonico, Morgane Le Guern, Cécile Quintin, Laure Carcaillon, et al. 2016. "Parkinson Disease Male-to-Female Ratios Increase with Age: French Nationwide Study and Meta-Analysis." *Journal of Neurology, Neurosurgery & Psychiatry* 87 (9): 952–57. <https://doi.org/10.1136/jnnp-2015-312283>.
- Molodecky, Natalie A., Ing Shian Soon, Doreen M. Rabi, William A. Ghali, Mollie Ferris, Greg Chernoff, Eric I. Benchimol, et al. 2012. "Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review." *Gastroenterology* 142 (1): 46–54.e42. <https://doi.org/10.1053/j.gastro.2011.10.001>.
- Moore, Darren J., Andrew B. West, Valina L. Dawson, and Ted M. Dawson. 2005. "MOLECULAR PATHOPHYSIOLOGY OF PARKINSON'S DISEASE." *Annual Review of Neuroscience* 28 (1): 57–87. <https://doi.org/10.1146/annurev.neuro.28.061604.135718>.
- Mukherjee, Bhramar, Jaeil Ahn, Stephen B. Gruber, and Nilanjan Chatterjee. 2012. "Testing Gene-Environment Interaction in Large-Scale Case-Control Association Studies: Possible Choices and Comparisons." *American Journal of Epidemiology* 175 (3): 177–90. <https://doi.org/10.1093/aje/kwr367>.
- Mukherjee, Bhramar, Jaeil Ahn, Stephen B. Gruber, Gad Rennert, Victor Moreno, and Nilanjan Chatterjee. 2008. "Tests for Gene-Environment Interaction from Case-Control Data: A Novel Study of Type I Error, Power and Designs." *Genetic Epidemiology* 32 (7): 615–26. <https://doi.org/10.1002/gepi.20337>.
- Mukherjee, Bhramar, and Nilanjan Chatterjee. 2008. "Exploiting Gene-Environment Independence for Analysis of Case-Control Studies: An Empirical Bayes-Type Shrinkage Estimator to Trade-Off between Bias and Efficiency." *Biometrics* 64 (3): 685–94. <https://doi.org/10.1111/j.1541-0420.2007.00953.x>.
- Murcray, C. E., J. P. Lewinger, and W. J. Gauderman. 2008. "Gene-Environment Interaction in Genome-Wide Association Studies." *American Journal of Epidemiology* 169 (2): 219–26. <https://doi.org/10.1093/aje/kwn353>.
- Naj, Adam C. 2019. "Genotype Imputation in Genome-Wide Association Studies." *Current Protocols in Human Genetics* 102 (1). <https://doi.org/10.1002/cphg.84>.

- Nalls, Mike A, Cornelis Blauwendraat, Costanza L Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, Manuela Tan, et al. 2019. "Identification of Novel Risk Loci, Causal Insights, and Heritable Risk for Parkinson's Disease: A Meta-Analysis of Genome-Wide Association Studies." *The Lancet Neurology* 18 (12): 1091–1102. [https://doi.org/10.1016/S1474-4422\(19\)30320-5](https://doi.org/10.1016/S1474-4422(19)30320-5).
- Narendra, Derek P., Seok Min Jin, Atsushi Tanaka, Der-Fen Suen, Clement A. Gautier, Jie Shen, Mark R. Cookson, and Richard J. Youle. 2010. "PINK1 Is Selectively Stabilized on Impaired Mitochondria to Activate Parkin." Edited by Douglas R. Green. *PLoS Biology* 8 (1): e1000298. <https://doi.org/10.1371/journal.pbio.1000298>.
- Ng, Siew C, Hai Yun Shi, Nima Hamidi, Fox E Underwood, Whitney Tang, Eric I Benchimol, Remo Panaccione, et al. 2017. "Worldwide Incidence and Prevalence of Inflammatory Bowel Disease in the 21st Century: A Systematic Review of Population-Based Studies." *The Lancet* 390 (10114): 2769–78. [https://doi.org/10.1016/S0140-6736\(17\)32448-0](https://doi.org/10.1016/S0140-6736(17)32448-0).
- Ng, Siew C., Whitney Tang, Jessica Y. Ching, May Wong, Chung Mo Chow, A.J. Hui, T.C. Wong, et al. 2013. "Incidence and Phenotype of Inflammatory Bowel Disease Based on Results From the Asia-Pacific Crohn's and Colitis Epidemiology Study." *Gastroenterology* 145 (1): 158-165.e2. <https://doi.org/10.1053/j.gastro.2013.04.007>.
- Nussbaum, Robert L., and Christopher E. Ellis. 2003. "Alzheimer's Disease and Parkinson's Disease." Edited by Alan E. Guttmacher and Francis S. Collins. *New England Journal of Medicine* 348 (14): 1356–64. <https://doi.org/10.1056/NEJM2003ra020003>.
- Panaccione, Remo. 2013. "Mechanisms of Inflammatory Bowel Disease." *Gastroenterology & Hepatology* 9 (8): 529–32.
- Parkinson, James. 2002. "An Essay on the Shaking Palsy." *The Journal of Neuropsychiatry and Clinical Neurosciences* 14 (2): 223–36. <https://doi.org/10.1176/jnp.14.2.223>.
- Pecanka, Jakub, Marianne A. Jonker, International Parkinson's Disease Genomics Consortium (IPDGC), Zoltan Bochdanovits, and Aad W. Van Der Vaart. 2017. "A Powerful and Efficient Two-Stage Method for Detecting Gene-to-Gene Interactions in GWAS." *Biostatistics* 18 (3): 477–94. <https://doi.org/10.1093/biostatistics/kxw060>.
- Phillips, Patrick C. 2008. "Epistasis--the Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems." *Nature Reviews. Genetics* 9 (11): 855–67. <https://doi.org/10.1038/nrg2452>.
- Piegorsch, W. W., C. R. Weinberg, and J. A. Taylor. 1994. "Non-Hierarchical Logistic Models and Case-Only Designs for Assessing Susceptibility in Population-Based Case-Control Studies." *Statistics in Medicine* 13 (2): 153–62. <https://doi.org/10.1002/sim.4780130206>.
- Pinsk, Vared, Daniel A. Lemberg, Karan Grewal, Collin C. Barker, Richard A. Schreiber, and Kevan Jacobson. 2007. "Inflammatory Bowel Disease in the South Asian Pediatric Population of British Columbia." *The American Journal of Gastroenterology* 102 (5): 1077–83. <https://doi.org/10.1111/j.1572-0241.2007.01124.x>.
- Polgar, N., V. Csongei, M. Szabo, V. Zambo, B. I. Melegh, K. Sumegi, G. Nagy, Z. Tulassay, and B. Melegh. 2012. "Investigation of JAK2, STAT3 and CCR6 Polymorphisms and Their Gene-Gene Interactions in Inflammatory Bowel Disease: Investigation of JAK2, STAT3 and CCR6 Polymorphisms." *International Journal of Immunogenetics* 39 (3): 247–52. <https://doi.org/10.1111/j.1744-313X.2012.01084.x>.
- Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9. <https://doi.org/10.1038/ng1847>.
- Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3): 559–75. <https://doi.org/10.1086/519795>.
- Ramnarine, Shelina, Juan Zhang, Li-Shiun Chen, Robert Culverhouse, Weimin Duan, Dana B. Hancock, Sarah M. Hartz, et al. 2015. "When Does Choice of Accuracy Measure Alter Imputation

- Accuracy Assessments?" Edited by Chuhsing Kate Hsiao. *PLOS ONE* 10 (10): e0137601. <https://doi.org/10.1371/journal.pone.0137601>.
- Rees, Jonathan L. 2004. "The Genetics of Sun Sensitivity in Humans." *The American Journal of Human Genetics* 75 (5): 739–51. <https://doi.org/10.1086/425285>.
- Reich, Stephen G., and Joseph M. Savitt. 2019. "Parkinson's Disease." *Medical Clinics of North America* 103 (2): 337–50. <https://doi.org/10.1016/j.mcna.2018.10.014>.
- Roberts, R L, R K G Topless, A J Phipps-Green, R B Gearry, M L Barclay, and T R Merriman. 2010. "Evidence of Interaction of CARD8 Rs2043211 with NALP3 Rs35829419 in Crohn's Disease." *Genes & Immunity* 11 (4): 351–56. <https://doi.org/10.1038/gene.2010.11>.
- Rogler, Gerhard, and Stephan Vavricka. 2015. "Exposome in IBD: Recent Insights in Environmental Factors That Influence the Onset and Course of IBD." *Inflammatory Bowel Diseases* 21 (2): 400–408. <https://doi.org/10.1097/MIB.0000000000000229>.
- Rose, Geoffrey. 2001. "Sick Individuals and Sick Populations." *International Journal of Epidemiology* 30 (3): 427–32. <https://doi.org/10.1093/ije/30.3.427>.
- Rothman, Kenneth J., Sander Greenland, and Timothy L. Lash. 2008. *Modern Epidemiology*. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.
- Scheet, Paul, and Matthew Stephens. 2006. "A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase." *The American Journal of Human Genetics* 78 (4): 629–44. <https://doi.org/10.1086/502802>.
- Schulz, Kenneth F, and David A Grimes. 2002. "Case-Control Studies: Research in Reverse." *The Lancet* 359 (9304): 431–34. [https://doi.org/10.1016/S0140-6736\(02\)07605-5](https://doi.org/10.1016/S0140-6736(02)07605-5).
- Schurz, Haiko, Stephanie J. Müller, Paul David van Helden, Gerard Tromp, Eileen G. Hoal, Craig J. Kinnear, and Marlo Möller. 2019. "Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population." *Frontiers in Genetics* 10 (February): 34. <https://doi.org/10.3389/fgene.2019.00034>.
- Scott, L. J., K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, et al. 2007. "A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants." *Science* 316 (5829): 1341–45. <https://doi.org/10.1126/science.1142382>.
- Sherry, S. T. 2001. "dbSNP: The NCBI Database of Genetic Variation." *Nucleic Acids Research* 29 (1): 308–11. <https://doi.org/10.1093/nar/29.1.308>.
- Shiina, Takashi, Kazuyoshi Hosomichi, Hidetoshi Inoko, and Jerzy K Kulski. 2009. "The HLA Genomic Loci Map: Expression, Interaction, Diversity and Disease." *Journal of Human Genetics* 54 (1): 15–39. <https://doi.org/10.1038/jhg.2008.5>.
- Singh, Neeraj, Basu Banerjee, Kiran Bala, Mitrabasu Chhillar, and Neelam Chhillar. 2014. "Gene-Gene and Gene-Environment Interaction on the Risk of Parkinson's Disease." *Current Aging Science* 7 (2): 101–9. <https://doi.org/10.2174/1874609807666140805123621>.
- Singleton, Andrew, and John Hardy. 2016. "The Evolution of Genetics: Alzheimer's and Parkinson's Diseases." *Neuron* 90 (6): 1154–63. <https://doi.org/10.1016/j.neuron.2016.05.040>.
- Speelman, Arlène D., Jan T. Groothuis, Marlies van Nimwegen, Ellis S. van der Scheer, George F. Borm, Bastiaan R. Bloem, Maria T.E. Hopman, and Marten Munneke. 2012. "Cardiovascular Responses During a Submaximal Exercise Test in Patients with Parkinson's Disease." *Journal of Parkinson's Disease* 2 (3): 241–47. <https://doi.org/10.3233/JPD-2012-012111>.
- Stelzer, Gil, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, et al. 2016. "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses." *Current Protocols in Bioinformatics* 54 (1). <https://doi.org/10.1002/cpbi.5>.
- Stralen, Karlijn J. van, Friedo W. Dekker, Carmine Zoccali, and Kitty J. Jager. 2010. "Case-Control Studies ‐ An Efficient Observational Study Design." *Nephron Clinical Practice* 114 (1): c1–4. <https://doi.org/10.1159/000242442>.
- Tanner, Caroline M., Freya Kamel, G. Webster Ross, Jane A. Hoppin, Samuel M. Goldman, Monica Korell, Connie Marras, et al. 2011. "Rotenone, Paraquat, and Parkinson's Disease." *Environmental Health Perspectives* 119 (6): 866–72. <https://doi.org/10.1289/ehp.1002839>.

- The 1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. <https://doi.org/10.1038/nature15393>.
- the Haplotype Reference Consortium. 2016. "A Reference Panel of 64,976 Haplotypes for Genotype Imputation." *Nature Genetics* 48 (10): 1279–83. <https://doi.org/10.1038/ng.3643>.
- Thia, Kelvin T., William J. Sandborn, William S. Harmsen, Alan R. Zinsmeister, and Edward V. Loftus. 2010. "Risk Factors Associated With Progression to Intestinal Complications of Crohn's Disease in a Population-Based Cohort." *Gastroenterology* 139 (4): 1147–55. <https://doi.org/10.1053/j.gastro.2010.06.070>.
- Torres, Joana, Saurabh Mehandru, Jean-Frédéric Colombel, and Laurent Peyrin-Biroulet. 2017. "Crohn's Disease." *The Lancet* 389 (10080): 1741–55. [https://doi.org/10.1016/S0140-6736\(16\)31711-1](https://doi.org/10.1016/S0140-6736(16)31711-1).
- Trinh, Joanne, Florentine M.J. Zeldenrust, Jana Huang, Meike Kasten, Susen Schaaque, Sonja Petkovic, Harutyun Madoev, et al. 2018. "Genotype-Phenotype Relations for the Parkinson's Disease Genes SNCA, LRRK2, VPS35: MDSGene Systematic Review: MDSGene Systematic Review: SNCA, LRRK2, VPS35." *Movement Disorders* 33 (12): 1857–70. <https://doi.org/10.1002/mds.27527>.
- Tsironi, Eftychia, Roger M Feakins, Chris SJ Roberts, David S Rampton, and D Phil. 2004. "Incidence of Inflammatory Bowel Disease Is Rising and Abdominal Tuberculosis Is Falling in Bangladeshis in East London, United Kingdom." *American Journal of Gastroenterology* 99 (9): 1749–55. <https://doi.org/10.1111/j.1572-0241.2004.30445.x>.
- Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the **Metafor** Package." *Journal of Statistical Software* 36 (3). <https://doi.org/10.18637/jss.v036.i03>.
- Wan, Xiang, Can Yang, Qiang Yang, Hong Xue, Xiaodan Fan, Nelson L.S. Tang, and Weichuan Yu. 2010. "BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-Wide Case-Control Studies." *The American Journal of Human Genetics* 87 (3): 325–40. <https://doi.org/10.1016/j.ajhg.2010.07.021>.
- Weiss, Robert B., Timothy B. Baker, Dale S. Cannon, Andrew von Niederhausern, Diane M. Dunn, Nori Matsunami, Nanda A. Singh, et al. 2008. "A Candidate Gene Approach Identifies the CHRNA5-A3-B4 Region as a Risk Factor for Age-Dependent Nicotine Addiction." Edited by Jonathan Flint. *PLoS Genetics* 4 (7): e1000125. <https://doi.org/10.1371/journal.pgen.1000125>.
- Wider, C., C. Vilariño-Güell, M. G. Heckman, B. Jasinska-Myga, A. I. Ortolaza-Soto, N. N. Diehl, J. E. Crook, et al. 2011. "SNCA, MAPT, and GSK3B in Parkinson Disease: A Gene-Gene Interaction Study: SNCA, MAPT and GSK3B in Parkinson Disease." *European Journal of Neurology* 18 (6): 876–81. <https://doi.org/10.1111/j.1468-1331.2010.03297.x>.
- Wong, M. Y., N. E. Day, J. A. Luan, and N. J. Wareham. 2004. "Estimation of Magnitude in Gene–Environment Interactions in the Presence of Measurement Error." *Statistics in Medicine* 23 (6): 987–98. <https://doi.org/10.1002/sim.1662>.
- Wong, Suzy L., Heather Gilmour, and Pamela L. Ramage-Morin. 2014. "Parkinson's Disease: Prevalence, Diagnosis and Impact." *Health Reports* 25 (11): 10–14.
- Yadav, Pankaj, David Ellinghaus, Gaëlle Rémy, Sandra Freitag-Wolf, Anabelle Cesaro, Frauke Degenhardt, Gabrielle Boucher, et al. 2017. "Genetic Factors Interact With Tobacco Smoke to Modify Risk for Inflammatory Bowel Disease in Humans and Mice." *Gastroenterology* 153 (2): 550–65. <https://doi.org/10.1053/j.gastro.2017.05.010>.
- Yadav, Pankaj, Sandra Freitag-Wolf, Wolfgang Lieb, Astrid Dempfle, and Michael Krawczak. 2015. "Allowing for Population Stratification in Case-Only Studies of Gene-Environment Interaction, Using Genomic Control." *Human Genetics* 134 (10): 1117–25. <https://doi.org/10.1007/s00439-015-1593-y>.
- Yadav, Pankaj, Sandra Freitag-Wolf, Wolfgang Lieb, and Michael Krawczak. 2015. "The Role of Linkage Disequilibrium in Case-Only Studies of Gene-Environment Interactions." *Human Genetics* 134 (1): 89–96. <https://doi.org/10.1007/s00439-014-1497-2>.
- Yang, Q., and M. J. Khoury. 1997. "Evolving Methods in Genetic Epidemiology. III. Gene-Environment Interaction in Epidemiologic Research." *Epidemiologic Reviews* 19 (1): 33–43. <https://doi.org/10.1093/oxfordjournals.epirev.a017944>.

- Yang, Q., M. J. Khoury, and W. D. Flanders. 1997. "Sample Size Requirements in Case-Only Designs to Detect Gene-Environment Interaction." *American Journal of Epidemiology* 146 (9): 713–20. <https://doi.org/10.1093/oxfordjournals.aje.a009346>.
- Yates, Andrew D, Premanand Achuthan, Wasiru Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, et al. 2019. "Ensembl 2020." *Nucleic Acids Research*, November, gkz966. <https://doi.org/10.1093/nar/gkz966>.
- Zhang, Boshao, Degui Zhi, Kui Zhang, Guimin Gao, Nita N. Limdi, and Nianjun Liu. 2011. "Practical Consideration of Genotype Imputation: Sample Size, Window Size, Reference Choice, and Untyped Rate." *Statistics and Its Interface* 4 (3): 339–52. <https://doi.org/10.4310/sii.2011.v4.n3.a8>.
- Zhang, Jie, Zhi Wei, Christopher J. Cardinale, Elena S. Gusareva, Kristel Van Steen, Patrick Sleiman, International IBD Genetics Consortium, and Hakon Hakonarson. 2019. "Multiple Epistasis Interactions Within MHC Are Associated With Ulcerative Colitis." *Frontiers in Genetics* 10 (April): 257. <https://doi.org/10.3389/fgene.2019.00257>.
- Zuk, Or, Eliana Hechter, Shamil R. Sunyaev, and Eric S. Lander. 2012. "The Mystery of Missing Heritability: Genetic Interactions Create Phantom Heritability." *Proceedings of the National Academy of Sciences of the United States of America* 109 (4): 1193–98. <https://doi.org/10.1073/pnas.1119675109>.

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Dr. Astrid Dempfle, for the opportunity to work on this thesis at the Institute of Medical Informatics and Statistics (IMIS). I greatly appreciate the time, effort, patience and benefit of the doubt she put in to teach someone with a background in economics all about genetics and medical statistics. It was a pleasure to work for and learn from someone who has so much guidance to give in scientific, career related topics and beyond.

My utmost gratitude goes to Prof. Dr. Michael Krawczak for the invaluable discussions regarding my research as well as his friendly and attentive manner. I am grateful for his input, support and the many doors to opportunities he opened for me.

A special thank you goes to Prof. Dr. Hinrich Schulenburg for his guidance and advice. Moreover, I would like to thank the members of my dissertation committee, Prof. Dr. Tal Dagan and committee Chair Prof. Dr. Axel Scheidig for their time and intellectual contributions to my development as a scientist.

Furthermore, I would like to express my sincere gratitude to the funding source of my position, namely the Research Training Group 1743 "*Genes, Environment and Inflammation*", which is part of the *Deutsche Forschungsgemeinschaft*. Not only was the financial support family friendly, but the opportunity to meet doctoral candidates and their supervisors from different labs as well as numerous soft-skill courses enriched my experience.

Many thanks go to all my colleagues at IMIS. In particular to Prof. Dr. Silke Szymczak and Dr. Sandra Freitag-Wolf for their useful discussions and involvement. A special thanks go to Olaf Junge and Markus Schilhabel for their technical support and the time they saved all my work product after my computer decided to kick the bucket. A warm thank you goes to Petra Neumann for all her help and kind encouragement. It is a pleasure that some work relationships formed into friendships. I also thank my current colleagues Dr. Christoph Borzikowsky, Dr. Amke Caliebe, Dr. Carolin Knecht, Pegah Rahmati, Arunabh Sharma and my former colleagues Dr. Kristina Schlicht, Prof. Dr. Stephan Seifert and Brittany Burmester for their support and valuable life lessons learned. I hope they receive as much support in their future endeavours as they gave me.

Finally, I would especially like to thank my family and friends. My husband, Jan has been extremely supportive of me throughout the whole journey and has made countless sacrifices to help me. My daughter Laura has been very patient and always found the time to listen to mock presentations after kindergarten as well as provided the motivation to finish my degree with expediency. Jonas Danielius had very helpful and quick IT solutions. It goes without saying that my parents, Rasa and Prof. Dr. Gintaras Aleknoniai deserve a special thank you for their continued support and encouragement throughout my life. Without this strong team behind me, I doubt that I would be in this place at this time today.

Declaration

I hereby declare that this thesis is the outcome of my own work and effort. Apart from the advice and guidance of my supervisors, all third party help and any text or contents from other sources have been cited appropriately. I affirm in lieu of oath that up to this date I have not failed any dissertation procedures and that this thesis has not been previously submitted elsewhere. This work has been carried out in strict accordance with the rules of good scientific practice of the *Deutsche Forschungsgemeinschaft*.

Place, Date

Milda Aleknonytė-Resch

Curriculum Vitae

Personal Information

Name Milda Aleknonytė-Resch
Nationality Lithuanian
Birth Date 25.03.1990 in Vilnius, Lithuania
Marital Status Married, one daughter

Academic Training

06/2017 - now **Doctoral Studies in Natural Sciences**

Thesis title: The Validity and Statistical Power of the Case-Only Study Design for Interaction Analysis: Gene-Gene Interaction and the Role of Genotype Imputation in Gene-Environment Interaction
Kiel University, Germany

09/2012 – 02/2015 **Master of Science in Quantitative Economics**

Thesis title: Inequality and Risk Taking
Kiel University, Germany

09/2008 – 06/2012 **Bachelor in Economics**, specializing in Economic Analysis
Vilnius University, Lithuania

09/2010 – 09/2011 Exchange year abroad
Kiel University, Germany

09/2004 – 06/2008 **Secondary education**, with honours
Vilniaus licėjus, Lithuania

08/1997 - 06/2004 *International School of Prague, Czech Republic*

Work Experience

06/2017 – now **Research Associate**, Institute of Medical Informatics and Statistics, Kiel, Germany

04/2015 – 06/2017 **Data Scientist**, *meteolytix GmbH*, Kiel, Germany

09/2011 – 05/2012 **Junior Analyst**, *Civitta (business consulting)*, Vilnius, Lithuania

08/2008 – 07/2009 **Assistant (Interpreter)**, *Vilniaus Baldai*, Vilnius, Lithuania

Other Work Activities

11/2014 – 03/2015 **Student Assistant**, *Kiel Institute for the World Economy, Research group on Social and Behavioural Approaches to Global Problems*, Kiel, Germany

11/2013 – 07/2014 **Mathematics Tutor**, *Christian-Albrechts-Universität zu Kiel*, Kiel, Germany

04/2013 – 02/2014 **Student Assistant**, *Kiel Institute for the World Economy, Global Economic Symposium Department*, Kiel, Germany

06/2012 – 09/2012 **Working Student**, *MC Services*, Munich, Germany

06/2010 – 09/2010 **Intern**, *MC Services*, Munich, Germany

■ Computer skills

Microsoft Office	Professional
R	Advanced
SPSS	Basic
Typing speed	65-75 WPM

■ Languages

Lithuanian	Mother tongue
German	Proficient, Sprachdiplom (C1)
English	Proficient, TOEFL iBT score: 118/120

■ Publications

Aleknonytė-Resch M., Freitag-Wolf S., The International Inflammatory Bowel Disease Genetics Consortium, Schreiber S., Krawczak M., Dempfle A. Case-only analysis of gene–gene interactions in inflammatory bowel disease. *Scand J Gastroenterol* 2020;1–10. Doi: 10.1080/00365521.2020.1790646.

Klosa K., Shahid W., **Aleknonytė-Resch M.**, Kern M. Cleaning and Conditioning of Contaminated Core Build-Up Material before Adhesive Bonding. *Materials* 2020;13(12):2880. Doi: 10.3390/ma13122880.

Schmidt U., Neyse L., **Aleknonyte M.** Income inequality and risk taking: the impact of social comparison information. *Theory Decis* 2019;87(3):283–97. Doi: 10.1007/s11238-019-09713-8.

■ Awards and Certifications

- 2014 Tutor qualification certification, PerLe, Christian-Albrechts-Universität zu Kiel
- 2012 DAAD / ERP (Deutsch Akademischer Austauschdienst / European Recovery Program) scholarship for Master studies in Germany
- 2010 City of Kiel Scholarship to study for a full academic year (2010-11)
- 2009 Two scholarships (one each semester) for academic achievements
- 2008 The European Law Student's Association (elsa) law competition 1st place in Lithuanian finals
- 2007 International Youth Debates (Jugend debattiert international) 1st place in Vilnius tournament, 2nd place in Lithuanian finals, took part in European finals

■ Volunteering

- 2016 – now Member of International Youth Debates Alumni Club
- 2016 Managed a volunteer team made up of more than 70 volunteers during the Global Economic Symposium in Kiel
- 2011 Helped organize the Global Economic Symposium in Kiel
- 2010 Participated in a podium discussion with German Chancellor Dr. Angela Merkel
- 2009 Volunteered at the World Lithuanian Economic Forum (responsible for web translation and organization)