

The Genetics of IBD: From Susceptibility to Drug Response and Patient Outcome



Aleksejs Sazonovs

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Pembroke College

September 2019

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Aleksejs Sazonovs
September 2019

The Genetics of IBD: From Susceptibility to Drug Response and Patient Outcome

Aleksejs Sazonovs

Abstract

Inflammatory bowel disease (IBD) is a group of immune-mediated autoinflammatory disorders, primarily manifesting in the gastrointestinal tract. Affecting millions of people around the world, IBD has a severe impact on patients' quality of life. Several pharmacologic treatments have been available since the 1950s. However, the majority of patients either do not respond to a given therapy or lose response to a previously effective treatment and thus require therapeutic escalation.

In the first research chapter of my thesis, I describe the results of the Personalised Anti-TNF Therapy in Crohn's disease study. Immunogenicity to anti-TNF therapy is a major cause of loss of response, hypersensitivity reactions, and discontinuation of treatment in patients. Currently, immunogenicity cannot be predicted prior to treatment. My analysis has identified a strong dominant association in the HLA region on chromosome 6 (HLA-DQA1*05, $P=5.9 \times 10^{-13}$; HR=1.90; 95% CI, 1.60 to 2.25). Around 40% of individuals of European ancestry carry HLA-DQA1*05, and the data suggest that around 95% of these would develop immunogenicity within the first year of infliximab monotherapy treatment (a common anti-TNF treatment regime).

In the second research chapter of my thesis, I describe a genome-wide association study of thiopurine-induced liver damage (TILI). Ultimately, the study was underpowered to detect

any associations of moderate effect size and did not detect any associations of high effect size amongst the common genetic variants. Interestingly, I was not able to replicate the association in *PTPN22*, which was reported to be a risk factor for drug-induced liver damage by Cirulli et al. [39] – suggesting that its effect might be heterogeneous depending on the therapy.

Finally, the third research chapter describes the initial analysis of the IBD 15x dataset – a whole-genome sequencing association study of around 7,000 IBD patients paired with 12,000 matching controls. I provide an overview of the sample quality control procedures and describe some of the novel challenges that sequencing studies bring in comparison to standard GWAS (e.g., sample cross-contamination due to index mismatching). Finally, I also provide the results of the initial meta-analysis of the exome-sequencing dataset produced by the Broad Institute. The results demonstrate that rare coding genetic variants play a role in IBD pathogenesis.

I would like to dedicate this thesis to my parents – Galina Sazonova and Vadims Sazonovs

Acknowledgements

First of all, I would like to thank my supervisor Carl Anderson who has guided this work. I could not have wished for a more compassionate and brilliant mentor. Jeff Barrett, with whom I started my PhD, has shown me how passion for a topic can coincide with relentless rigour. The union of teams 143 and 152 – you have made me feel welcome here and I have learnt so much from each of you. The computational work (i.e. all) was enabled by the Human Genome Informatics team – Pavlos Antoniou, Chris Harrison, Colin Nolan, Alan Daly, Vivek Iyer, and Josh Randall. I've broken so many things and yet you seem to tolerate me. Nicole and other members of the Soranzo team, especially Klaudia Walter and Kousik Kundu for stimulating discussions and lively debates about 15x. Paris Litterick, Eloise Stapleton, Sally Bygraves, without whom these projects would have ground to a halt. The PANTS and TILI projects would not have been possible without the Exeter team – Nick Kennedy, Tariq Ahmad, Gareth Walker, Claire Bewshea and others. The last three years would not have been so joyful without my colleagues, who later became friends, then stopped being colleagues, but remained friends: Dan Rice, Loukas Moutsianas, Liu He, Mari Niemi, Fernando Riveros Mckay Aguilera, Scott Shooter, Tejas Shah, Arthur Gilly, and Sophie Hackinger. A heartfelt thank you to Emma Molloy for putting up with me during the final stretch of the PhD. Finally, I would like to express my gratitude to my family who have selflessly supported me on this journey. I owe so much to you.

Table of contents

List of figures	xvii
List of tables	xxi
Nomenclature	xxiii
1 Introduction	1
1.1 Common and rare variant studies of complex traits	1
1.1.1 Genome-wide association studies	1
1.1.2 Next-generation sequencing	2
1.2 Current approaches to resolving genome-wide association study signals . .	3
1.2.1 Statistical fine-mapping	3
1.2.2 Trans-ethnic association studies	5
1.2.3 Regulatory target analysis	6
Genetic variation and gene expression	6
Functional genomics and regulatory connections	7
1.3 Using rare variants to resolve genome-wide association studies	8

1.3.1	Measuring rare variation	9
	Genotyping and genotype imputation	9
	Targeted sequencing	10
	Whole-exome and whole-genome sequencing	11
1.3.2	Testing for statistical association with rare variants	13
	Single-variant tests	14
	Burden tests	15
1.3.3	Special study designs in the sequencing era	18
	Families and population isolates	18
	Biobanks and risk prioritisation	19
1.4	Inflammatory bowel disease	19
1.4.1	Manifestation	20
1.4.2	Epidemiology	22
	Prevalence and incidence	22
1.4.3	Disease aetiology	22
	Environmental factors	22
	Diet and lifestyle	23
1.4.4	Therapies and treatment options	24
1.4.5	Genetic component of inflammatory bowel disease	25
	Clues from observational epidemiology	27
	Linkage studies	27

Genome-wide association studies	28
Whole-genome and whole-exome sequencing association studies	29
Prediction of IBD via polygenic risk scores	29
Pharmacogenetic studies	31
2 HLA-DQA1*05 is associated with immunogenicity to anti-TNF therapy	33
2.1 Introduction	34
2.2 Methods	36
2.2.1 PANTS study: patient recruitment and phenotyping	36
2.2.2 Measurement of drug and anti-drug antibody levels	37
2.2.3 Genotyping, quality control, and imputation	38
2.2.4 Statistical and genome-wide association analyses	42
2.3 Results	43
2.3.1 A locus within the HLA region is associated with time to immunogenicity	43
2.3.2 Fine-mapping of the signal in the HLA region	43
2.3.3 The effect of HLA-DQA1*05 across drug and treatment regimes	47
2.4 Discussion	50
3 Attempting to identify the genetic determinants of thiopurine-induced liver injury	59
3.1 Introduction	59
3.2 Methods	62

3.2.1	Genotyping, cohort assembly, and quality control	62
	Assembling the dataset	62
	Sample quality control	63
	Variant QC	65
3.2.2	Statistical testing	65
3.2.3	Power calculation	67
3.3	Results	67
3.3.1	Case-control analysis	67
3.3.2	rs2476601 in <i>PTPN22</i> does not appear to be associated with TILI .	71
3.4	Discussion	72
4	Quality control and the initial analysis of the IBD 15x cohort	75
4.1	Introduction	75
4.2	Methods	78
4.2.1	Power modelling	78
	Variant calling sensitivity at different depths	79
	Estimating the statistical power of sequencing studies	80
4.2.2	Sample selection	81
	Cases: IBD 15x	81
	Controls: INTERVAL 15x	82
4.2.3	Sequencing	82
4.2.4	Alignment and variant calling	83

4.2.5	Computational analysis pipeline	83
4.2.6	Dataset overview and pre-processing	85
4.2.7	Sample quality control	85
	Hard filters	86
	Distribution-based filters	87
	Identifying batch effects via metric-based PCA	87
	Identifying genetic ancestry outliers via 1000G PCA loading projection	90
	Removal of duplicated and related samples	91
	Within-cohort principal component analysis	96
4.2.8	Variant and site quality control	101
4.3	Results	104
4.3.1	Depth and the statistical power trade-off	104
	Optimal sequencing depth for case-control experiments	104
	Choice of sequencing depth for biobank-scale projects	105
4.3.2	Index misassignment impacts multiplexed sequencing	108
4.3.3	Estimating the impact of the covariates	111
4.3.4	Meta-analysis with the Broad IBD WES results	114
4.4	Discussion	116
5	Discussion	123
5.1	Longitudinal studies for drug response leveraging expression data	124
5.2	Host-microbiome interactions	125

5.3	Extending anti-TNF pharmacogenetic analysis	126
5.3.1	Analysing retrospective data from the NIHR IBD BioResource . . .	127
5.3.2	Prescription records from the UK Biobank	128
	References	131

List of figures

1.1	Matching SNPs to genes via eQTLs	7
1.2	Comparison of whole-exome and whole-genome sequencing	12
1.3	One-stage association study power calculations for single-variant tests	14
1.4	Sliding window technique	17
1.5	Grouping promoters and enhancers	17
1.6	Primary IBD sub-types: UC and CD	21
1.7	IBD treatment pyramid	25
2.1	Principal component analysis (PCA) on imputed data from 1,323 individuals in the PANTS study	39
2.2	PANTS cohort flowchart	40
2.3	Manhattan plot for Cox proportional hazards model analysis of time to immunogenicity	44
2.4	Quantile-quantile plot for Cox proportional hazards model analysis of time to immunogenicity	46
2.5	MHC regional plot	47
2.6	Effect sizes SNP, HLA alleles, and amino acids	48

2.7	Rate of anti-drug antibody development stratified by the number of HLA-DQA1*05 alleles carried	49
2.8	Residual association signal in the MHC region, after conditioning on HLA-DQA1*05	50
2.9	Sensitivity analysis of the HLA-DQA1*05 association	51
2.10	Effect of HLA-DQA1*05 in different patient subgroups	52
2.11	HLA-DQA1*05 has a consistent effect on immunogenicity between patients treated with the infliximab originator, Remicade, and its biosimilar CT-P13	53
2.12	Kaplan–Meier estimator showing the rate of anti-drug antibody development	54
2.13	Kaplan–Meier estimator showing the rate of drug persistence	55
3.1	Projection of weights derived from the 1000 Genome Project PCA	64
3.2	PCA of the PRED4 TILI cohort	66
3.3	Power to detect single-variant associations for the PRED4 TILI cohort	68
3.4	Quantile-quantile plot for the case-control analysis of the PRED4 TILI cohort	69
3.5	Manhattan plot for case-control analysis of the PRED4 TILI cohort	70
4.1	Histogram of mean per-sample coverage for samples in the combined 15x cohort	79
4.2	Influence of sequencing depth on the ability to detect SNPs and INDELS	80
4.3	Set intersection of the samples failing different QC metric thresholds in the 15x dataset	88
4.4	The first four principal components built based on the QC metrics of the samples in the 15x cohort	89

4.5	The first four principal components of the 1000G cohort, alongside the 15x cohort samples projected onto them	92
4.6	Set intersection plot for the 434 genetic ancestry outlier samples from 15x. .	93
4.7	The first four principal components of the European subset of the 1000G cohort, alongside the 15x cohort samples projected onto them	94
4.8	Violin plot comparing the distributions of the projected 1000G PC scores for IBD and INTERVAL samples that pass the ancestry filters	95
4.9	The first four principal components of 15x cohort samples	97
4.10	Set intersection plot of PCA outliers	98
4.11	Correlation between PC1 and PC6 _{1000G}	99
4.12	Violin plot comparing the distributions of the projected PC scores for IBD and INTERVAL (without outlier removal)	100
4.13	VQSR calibration results for IBD & INTERVAL 15x cohorts	103
4.14	Power to discover rare SNP variants in a case-control experiment setup . . .	104
4.15	Power to discover associations using burden tests	106
4.16	Power to discover associations by aggregated rare variants in the UK Biobank ($\alpha = 1 \times 10^{-9}$)	107
4.17	FREEMIX scores of read groups	110
4.18	Mean FREEMIX	111
4.19	Influence of the inclusion of different covariate types on the power to identify known Crohn's associations in the IBD 15x cohort	112
4.20	Betas of the known Crohn's disease associations estimated in the 15x cohort	113

List of tables

2.1	Baseline demographic and clinical characteristics	41
2.2	Covariates used in the final model	42
2.3	Comparison between different models for the observed effect in the MHC region	45
4.1	Summary statistics for the meta-analysed variants within the known IBD-associated regions	117
4.2	Summary statistics for the meta-analysed variants outside of the known IBD-associated regions	118

Nomenclature

Acronyms / Abbreviations

ADA anti-drug antibody

ADR adverse drug reaction

EWAS exome-wide association study

GWAS genome-wide association study

HLA human leukocyte antigen

CD Crohn's disease

IBD inflammatory bowel disease

IBDU unclassified IBD

UC ulcerative colitis

LD linkage disequilibrium

MAF minor allele frequency

MHC major histocompatibility complex

TNF tumour necrosis factor

VEO very early onset

WES whole-exome sequencing

WGS whole-genome sequencing

