

Understanding Inflammatory Bowel
Disease using High-Throughput
Sequencing



Katrina Melanie de Lange
Trinity College
University of Cambridge

May 2017

Dissertation submitted for the degree of Doctor of Philosophy

Understanding Inflammatory Bowel Disease using High-Throughput Sequencing

Katrina Melanie de Lange, Trinity College, University of Cambridge

For over two decades, the study of genetics has been making significant progress towards understanding the causes of common disease. Across a wide range of complex disorders there have been hundreds of associated loci identified, largely driven by common genetic variation. Now, with the advent of next-generation sequencing technology, we are able to interrogate rare and low frequency variation in a high throughput manner for the first time. This provides an exciting opportunity to investigate the role of rarer variation in complex disease risk on a genome-wide scale, potentially offering novel insights into the biological mechanisms underlying disease pathogenesis. In this thesis I will assess the potential of this technology to further our understanding of the genetics of complex disease, using inflammatory bowel disease (IBD) as an example.

After first reviewing the history of genetic studies into IBD, I will describe the analytical challenges that can occur when using sequencing to perform case-control association testing at scale, and the methods that can be used to overcome these. I then test for novel IBD associations in a low coverage whole genome sequencing dataset, and uncover a significant burden of rare, damaging missense variation in the gene *NOD2*, as well as a more general burden of such variation amongst known inflammatory bowel disease risk genes. Through imputation into both new and existing genotyped cohorts, I also describe the discovery of 26 novel IBD-associated loci, including a low frequency missense variant in *ADCY7* that approximately doubles the risk of ulcerative colitis. I resolve biological associations underlying several of these novel associations, including a number of signals associated with monocyte-specific changes in integrin gene expression following immune stimulation.

These results reveal important insights into the genetic architecture of inflammatory bowel disease, and suggest that a combination of continued array-based genome-wide association studies, imputed using substantial new reference panels, and large scale deep sequencing projects will be required in order to fully understand the genetic basis of complex diseases like IBD.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as declared in the contributions section of each chapter and/or specified in the text. It is not being concurrently submitted for a degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation 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. It does not exceed the prescribed word limit of 60,000 words.

Katrina M. de Lange
May 2017

Acknowledgements

First and foremost, I would like to thank my supervisor, Jeffrey Barrett, for the endless help, advice and support over the past few years. Your commitment to statistical rigour, reproducible research, data sharing and clear scientific writing has been inspiring, and you have taught me more things than I can count. Perhaps most importantly, you have shown me that it is possible to be an amazing scientist and still have fun. I would also like to thank my secondary supervisor, Miles Parkes, and my thesis committee members, Richard Durbin, Gosia Trynka and Trevor Lawley, for their help and guidance along the way.

I would like to extend a particular thank you to Yang Luo, for taking me under her wing and sharing her extensive knowledge on all things sequencing. Your constant smile and boundless enthusiasm made working together both enjoyable and incredibly rewarding. To everyone else who contributed so much to the projects discussed in this thesis - especially Carl Anderson, Loukas Moutsianas, and Luke Jostins - thank you for all your hard work and dedication, and for the jokes and laughter that got us through the difficult bits. Finally, this work would not have been possible without the funding of the Wellcome Trust, and the countless individuals who donated samples for us to study; thank you for your generosity and your commitment to science.

I am forever grateful to those who got me here in the first place. To Tony Smith, whose infectious enthusiasm for bioinformatics was the trigger I needed to start down this path, and to all my supervisors at the University of Waikato who gave me the opportunity to find my fit in the world of research. A big thank you also goes to the Woolf Fisher Trust, without whom I never would have had the fantastic opportunity to pursue a PhD at the University of Cambridge. Thank you not only for giving me this chance, but for continuing to believe in me and the value of my work.

I would also like to thank those people who have made my time here in Cambridge so enjoyable. To the rest of the Barrett team: thank you for the laughs, lunches, and fascinating discussions. It has been a pleasure to work with you all! To the

other PhD students I have gotten to know so well at Sanger: the experience would not have been the same without you. To Ellese, Mariel, Nicola, Sumana, John and Alice: thank you for all the fun times, and all the support in the not-so-fun times. I wouldn't have made it through without you. Finally, to Julian. For everything.

Lastly, I want to thank my family, especially my siblings, Andrew and Robyn, and my parents, Vicki and Willem. Thank you for supporting me in everything I do, and for putting things in perspective. Because, sometimes, it really is more important that you go climb a mountain rather than write your thesis.

Publications

Arising from this dissertation

de Lange, K. M. & Barrett, J. C. (2015). Understanding inflammatory bowel disease via immunogenetics. *Journal of Autoimmunity*. 64, pp. 91–100

Luo, Y.* , **de Lange, K. M.***, Jostins, L., Moutsianas, L., Randall J. et al (2017). Exploring the genetic architecture of inflammatory bowel disease by whole genome sequencing identifies association at *ADCY7*. *Nature Genetics* 49, pp. 186–192

de Lange, K. M.*, Moutsianas, L.* , Lee, J. C.* , Lamb, C. A., Luo, Y. et al (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics* 49, pp. 256–261

Contents

1	Background and historical perspective	1
1.1	Inflammatory bowel disease	2
1.1.1	Clinical presentation	2
1.1.2	Epidemiology	5
1.2	The early days of IBD genetics	6
1.2.1	Twin studies	6
1.2.2	Linkage studies	7
1.2.3	Limitations of linkage studies and the common disease, common variant hypothesis	10
1.3	The GWAS era	11
1.3.1	Technological developments that made GWAS possible	11
1.3.2	GWAS: a revolution in IBD genetics	12
1.3.3	Meta-analyses and the importance of sample size	14
1.3.4	IBD genetics in the context of other diseases	16
1.3.5	Expanding into non-European populations	19
1.4	Beyond GWAS	20
1.4.1	Rare and low frequency variation	21
1.4.2	Identifying the casual mutations	23
1.5	Aims and overview	25
2	Case-control association testing using sequencing data	29
2.1	Introduction	29
2.1.1	Chapter overview	30
2.1.2	Contributions	31

2.2	Next-generation sequencing studies	32
2.2.1	Study design considerations	32
2.2.2	Challenges of performing case-control analyses	33
2.3	Low frequency and common variants	36
2.3.1	Joint calling across samples	36
2.3.2	Genotype refinement	37
2.3.3	Imputation of GWAS cohorts	38
2.4	Rare variant association testing	39
2.4.1	Increasing power using burden testing	40
2.4.2	Accounting for differences in sensitivity and specificity	43
2.4.3	Testing in a dataset with systematic read depth bias	47
2.4.4	Adjusting the quality control procedures	50
2.4.5	Increasing the size of the burden test	57
2.5	Discussion	59
3	The role of rare and low frequency variation in IBD risk	61
3.1	Introduction	61
3.1.1	Chapter overview	62
3.1.2	Contributions	63
3.2	Data preparation	64
3.2.1	Low coverage whole genome sequencing	64
3.2.2	Variant calling and imputation improvement	65
3.2.3	Quality control	69
3.3	Structural variation	73
3.4	Rare variation	74
3.4.1	Additional quality control	74
3.4.2	Burden testing across coding regions	75
3.4.3	Burden testing across non-coding regions	86
3.5	Low frequency variation	94
3.5.1	Imputation into GWAS	95
3.5.2	Quality control and association testing	97
3.5.3	p.Asp439Glu in <i>ADCY7</i> doubles risk of ulcerative colitis	98
3.6	Discussion	103

4	Uncovering the biological mechanisms driving association	107
4.1	Introduction	107
4.1.1	Chapter overview	108
4.1.2	Contributions	109
4.2	Data preparation	110
4.2.1	A new UK IBD genome wide association study	110
4.2.2	Imputation using an IBD-specific reference panel	114
4.2.3	Meta-analysis of sequencing and imputed genomes with existing summary statistics	114
4.2.4	Quality control	115
4.3	Unravelling common variant associations	117
4.3.1	Fine-mapping and functional annotation of new and known loci	117
4.3.2	Enrichment amongst IBD loci for genes associated with Mendelian disorders of inflammation and immunity	125
4.3.3	Co-localization of GWAS and eQTL associations	128
4.3.4	Therapeutic relevance of genetic associations	133
4.4	Discussion	136
5	Discussion and future directions	139
5.1	Studying complex genetic disease in the sequencing era	141
5.1.1	Exome vs whole genome sequencing	142
5.1.2	Combining and analysing data across multiple studies	146
5.1.3	Overcoming computational limitations	148
5.1.4	The future of locus discovery	149
5.2	Prospects for translation into the clinic	150
5.2.1	Integration with functional datasets	150
5.2.2	Informing treatment	151
5.2.3	Environmental factors: the microbiome	154
5.3	Concluding remarks	156

Contents

A	Contributing members of the UKIBDGC	157
B	Meta-analysis association statistics at all 241 known loci	161
	List of Tables	173
	List of Figures	175
	Bibliography	179