Using Next-Generation Genomic Datasets In Disease Association

Luke Jostins

King's College
University of Cambridge
September 2012

This dissertation is submitted for
the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the Attributions section or the text. It does not exceed the word limit set out by the Degree Committee for the Faculty of Biology, and is not substantially the same as any work that has been, or is being, submitted to any other university for any degree, diploma or any other qualification.

Luke Jostins
18^{th} September 2012

This is a post-viva dissertation, containing some minor corrections to that submitted on 18^{th} September 2012. The corrections were suggested by Stephen Sawcer and Peter Donnelly.

Luke Jostins
24^{th} January 2012

How are you going to be successful in treatment, if you do not understand the real essence of each disease?

— Galen, On Natural Faculties

Using Next-Generation Genomic Datasets In Disease Association

Luke Jostins
King's College, Cambridge
Ph.D Thesis, September 2012

Abstract

The first generation of genome-wide association studies (GWAS) uncovered thousands of genetic risk factors for hundreds of complex human diseases. However, over the past five years new high-throughput techniques, including next-generation sequencing and low-cost custom genotyping, have allowed us to expand disease association studies into larger sample sizes and across the entire spectrum of human variation. This thesis will explore the potential of these new technologies, and in particular their application to the study of Inflammatory Bowel Disease (IBD) genetics. After reviewing the historical context of complex disease genetics, I introduce the statistical methods and models used in this thesis, and demonstrate how they can be placed into a unified framework of genetic risk models. I then detail three analysis projects that focus on identifying risk variants that the first generation of GWAS was unable to study. The first investigates how genotype imputation, coupled with high-density sequencing reference sets, can aid locus discovery in both European and African populations. The second discusses the use of a custom genotyping chip (the Immunochip) to discover risk variants with low effect sizes, by allowing low-cost genotyping of a very large number of samples. The third investigates the use of next-generation sequencing of multiply affected (or "multiplex") families in order to identify low-frequency, high penetrance risk alleles. Throughout these three projects I describe the discovery of a large number of novel IBD risk loci, and discuss how statistical and biological interrogation of these risk loci can help us to develop and expand biological hypotheses.

Acknowledgements

First and foremost I would like to thank my adviser Jeff Barrett for more things than I could reasonably list here. Needless to say, his dedication to statistical rigour, free sharing of data, clear scientific writing and public communication have provided me with a model scientist that I will always seek to emulate.

I would like express my gratitude to my other advisers, thesis committee members and first year examiners, including Richard Durbin, Julian Parkhill, Chris Tyler-Smith and Ralph McGinnis. I would especially like to thank my external adviser John Todd for the confidence he has shown in me over the last three years.

I would also like to thank the rest of Medical Genomics/ Barrett Group/ Statistical and Computational Genetics/ Team 143 (we'll settle on a name one day): Kate Morley, for teaching me most of what I know, James Morris, for providing the perfect combination of well-planned software development, wry humour and cake, Yang Luo for Chinese sweets and old-school Cambridge amiability (dan shi...), Iris Kolder for her irrepressible sense of mischief and Isabelle Claynen for her infectious passion for biology.

I'd also like to thank all the people that I have got to know during my time at Sanger: Annabel Smith, Christina Hedberg-Delouka and Alex Bateman for running such a tight ship. The Anderson Group, including Carl Anderson, Jimmy Liu and Jamie Floyd, for tea and laughs. Daniel MacArthur for acting as my perennial foil, and Liz Murchison for being a true friend. Sanger's army of pipeline developers past and present, including Thomas Keane, Shane McCarthy, Petr Danecek, Jim Stalker, Josh Randall and Martin Pollard, for

keeping my life (and the lives of hundreds of others) running smoothly, Chloe Noble and Anja Kolb-Kokocinski for somehow keeping Human Genetics from constantly falling over. Anne Wombwell, for taking the pain out of conference travel. All the other people who have made Sanger what is it for me, including (but definitely not limited to) Nicole Soranzo, Aaron Day-Williams, Aarno Palotie, Dan Gaffney, Matt Hurles and Helena Kilpinen.

This thesis would not have been possible without my collaborators, and I would like to thank them all: The clinicians and analysts of the UKIBDGC, for making me feel part of something important, including Miles Parks, Charlie Lees, James Lee, Tim Raine, John Mansfield, Christopher Matthew and Jack Satsangi. The many researchers of the IIBDGC, for showing me just how big "Big Science" can get, including Judy Cho, Andre Franke, Rinse Weersma, Sarah West, Yashoda Sharma, Phil Schumm, Kaida Ning, Severine Vermeire, and Dermot McGovern. All the people that make Boston a powerhouse of genetic analysis, including Stephan Ripke, Johan Essers, Hailiang Huang, Mark Daly, Soumya Raychaudhuri, Xinli Hu and Lizzy Rossin. Closer to home, the statistical know-how of the Diabetes and Inflammation Laboratory, including David Clayton and Chris Wallace. The dedication of the MalariaGEN researchers, Chris Spencer, Gavin Band, Katja Kivinen, Kirk Rockett, Quang Li and Dominic Kwiatkowski, for always doing the right thing, instead of the easy thing. I'd especially like to thank Adam Levine and Tony Segal from UCL for giving me a taste of real biology.

I would like to thank my fellow members of Genomes Unzipped, including Joe Pickrell, Don Conrad, Dan Vorhaus, Caroline Wright and Vincent Plagnol, for fighting all the good fights we so love to fight. And the many people I have met through blogging and twitter for drawing me into their community, and in particular Chris Gunter, Zoe McDougall and Razib Khan for all their support.

I'd also like to thank Daniel Stretch and Stephen Eglen from the Cambridge Computational Biology Institute, and Jules Griffin and Sarah Lummis from King's College, for letting me loose on five years worth of Cambridge students. And I'd like to thank everyone at King's, for taking in an overconfident Physics student just out of A-levels and producing an overconfident

geneticist eight years later.

I'd like to thank all of my friends in Cambridge who have been so important to forming how I think about the world, and my parents for putting me on it the first place. I'd also like to thank Hannah "Pixie" Price, for everything.

I would like to thank the research bodies who have funded the research in the thesis, including the Wellcome Trust, the Medical Research Council, the National Association for Colitis and Crohns disease, the Charles Wolfson Charitable Trust and the Gates Foundation, as well as the many international funders who supported the IIBDGC. I'd also like to thank Gil McVean, Anne Pratt, Ali Momin, Helen Thompson and the Wellcome Trust for ensuring that I have somewhere to go next.

Last, but certainly not least, I would like to thank the 104,341 (give or take a few) patients and volunteers who donated their DNA to all the studies that make up this thesis.

Attributions

Many of the projects described in this thesis are of a collaborative nature, and many were performed as part of large, international consortia. Below is a summary of the contributions of other scientists to the work described in this thesis.

Chapter 3

The African data used for testing imputation were generated as part of the MalariaGEN Consortial Project 1 (CP1). Samples were collected by partners from 12 centres, and full information on all sample collections is available from `http://www.malariagen.net/projects/cp1`. Genotyping was carried out at the Wellcome Trust Sanger Institute. Pre-imputation data processing, quality control and calling of genotypes was carried out by Kirk Rockett, Katja Kivinen, Gavin Band, Si Quang Le and Chris Spencer.

The list of loss-of-function (LoF) variants was generated by the 1000 Genomes LoF Group (now Functional Integration Group). The list of high quality, polymorphic LoF variants was generated by Daniel MacArthur.

Chapter 4

The samples used in the combined GWAS-Immonchip analysis were collected by the research groups of the International IBD Genetics Consortium (`http://www.ibdgenetics.org/groups.html`). Genotyping was performed across multiple centres, summarised in Table 4.8.

Quality control, genotype imputation and association testing for the

GWAS collections were performed by Stephan Ripke. The optiCall genotyping program used to call Immunochip samples was developed by Tejas Shah and Carl Anderson. The tag SNP meta-analysis technique was developed in collaboration with Stephan Ripke and Mark Daly, and implemented by Stephan Ripke. The phenotype likelihood modelling approach was developed and implemented by Jonah Essers.

The DAPPLE analysis was carried out by Stephan Ripke and Lizzie Rossin. The eQTL and cSNP analyses were carried out by Kaida Ning. The immune cell expression enrichment analysis was carried out by Xinli Hu and Soumya Raychaudhuri, and the gene expression network analysis was carried out by Ken Hui and Eric Shadt.

Immunochip cluster plots were manually inspected by Mitja Mitrovic, Jeff Barrett, Carl Anderson, Emilie Theatre, Tobias Balschun, Sarah West, Kaida Ning, Zhi Wei, Karin Fransen, Kyle Bailey, Isabelle Cleynen, Suzanne van Sommeren, Philippe Goyette, Sok Meng Evelyn Ng and Martin Ladouceur.

Chapter 5

All research described in Chapter 5 was carried in close collaboration with Adam Levine. There is no single element of this project that I carried out entirely without Adam's input, and virtually all of Section 5.4 was carried out jointly.

However, the following were carried out by Adam with little or no input from me:

- The collection of all samples and phenotype data
- The processing and analysis of the gene expression data
- The calling and quality control of genome-wide genotype data
- The calling of SNPs and indels in the exome data

Gavin Sewell selected the SNPs used to test the load of common IBD risk variants in the family. CytoSNP 12 genotyping was carried out at University College, London, and Sequenom genotyping, whole-genome sequencing

and whole-exome sequencing was carried out at the Wellcome Trust Sanger Institute. QC of the exomes was carried out by Martin Pollard.

Chapter 6

The whole genome sequencing experiment was designed in collaboration with the UK IBD Genetics Consortium, who provided the samples. Sample selection was performed by James Lee. Sequencing was carried out at the Wellcome Trust Sanger Institute, and sequence data was processed and quality controlled by Martin Pollard and Josh Randall. Variant calling, imputation, stringent sample and variant QC and association analysis were carried out by Yang Luo.

Publications

From this Dissertation

- Band, G., Quang, S. L., **Jostins, L.**, Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F., *et al* (2012) Imputation based meta-analysis of severe malaria in three African populations. (Under review).

- **Jostins, L.**, Ripke, S., Weersma, R., Duerr, R. H., McGovern, D. P, Hui, K. Y., Lee, J. C., *et al* (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease *Nature* **491**(7422):119-124.

- MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., **Jostins, L.**, *et al* (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**(6070):823-828

- **Jostins, L.** and Barrett, J. C. (2011) Genetic risk prediction in complex disease. *Hum Mol Genet.* **20** (R2):R182-8.

- **Jostins, L.**, Morley, K. I. and Barrett, J. C. (2011) Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet.* **9**(6):662-666.

Arising elsewhere

- **1000 Genomes Project Consortium** (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422):56-65.

- Liu, J. Z., Almarri, M. A., Gaffney, D. J., Mells, F. G., **Jostins, L.**, Cordell, H. J., Ducker, S. J., *et al* (2012) Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis *Nat Genet.* **44**(10):1137-1141.

- **Jostins, L.** (2012) Dispatches from the functional phase of genome biology *Genome Biol.* **13**(6):316.

- **Jostins, L.**, Pickrell, J. K., Macarthur, D. G. and Barrett J. C. (2012) Misuse of hierarchical linear models overstates the significance of a reported association between OXTR and prosociality. *Proc Natl Acad Sci USA* **109**(18):E1048.

- Sewell, G. W., Rahman, F. Z., Levine, A. P., **Jostins, L.**, Smith, P. J., Walker, A. P., Bloom, S. L., Segal, A. W. and Smith, A. M. (2012) Defective tumor necrosis factor release from Crohn's disease macrophages in response to toll-like receptor activation: Relationship to phenotype and genome-wide association susceptibility loci. *Inflamm Bowel Dis.* **8**(11):2120-2127.

- Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., ***et al*** (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet.* **42**(12):1118-1125.

- **1000 Genomes Project Consortium** (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**(7319):1061-1073.

- Krawitz, P., Rödelsperger, C., Jäger, M., **Jostins, L.**, Bauer, S. and Robinson, P.N. (2010) Microindel detection in short-read sequence data. *Bioinformatics* **26** (6):722-729.

Contents