

Understanding human disease using high-throughput sequencing



Eva Gonçalves Serra

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Declaration

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute between October 2012 and August 2016. 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 Colleagues 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.

This is a post-viva dissertation, containing some minor corrections to that submitted on August 2016. The corrections were suggested by Dr Helen Firth and Prof William Newman.

Eva Gonçalves Serra
November 2016

Hold the vision, trust the process.

Abstract

Next-generation sequencing (NGS) is revolutionising Mendelian and complex disease research by enabling variant information to single-base resolution in a high-throughput way, scalable to the size of the human genome. In this dissertation, I describe four distinct projects in which NGS technologies were employed, in combination with different study designs and analytical strategies, to identify genetic determinants, or modifiers, of diseases that have been poorly studied thus far.

In Chapter 1, I provide a historical background of our understanding of how genetic variation contributes to disease phenotypes, and the technological advances in the last twenty years that have led to the NGS-based gene-mapping studies of today.

In Chapter 2, I describe a NGS-based screening of genes that are known to cause thyroid hormone production defects in a congenital hypothyroidism (CH) cohort of patients with *gland-in-situ*. I show how a stringent variant filtering pipeline, combined with pedigree segregation analyses and *in silico* predictions of pathogenicity for candidate variants, led to the identification of likely causal mutations in 59% of the patients.

In Chapter 3, I describe a family-based exome and targeted-sequencing analysis to identify novel genetic causes of CH. I explore different variant filtering pipelines to map *de novo*, inherited and copy-number-variants segregating with disease within families. I find that no gene is recurrently mutated across families over what is expected by chance. I then explore how a candidate-gene screening approach, leveraging rare disruptive mutations mapped in families, can highlight novel genes potentially associated with CH or the extrathyroidal features of some patients.

In Chapter 4, I describe a series of analyses to better understand the genetic architecture of very-early-onset inflammatory bowel disease (VEO-IBD). This condition is currently viewed as a Mendelian form of inflammatory bowel disease (IBD), a complex disorder of adulthood onset. Using exome data, I identify likely causative defects in known primary-immunodeficiency genes and explore the broader contribution of rare variants to VEO-IBD through case-control enrichment analyses at the level of single genes,

genesets and biological pathways. Moving beyond rare alleles, I generate polygenic risk scores leveraging the set of known, adult-onset IBD-risk alleles discovered to date, and demonstrate a polygenic component operating in VEO-IBD.

In Chapter 5, I describe a meta-analysis combining low-coverage whole-genome sequencing data and three genome-wide-association studies to identify genetic modifiers of age at IBD diagnosis. Four loci were discovered associated at suggestive significance with Crohn's disease (CD), ulcerative colitis or both, one of which may have a pleiotropic effect, being associated with both the risk of CD and a decrease age at CD diagnosis.

Finally, in Chapter 6, I highlight the major lessons learnt with these projects, discuss some immediate impact some of these results had for patients, and look forward to the future NGS-based studies that will shape gene-mapping strategies over the next coming years.

Acknowledgements

I owe my gratitude to a great many people who helped me through this PhD journey. First and foremost, I thank my primary supervisor, Dr Carl Anderson, for giving me the opportunity to pursue this PhD and for his faith in me throughout these years. I remember knocking on your office door four years ago and asking for some sequencing data I could get to grips with – I was keen to learn more about genetics with you, and I absolutely did! You taught me the importance of critical thinking and statistical rigour and you helped me grow immensely as a research scientist. I am beyond grateful for that. This dissertation would not have been possible without your support and constant optimistic encouragement, especially during tough times.

I'd also like to give a heartfelt, special thanks to my secondary supervisor and clinical collaborator, Dr Nadia Schoenmakers. Thank you for the opportunity to work with you, in an exciting and fruitful collaboration. It has been an absolute privilege. Thank you for your guidance, infectious passion for biology, and for always putting our work in a bigger and fulfilling context.

I am especially indebted to Dr Inês Barroso, my first mentor at Sanger, for the confidence she has shown in me throughout the years. Thank you for taking me as a young and naive Master's student and for turning me into an excited human geneticist wannabe one year later. I still have much to learn, but I owe my first steps to you.

I'd also like to thank my other advisers, thesis committee members and first year examiners including: Professor Krishna Chatterjee, Dr Chris Tyler-Smith, Dr Matt Hurles and Dr Jeff Barrett. I am also grateful to Dr Annabel Smith, Christina Hedberg-Delouka and the Committee of Graduate Studies for keeping the Sanger PhD programme running smoothly.

The work presented in this thesis would not have been possible without the collaboration of many colleagues at Sanger, Cambridge, Oxford and other points of the world. Thank you to Professor Krishna Chatterjee, Professor Holm Uhlig, Adeline Nicholas, Dr Tobias Schwerd, Dr Erik Schoenmakers, Martin Howard, Dr Hakan Cangul, Dr Amir Babiker,

Dr Irfan Ullah, Dr Saif Alyarubi, Dr Asma Deeb, Dr Abdelhadi Habeb, Dr Justin Davies, Philip Murray, Dr Shenoy Savitha, Mehul Dattani, Dr Ruben Willemsen, Dr Ajay Thankamony, Dr Soo-Mi Park, Dr Ahmed Massoud, Dr John Gregory, Dr Vijaya Parthiban, Dr Shane McCarthy, Nicola Corton, Tarjinder Singh, Katie De Lange, Dr Yang Luo and Daniel Rice.

Thank you also to the legion of doctors, nurses, researchers and administrators of the UK10K and the UK IBD Genetics Consortia for their efforts in bringing large cohorts of patients around the world towards the noble goal of advancing disease research. I'd also like to extend a big thank you to all the colleagues at Sanger who work in the sample management and sequencing team pipelines, as well as the many other people who keep the laboratory and computational facilities running. I have been tremendously fortunate to have had your help throughout. Special mention goes to Martin Pollard, Irina Colgiu, Allan Daly and Colin Nolan from the Human Genetics Informatics team, for going far beyond the call of duty.

I'd like to express my sincere gratitude to the Wellcome Trust for generously funding the four years of my PhD, as well as the various sponsors of the above Consortia for making these projects possible.

I am forever grateful to all patients (and their families) who participated in these research studies, without whose generosity none of this work would have been possible.

To members of the Anderson Group (both past and present) – Sun-Gou Ji, Tejas Shah, Javier Archury, Loukas Moutsianas, Jimmy Liu, Jamie Floyd, Velislava Petrova and Carmen Diaz – you became family. Thank you for patiently listening (and provoking!) all my daily rants; you really made the difference.

All the other people who have made the Genome Campus what it is for me – Joanna Kaplanis, Chris Franklin, Electra Tapanari, Pedro Albuquerque, Luis Pureza, Scott Shooter, Manuela Menchi, Ricardo Antunes, Rui Pereira, Alina Farmaki, Luis De Figueiredo, Ricardo Miragaia, Maria Fernandez, Neneh Sallah, Pinky Langat, Michal Szpak, Mia Petljak, Paris Litterick and Carol Dunbar – you brightened up my days.

Last, but certainly not least, I thank my family, who has provided unwavering support and limitless pride. Particularly: my dad for setting the bar high; my mum for reminding me that life is so much more than work; my four younger sisters Marta, Maria, Sofia and Inês for always picking up the phone; and finally my endlessly supportive soon-to-be husband Stathi, whose cooking skills really did improve throughout the course of this PhD. I love you all.

Publications

From this dissertation

Nicholas, A. K.^{*}, Serra, E. G.^{*} *et al.*, (2016). Comprehensive screening of eight known causative genes in congenital hypothyroidism with *gland-in-situ*. *The Journal of Clinical Endocrinology and Metabolism*, 101(12): 4521–4531.

Serra, E. G. *et al.* Whole-exome sequencing and genome-wide genotyping defines the genetic architecture of very-early-onset inflammatory bowel disease. *In preparation*.

Arising elsewhere

Boudellioua, I., Razali, R. B. M., Kulmanov, M., Vladimir, B. B, Serra, E. G., Schoenmakers, N., Gkoutos, G., Schofield, P. and Hoehndorf, R. (2016). Genome-scale identification of causative variants involved in human disease. *Under review*.

Luo, Y., de Lange, K. M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N. A., Lamb, C. A., McCarthy, S., Ahmad, T., Edwards, C., Serra, E. G., Hart, A., Hawkey, C., Mansfield, J. C., Mowat, C., Newman, W. G., Nichols, S., Pollard, M., Satsangi, J., Simmons, A., Tremelling, M., Uhlig, H., Wilson, D. W., Lee, J. C., Prescott, N. J., Lees, C. W., Mathew, C. G., Parkes, M., Barrett, J. C., and Anderson, C. A. (2016). Exploring the genetic architecture of inflammatory bowel disease by whole genome sequencing identifies association at *ADCY7*. *Under review*.

^{*} Jointly contributing authors.

Table of contents

List of figures	xix
List of tables	xxi
1 Introduction and historical perspective	1
1.1 The genetic architecture of disease	2
1.2 Gene-mapping in human disease	4
1.3 The start of gene-mapping: linkage analysis	4
1.4 Genome-wide association studies	7
1.5 The next-generation sequencing revolution	11
1.6 A standard NGS workflow	16
1.6.1 Sequence generation	16
1.6.2 Alignment and variant calling	18
1.6.3 Data annotation	19
1.7 NGS genetic analyses in Mendelian diseases	20
1.8 NGS genetic analyses in complex diseases	25
1.9 Outline of dissertation	28
2 NGS-based screening of known causative genes in CH with <i>gland-in-situ</i>	31
2.1 Introduction	31
2.1.1 What is congenital hypothyroidism?	31
2.1.2 The known genetics of CH with <i>gland-in-situ</i>	32
2.1.3 Previous genetic studies of CH with <i>gland-in-situ</i>	33
2.2 Aims	36
2.3 Colleagues	36
2.4 Methods	36
2.4.1 Patients	36

2.4.2	Next-generation DNA sequencing	37
2.4.3	Sequencing efficiency of WES and HiSeq-TS experiments	41
2.4.4	Variant annotation	41
2.4.5	Identifying likely damaging variants per sample	41
2.4.6	Capillary sequencing for variant validation	42
2.5	Results	43
2.5.1	Sequencing data quality	43
2.5.2	Genetic diagnostic yield	46
2.5.3	‘Solved’ families with mutations in one gene (monogenic families)	50
2.5.4	‘Solved’ families with mutations in two genes (digenic families)	56
2.5.5	‘Ambiguous’ and ‘unsolved’ families	59
2.6	Discussion	60
2.6.1	The significance of the causative variants identified	61
2.6.2	Clinical phenotypes of mutation carriers	62
2.6.3	The role of digenicity in disease development	62
2.6.4	Limitations	63
2.6.5	Future work	64
3	Exome and targeted-sequencing of families with congenital hypothyroidism	65
3.1	Introduction	65
3.1.1	Thyroid developmental defects	66
3.1.2	Arguments for a genetic involvement in TD	67
3.1.3	Genetic studies of thyroid dysgenesis	71
3.1.4	Why exome-sequence CH cases	72
3.2	Aims	72
3.3	Colleagues	73
3.4	Methods	73
3.4.1	Patients	73
3.4.2	Sequencing	75
3.4.3	Data quality control	76
3.4.4	Gene mapping within CH families	83
3.4.5	Predicting the impact of splice donor mutations	88
3.5	Results	90
3.5.1	Inherited variants in CH families	90

3.5.2	<i>De novo</i> variation in CH trios	96
3.5.3	Copy-number-variants in the WES dataset	102
3.5.4	Searching for novel genetic causes of CH across families	104
3.5.5	Searching for likely damaging variants in candidate genes	106
3.6	Discussion	110
3.6.1	A putative causative gene for CH with <i>gland-in-situ</i>	110
3.6.2	<i>De novo</i> and CNVs in TD and syndromic CH	111
3.6.3	Limitations	111
3.6.4	Future work	113
4	The genetic architecture of <i>very-early-onset</i> inflammatory bowel disease	115
4.1	Introduction	115
4.1.1	What is inflammatory bowel disease?	115
4.1.2	The genetics of IBD	116
4.1.3	Paediatric-IBD	120
4.1.4	Very-early-onset IBD	120
4.1.5	The genetics of VEO-IBD: the rare-variant hypothesis	121
4.1.6	Another hypothesis for the aetiology of VEO-IBD	125
4.2	Aims	126
4.3	Colleagues	127
4.4	Methods	127
4.4.1	Patients	127
4.4.2	Controls	128
4.4.3	Exome sequencing and variant calling	128
4.4.4	Data quality control	128
4.4.5	Annotations	134
4.4.6	Screening of IBD-like inflammatory genes	136
4.4.7	Gene-based association analysis	136
4.4.8	Geneset and pathway enrichment analysis	138
4.4.9	Calculation of polygenic risk scores in VEO-IBD cases	139
4.5	Results	144
4.5.1	Identification of causative defects in IBD-like inflammatory genes	144
4.5.2	Gene-based association analyses	149
4.5.3	Genesets and pathway enrichment analyses	152
4.5.4	The polygenic component of VEO-IBD	154

4.6	Discussion	156
4.6.1	The importance of screening IBD-like genes in VEO-IBD cohorts	156
4.6.2	The role of rare variants in VEO-IBD	157
4.6.3	VEO-IBD has a polygenic component	159
4.6.4	Limitations	161
5	A meta-analysis to map loci associated with age at IBD diagnosis	163
5.1	Introduction	163
5.1.1	The role of genetic variation in the age at IBD diagnosis	163
5.2	Aims	164
5.3	Methods	165
5.3.1	Association analyses	165
5.3.2	Meta-analysis within CD and UC studies	167
5.3.3	Meta-analysis for IBD	167
5.3.4	Post meta-analysis quality control	168
5.3.5	Power to detect previous ImmunoChip signals	168
5.4	Results	169
5.4.1	Suggestive association for the age at CD diagnosis	173
5.4.2	Suggestive associations for the age at UC diagnosis	177
5.4.3	Suggestive association for the age at IBD diagnosis	178
5.4.4	Comparison with the previous ADD ImmunoChip study	180
5.5	Discussion	182
5.5.1	The advantage of imputation	182
5.5.2	The pitfall and advantage of my genome-wide analysis	182
5.5.3	The possible pleiotropy of <i>FOSL2</i>	183
6	Conclusions and future prospects	187
6.1	Summary of my research	187
6.2	NGS: from bench to bedside	189
6.3	Common themes emerging from my research	190
6.3.1	Sample size	190
6.3.2	Phenotypic heterogeneity	192
6.3.3	Diverse ethnic origin	193
6.4	Future studies of rare and complex diseases	195
6.5	From variant discovery to disease mechanisms	197
6.6	Translation	200

6.6.1	Novel drug targets	200
6.6.2	Personalised treatments	200
6.6.3	Genetic risk prediction	201
6.7	Concluding remarks	202
References		203
Appendix A Appendix		271

List of figures

1.1	Inheritance of Mendelian and complex disorders	3
1.2	Linkage analysis within a family	5
1.3	Rate of Mendelian disease gene discovery between 1988-2012	7
1.4	Pace of GWAS publications since 2005	8
1.5	Schematic representation of a case-control GWAS study	10
1.6	Locus discovery in IBD over the past 15 years	11
1.7	High-throughput sequencing technology	12
1.8	Computational steps involved in NGS data generation	17
1.9	The functional impact of genetic variants at the protein level	20
1.10	A NGS-based study design for complex disease studies	27
2.1	Thyroid hormone synthesis	34
2.3	Sequencing efficiency per gene	44
2.4	Proportion of exons poorly covered per gene	45
2.5	Summary and distribution of mutations observed in the CH-GIS cohort	48
2.6	Causative variants identified in CH cohort with GIS	49
2.7	Mutations identified in <i>TG</i>	51
2.8	Mutations identified in <i>TPO</i>	53
2.9	Mutations identified in <i>DUOX2</i>	55
2.10	Genotype-phenotype segregation in families with oligogenic variants. . .	57
3.1	Gene expression during thyroid gland development	68
3.2	Phenotype categories within the CH cohort	74
3.3	Principal component analysis of exome and targeted-sequencing CH samples	77
3.4	Consanguinity status for exome-sequenced CH cases	79
3.5	Quality control metrics for the WES experiment	81
3.6	Population genetics metrics for the WES experiment	82

3.7	Variant filtering pipeline to identify inherited variation within CH families	85
3.8	MaxEntScan scores for the wild-type and mutant splice donor sequences of <i>TBX1</i>	92
3.9	Functional impact of inherited alleles identified in WES CH families	95
3.10	Distribution of inherited rare functional variants across CH families	96
3.11	Posterior probabilities of <i>de novo</i> events called by DeNovoGear	97
3.12	Intolerance to loss-of-function mutations for <i>HNRNPD</i> vs. the exome	101
3.13	Copy-number-variants identified in exome-sequenced CH families	103
3.14	<i>SLC26A7</i> expression in GTEx tissues	109
4.1	Epidemiological and clinical features of CD and UC	116
4.2	The genetic architecture of CD and UC	117
4.3	Overview of intestinal immunity in health and disease	119
4.4	Monogenic defects associated with VEO-IBD	124
4.5	Contamination analysis	130
4.6	Principal component analysis of VEO-IBD cases and controls	131
4.7	Mean genotype quality and mean depth per sample	132
4.8	Number of singletons per sample	133
4.9	Pre- and post- variant QC metrics for North European cases and controls	135
4.10	Per-sample heterozygosity rate at autosomal sites	141
4.11	Mutation identified in <i>XIAP</i>	146
4.12	Sequencing coverage for IBD-like genes vs. the exome	148
4.13	Burden of rare functional variants	150
4.14	Burden of rare disruptive variants at four allele frequency thresholds	151
4.15	Geneset enrichment of disruptive variants stratified by allele frequency	152
4.16	Distribution of CD-based and UC-based risk scores in VEO-IBD, CD, UC cases and control individuals	155
5.1	Distribution of age at disease diagnosis across the different studies	168
5.2	QQ plots of the individual association studies	171
5.3	QQ plots of the meta-analysis results for CD, UC and IBD	172
5.4	Regional association plot for 2p28	175
5.5	Genotype cluster plot for a directly genotyped proxy of rs2879179	176
5.6	Regional association plots for the associations with age at UC diagnosis	179
5.7	Regional association plots for the association with age at IBD diagnosis	180
5.8	Effect size estimations for <i>NOD2</i> rs5743293 across the studies	181

List of tables

1.1	Family-study designs used in NGS-based studies of Mendelian diseases	22
2.1	Known gene defects causing CH with <i>gland-in-situ</i>	35
2.2	Summary of CH samples sequenced for each NGS protocol	37
2.3	Known and novel mutations detected in the CH-GIS cohort	47
3.1	Phenotypes associated with mutations in thyroid transcription factors	69
3.2	Mouse models of thyroid dysgenesis	70
3.3	Pedigree structures available in the studied CH cohort	73
3.4	Candidate GIS and TD genes selected for targeted-sequencing	75
3.5	Pedigree segregation rules for different pedigree structures	86
3.6	Rare functional variants identified in three targeted-sequenced CH patients	90
3.7	Summary of <i>de novo</i> calls per family along each filtering step	98
3.8	<i>De novo</i> mutations identified in nine CH trios	99
3.9	Case-control analysis for recurrently mutated genes across CH families	106
3.10	List of CH candidate genes used in the candidate-gene approach	107
4.1	Biological processes involved in the pathology of IBD	118
4.2	Subgroups of paediatric-IBD according to age	120
4.3	Disease-causative genes discovered in the first WES studies of VEO-IBD	123
4.4	Genetic defects associated with IBD-inflammatory phenotypes	137
4.5	Variant subsets used in case-control enrichment tests	138
4.6	Genesets tested in case-control burden analyses	139
4.7	Predicted damaging and conserved variants identified in IBD-like inflammatory genes in VEO-IBD cases	144
4.8	Enrichment of disruptive variants in KEGG pathways at $\alpha = 0.05$	153
5.1	UKIBDGC sample breakdown per contributing study	165
5.2	Number of high-quality SNPs tested in each UKIBDGC study	166

5.3	Genetic loci associated at suggestive significance ($P_{\text{META-value}} \leq 5 \times 10^{-7}$) with age at CD, UC or IBD diagnosis	170
5.4	Power to detect previous loci associated with age at CD and UC diagnosis	181
A.1	VQSR training sets used in WES variant QC	271
A.2	Targeted-sequencing QC filters	272
A.3	Genotype and phenotype information for solved CH cases	273
A.4	Genotype and phenotype information for ambiguous and unsolved CH cases	274
A.5	List of CH candidate genes, part 1	275
A.6	List of CH candidate genes, part 2	276
A.7	Disease phenotype and therapy characteristics for VEO-IBD cohort . .	277