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

Queens' College

November 2016

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 is it 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

List of tables