# Fine-Mapping and Functional Analyses of Genetic Variants Driving Local Adaptations in Humans

University of Cambridge

Corpus Christi College

A thesis submitted for the degree of

*Doctor of Philosophy*

Michał Szpak

The Wellcome Trust Sanger Institute
Wellcome Genome Campus
Hinxton, Cambridge
CB10 1SA, UK

September 2016

*To my parents and siblings,*

*my granny and auntie AM*

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work carried out under the supervision of Prof. Chris Tyler-Smith at the Wellcome Trust Sanger Institute, while member of Corpus Christi College, University of Cambridge, and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words excluding bibliography, figures, tables, equations and appendices.

Michał Szpak
September 2016

# Acknowledgements

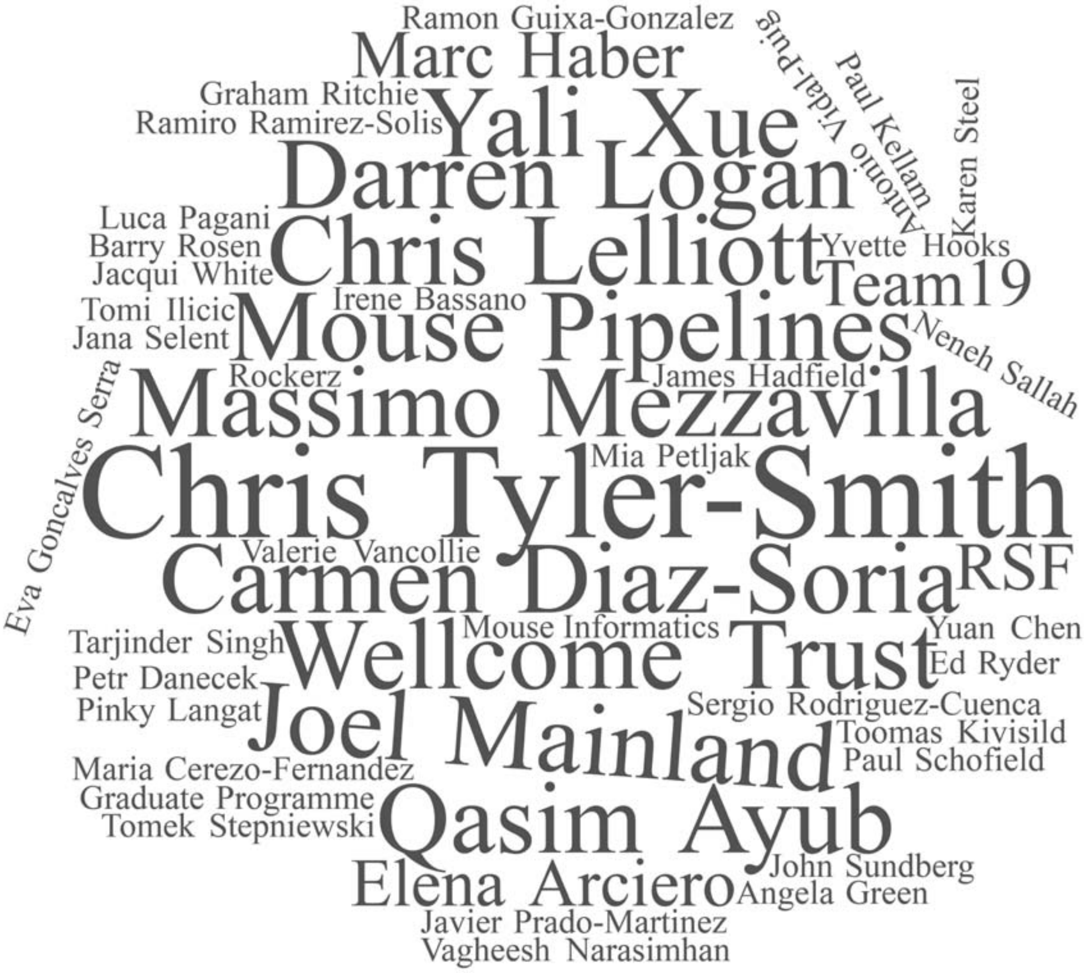

Figure 0. Acknowledgements. First and foremost, I would like to thank my supervisor Prof. Chris Tyler-Smith for giving me the opportunity to carry out this project and for all his invaluable advice, support and encouragement. Many thanks also to my secondary supervisors Dr Yali Xue and Dr Qasim Ayub for their critical and constructive assessment of my work. I would also like to thank current and former team19 members for their continual advice and guidance, especially Dr Massimo Mezzavilla, Dr Marc Haber, Dr Javier Prado-Martinez, Elena Arciero, Yuan Chen, Vagheesh Narasimhan, Dr Luca Pagani and Dr Maria Cerezo-Fernandez, who contributed to this project. I also thank my internal and external collaborators including Carmen Diaz-Soria, Dr Irene Bassano, Prof. Paul Kellam, Dr Joel Mainland, Dr Darren Logan, Dr Toomas Kivisild, Dr Petr Danecek, Dr Graham Ritchie and Lukasz Szpak for providing their expertise and fruitful collaborations. I also thank the WTSI Research Support Facility, Mouse Informatics and Mouse Pipelines, and in particular Dr Chris Lelliott, Dr Ramiro Ramirez-Solis, Dr Jacqui White, Dr Barry Rosen, Dr Ed Ryder, Yvette Hooks, Valerie Vancollie, Dr Angela Green and Hannah Wardle-Jones for productive discussions, generation and phenotyping of the mouse models and free dissemination of the data used in this project. I am also very grateful to my thesis committee members Dr Paul Schofield and Dr Darren Logan, WTSI Graduate Programme and the Committee of Graduate Studies, as well as the Wellcome Trust for my PhD studentship. I also wish to acknowledge external collaborators who joined our efforts in future development of this project, namely Prof. Karen Steel, Dr John Sundberg, Dr Sergio Rodriguez-Cuenca, Prof. Antonio Vidal-Puig, Tomek Stepniewski, Dr Ramon Guixa-Gonzalez and Dr Jana Selent. Lastly, a special thanks to my fellow PhD students, especially Tomi, Carmen, Neneh, Mia, Pinky, TJ, James and Eva, as well as the members of the 'Murray's breakfast table' and my non-Sanger friends for their support and being the best.

# Abstract

The genetic basis of human evolutionary adaptation and the resulting population diversification has been of great interest. A common approach has been to scan genomes for population-genetic signatures of positive selection, yielding vast lists of thousands of candidates. Here, we first took advantage of these data to perform a meta-analysis of published selection screens and assessed their concordance using a Selection Support Index (*SSI*) which weights, combines and evaluates signals of selection on a per-gene basis. Our analysis revealed both the low overall agreement of previous genome-wide selection scans and some strong candidates. The focus of positive selection studies in humans thus needs to move from candidate locus discovery to pinpointing underlying causal variants and further investigation of their biological significance. We developed a new computational method for this, Fine-Mapping of Adaptive Variation (*FineMAV*), which combines population differentiation, derived allele frequency and a measure of molecular functionality to prioritise candidate selected variants for functional follow-up. We calibrated and tested *FineMAV* using eight 'gold standard' examples of experimentally-validated causal variants underlying positive selection, and were able to pick out the known functional allele in all instances. We used this approach to identify the best candidate variants driving local adaptations in the 1000 Genomes Project Phase 3 SNP dataset including Africans, admixed Americans, Europeans, and East and South Asians. *FineMAV* top hits were overall enriched for high *SSI* scores, and we also report many novel examples, including rs6048066 in *TGM3* associated with curly hair and rs7547313 in *SPTA1* associated with erythrocyte shape and possibly malaria resistance in Africa, as well as rs201075024 in *PRSS53* linked to hair shape in South Asia. We extended our analyses to additional populations including Egyptians, Ethiopians, Greeks, Lebanese and non-admixed Native Americans, picking up interesting hits in Peruvian Quechua and Ethiopian Gumuz in genes involved in immunity and energy metabolism. The highest scoring *FineMAV* variant in Native Americans was rs34890031 in *LRGUK* associated with spermatogenesis. We then performed functional follow-up on chosen candidates. Our *in vitro* studies focused on comparison of the ancestral and derived forms of the

*OR10H3* olfactory receptor, and of *FUT2* involved in susceptibility to viruses, but were limited by technical issues. We also investigated the functions of six genes showing strong signals of selection using mouse knock-outs. The curly vibrissae (whiskers) of *Prss53* knock-out mice supports our hypothesis of selection in *PRSS53* due to hair shape in humans, while *Herc1* knock-out mice show a range of abnormalities affecting hearing, blood plasma chemistry and energy metabolism. Finally, we initiated the generation of nine mouse knock-ins carrying a human selected allele, which will be subjected to future collaborative phenotyping, focusing on hair shape, reproduction, energy metabolism and hearing as appropriate. Our work is thus facilitating the identification of causative alleles driving human adaptations.

# Table of contents

**LIST OF FIGURES** **17**

**LIST OF TABLES** **19**

**LIST OF EQUATIONS** **21**

**1. INTRODUCTION** **23**

**1.1. NATURAL SELECTION** **23**
1.1.1. TYPES OF NATURAL SELECTION 23
1.1.2. MODES OF POSITIVE SELECTION 26
1.1.3. HUMAN POPULATION DIVERSIFICATION 29
**1.2. METHODS TO DETECT POSITIVE SELECTION USING GENOMIC DATA** **33**
1.2.1. MACROEVOLUTIONARY RELATIVE-RATE APPROACHES 35
1.2.2. MICROEVOLUTIONARY POPULATION GENOMIC APPROACHES 37
1.2.2.1. Population differentiation-based methods 37
1.2.2.2. Site frequency spectrum-based methods 39
1.2.2.3. Linkage disequilibrium-based methods 41
1.2.2.4. Composite methods 42
1.2.2.5. Methods to detect adaptive introgression 43
**1.3. THESIS STRUCTURE** **45**

**2. FINE-MAPPING OF ADAPTIVE VARIATION *IN SILICO*** **47**

**2.1. INTRODUCTION** **47**
**2.2. META-ANALYSIS OF PREVIOUS SELECTION SCANS** **51**
2.2.1. MATERIALS AND METHODS 51
2.2.2. RESULTS 53
2.2.3. DISCUSSION 56
**2.3. FINE-MAPPING OF ADAPTIVE VARIATION** **59**
2.3.1. MATERIALS AND METHODS 59
2.3.1.1. *FineMAV* 59
2.3.1.2. Measure of population differentiation 60
2.3.1.3. Measure of allele prevalence 61
2.3.1.4. Measure of functionality 61
2.3.1.5. *FineMAV* calibration 62
2.3.1.6. *FineMAV* calculation in 1000 Genomes Project 63
2.3.1.7. Simulation analysis 65
2.3.2. RESULTS 67
2.3.2.1. *FineMAV* power analyses using simulation 67
2.3.2.2. *FineMAV* evaluation using 1000 Genomes Project 67
2.3.2.2.1. Top *FineMAV* hits classification and enrichment analysis 77
2.3.2.2.2. Functional validation *in silico* 79
2.3.2.3. Novel candidate variants across Africa, East Asia and Europe 81
2.3.2.3.1. Nonsense variants 82
2.3.2.3.2. Missense variants 83

2.3.2.3.3. Regulatory variants 84
2.3.3. DISCUSSION 86
**2.4. *FINEMAV* APPLICATION TO VARIOUS POPULATIONS 91**
2.4.1. MATERIALS AND METHODS 91
2.4.1.1. Admixed Americans and South Asians 91
2.4.1.2. Non-admixed Native Americans 91
2.4.1.3. Greeks, Lebanese, Egyptians and Ethiopians (GLEE) 92
2.4.2. RESULTS 94
2.4.2.1. *FineMAV* analysis in Native Americans and South Asians 94
2.4.2.1.1. AMR and SAS from 1000 Genomes Project 94
2.4.2.1.2. Non-admixed Native Americans 97
2.4.2.2. GLEE 100
2.4.3. DISCUSSION 106

**3. FUNCTIONAL FOLLOW-UP OF SELECTED CANDIDATES 109**

**3.1. FUNCTIONAL STUDIES *IN VITRO* 111**
3.1.1. POSITIVE SELECTION IN THE HUMAN OLFACTORY RECEPTOR GENE FAMILY 112
3.1.1.1. Introduction 112
3.1.1.2. Materials and methods 113
3.1.1.3. Results 114
3.1.1.4. Discussion 119
3.1.1.4.1. Lack of *OR10H3* activation 119
3.1.1.4.2. Ectopic expression of olfactory receptors 120
3.1.2. SELECTION ON FUCOSYLTRANSFERASE 2 (*FUT2*) 123
3.1.2.1. Introduction 123
3.1.2.2. Material and methods 125
3.1.2.2.1. Construct design with GeneArt and Site-Directed Mutagenesis 125
3.1.2.2.2. Construction of plasmids 126
3.1.2.2.3. Making lentivirus stocks 127
3.1.2.2.4. Production of stable cell lines and cell culture 128
3.1.2.2.5. Western blotting 128
3.1.2.3. Results 129
3.1.2.4. Discussion 131
**3.2. FUNCTIONAL STUDIES *IN VIVO* 135**
3.2.1. INTRODUCTION 135
3.2.2. MATERIALS AND METHODS 138
3.2.2.1. Candidate variant selection for *in vivo* studies 138
3.2.2.2. Mouse strain generation and phenotyping 139
3.2.3. RESULTS AND DISCUSSION 141
3.2.3.1. Knock-outs 141
3.2.3.1.1. Selected candidates 142
*CPSF3L* 142
*GPATCH1* 143
*LRRC36* 145
*MAGEE2* 147
*PRSS53* 147
*HERC1* 148
3.2.3.2. Knock-ins 159
3.2.3.2.1. Hair shape: *PRSS53* and *TGM3* 160

3.2.3.2.2. Reproduction: *LRGUK* and *VRK1* 161
3.2.3.2.3. Energy metabolism: *HERC1* and *CPT1A* 164
3.2.3.2.4. Hearing: *OTOF* and *PCDH15* 166

**4. GENERAL DISCUSSION 169**

**4.1. SUMMARY 169**
**4.2. NEXT STEPS 171**
**4.3. FUTURE DIRECTIONS 175**

**REFERENCES 179**

**APPENDIX A 221**

**APPENDIX B 225**

**APPENDIX C 231**

**APPENDIX D 233**

# List of figures

*Figure 1. Hard sweep model* ........ *27*
*Figure 2. Global human hair texture distribution* ........ *30*
*Figure 3. Site frequency spectrum under selection and neutrality* ........ *40*
*Figure 4. Workflow for prioritization of candidate variants for functional studies* ... *49*
*Figure 5. Meta-analysis of published genome-wide selection scans* ........ *54*
*Figure 6. Recommended minimal values of x for given n* ........ *64*
*Figure 7. Simulated distribution of FineMAV scores for variants under selection* ....... *68*
*Figure 8. Comparison of three different approaches for pinpointing selected variants in the calibration set* ........ *70*
*Figure 9. Comparison of three different approaches for pinpointing selected variants in the replication set* ........ *71*
*Figure 10. Manhattan plot of genome-wide FineMAV scores* ........ *72*
*Figure 11. Distribution of FineMAV scores* ........ *73*
*Figure 12. Derived allele frequency distribution among the top 100 FineMAV hits within each population* ........ *74*
*Figure 13. Derived allele purity distribution among the top 100 FineMAV hits within each population* ........ *75*
*Figure 14. CADD score distribution among the top 100 FineMAV hits within each population* ........ *76*
*Figure 15. Functional consequences of FineMAV top outliers as compared to random expectation* ........ *78*
*Figure 16. Distribution of FineMAV scores in SSI outlier genes* ........ *80*
*Figure 17. Genotypes of putatively introgressed SNPs identified by FineMAV* ........ *88*
*Figure 18. Manhattan plot of genome-wide FineMAV scores in Native Americans and South Asians* ........ *95*
*Figure 19. Signal of selection in the PRSS53* ........ *96*
*Figure 20. Manhattan plot of genome-wide FineMAV scores in Greeks* ........ *102*
*Figure 21. Manhattan plot of genome-wide FineMAV scores in Lebanese* ........ *103*
*Figure 22. Manhattan plot of genome-wide FineMAV scores in Egyptians and Ethiopians* ........ *104*
*Figure 23. Signal of selection in OR10H3 according to three different approaches* *115*
*Figure 24. Haplotype network of OR10H3* ........ *117*
*Figure 25. Dose-response curve of the derived (14Ile) and ancestral (14Leu) version of the OR10H3 receptor to the aurantiol odorant* ........ *118*
*Figure 26. Signal of selection in FUT2 according to three different approaches* ...... *130*
*Figure 27. Western Blot analysis cropped to show regions of interest* ........ *132*
*Figure 28. Comparison of three different approaches for pinpointing selected variants* ........ *144*
*Figure 29. Comparison of three different approaches for pinpointing selected variants* ........ *146*
*Figure 30. Comparison of three different approaches for pinpointing selected variants* ........ *149*
*Figure 31. Vibrissae shape in Prss53$^{-/-}$ mice* ........ *150*
*Figure 32. Vibrissae shape in Prss53$^{-/-}$ mice* ........ *151*

*Figure 33. Comparison of three different approaches for pinpointing selected variants ........ 153*
*Figure 34. Increased body weight in Herc1$^{-/-}$ mice ........ 154*
*Figure 35. Increased total body fat amount in Herc1$^{-/-}$ mice ........ 155*
*Figure 36. Body Composition X-ray imagining of a Herc1$^{-/-}$ mouse ........ 156*
*Figure 37. Increased circulating insulin level in Herc1$^{-/-}$ mice ........ 157*
*Figure 38. Signal of selection in LRGUK according to different approaches ........ 163*

# List of tables

*Table 1. Comparison of time-scale and modes of positive selection detected by different methods*......34
*Table 2. Selection support index values calculated for different scenarios*......52
*Table 3. List of 'gold standard' selected variants used for FineMAV calibration and replication*......62
*Table 4. Top-scoring candidates for positive selection in the olfactory receptor family*......116
*Table 5. Primers used in this study*......126
*Table 6. PCR cycling conditions for Site Directed Mutagenesis*......126
*Table 7. PCR cycling conditions for colony PCR and to generate DNA for Sanger sequencing*......127
*Table 8. List of databases used for functional annotation of candidate variants and genes*......138
*Table 9. List of mouse knock-out strains generated in this study*......141
*Table 10. List of humanized mouse strains generated in this study*......159

# List of equations

*Equation 1. Wright's fixation index ($F_{ST}$)* ........ *38*
*Equation 2. Selection Support Index* ........ *52*
*Equation 3. Fine-Mapping of Adaptive Variation* ........ *59*
*Equation 4. Derived allele purity* ........ *60*