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

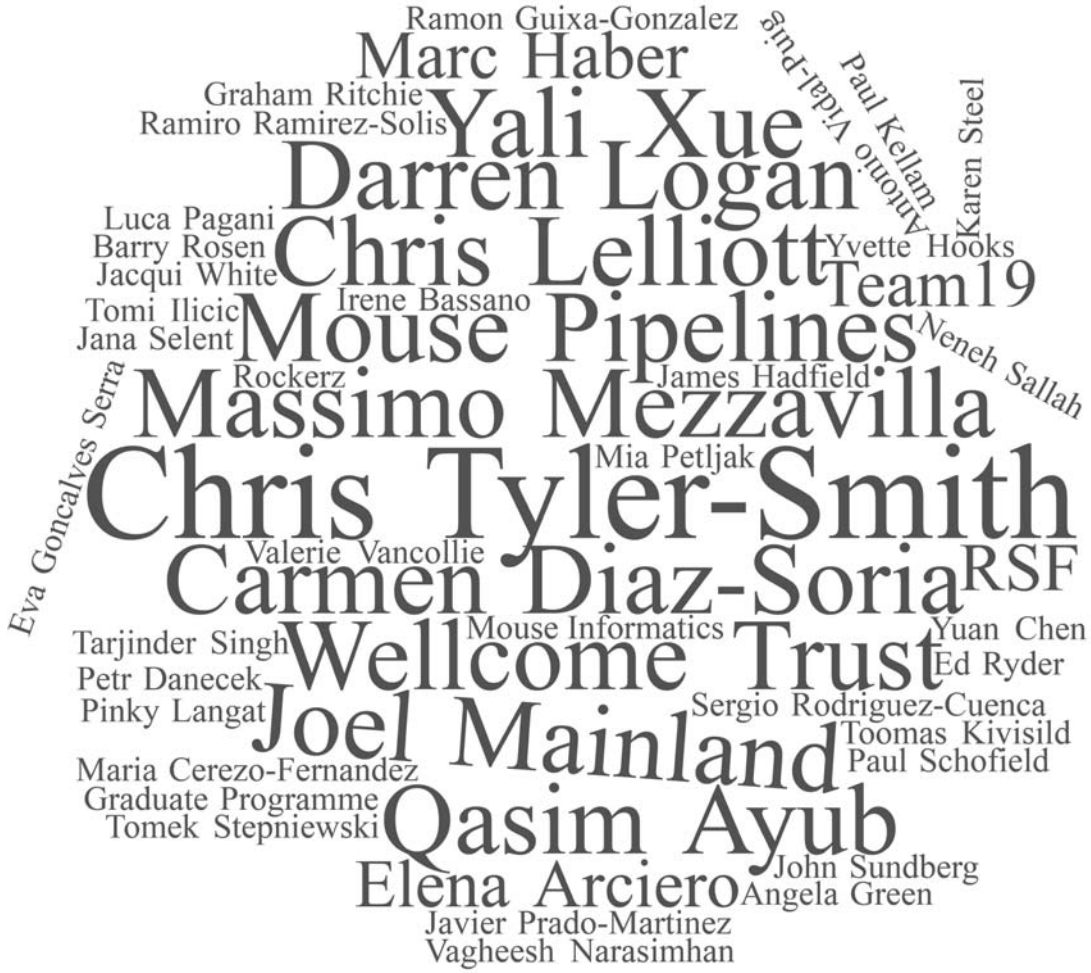

Figure 0. Acknowledgements. First and foremost, I would like to thank my supervisor Prof. Chris Tyler-Smith for giving me the opportunity to carry out this project and for all his invaluable advice, support and encouragement. Many thanks also to my secondary supervisors Dr Yali Xue and Dr Qasim Ayub 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*

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

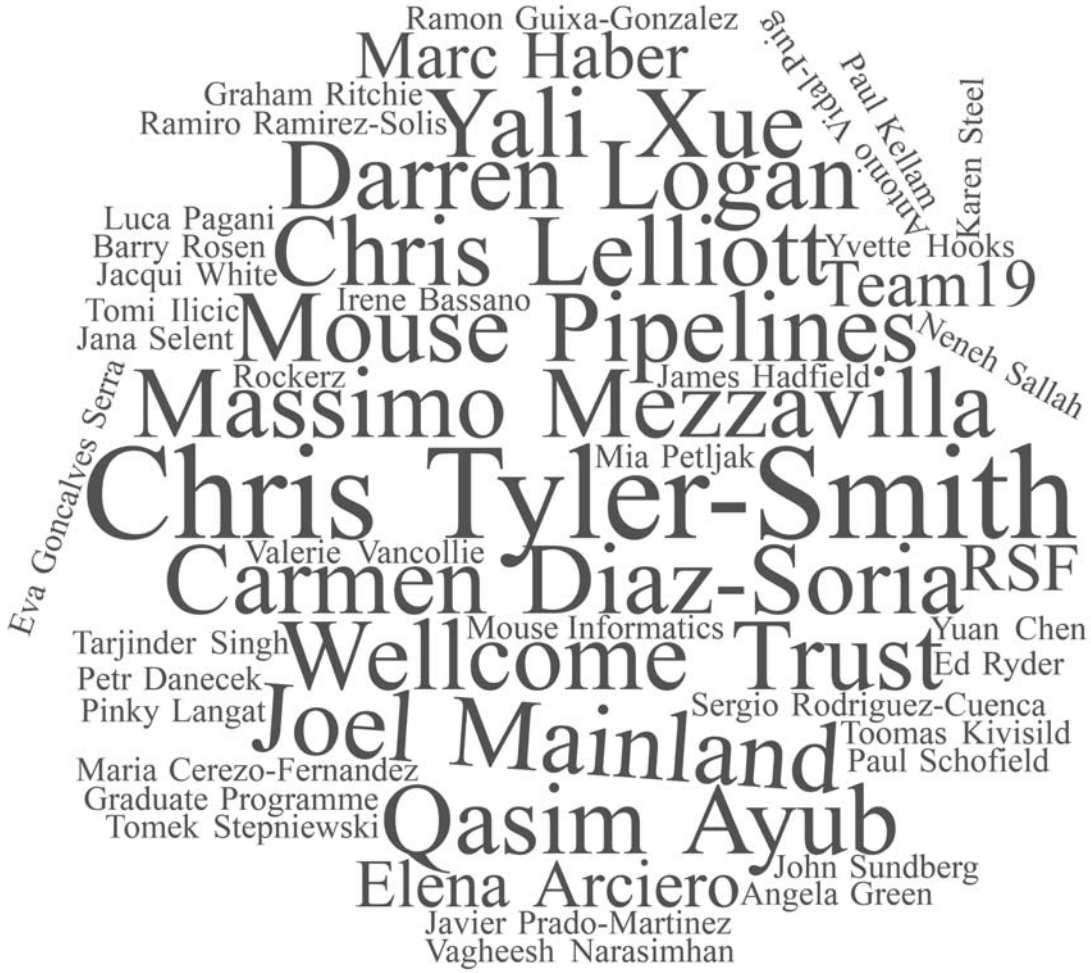

Figure 0. Acknowledgements. First and foremost, I would like to thank my supervisor Prof. Chris Tyler-Smith for giving me the opportunity to carry out this project and for all his invaluable advice, support and encouragement. Many thanks also to my secondary supervisors Dr Yali Xue and Dr Qasim Ayub 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*

1. Introduction

1.1. Natural selection

1.1.1. Types of natural selection

The theory of natural selection was introduced by Charles R. Darwin and Alfred R. Wallace in 1858, later described in detail in Darwin's book, 'On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life' (1). The process of natural selection is based on the assumption that individuals more suited to the environment are more likely to survive and reproduce, passing on their heritable traits to future generations (so-called survival of the fittest), so that the frequency of fitness-enhancing traits increases in the population over time (1, 2). Individuals with less favourable phenotypes are less likely to survive and reproduce, as all organisms are exposed to severe competition (2). According to the natural selection theory, populations change to adapt to their environments, which leads to the accumulation of variation over time and perhaps formation of a new species: a process of divergence that explains the diversity of living organisms (1). The key elements of the natural selection theory were inter-individual variation, inheritance with modification and the multiplication of new forms, although the hereditary mechanism and mutagenesis were unknown at that time (1). The advances in understanding the mechanism of inheritance made by Gregor J. Mendel and Thomas H. Morgan became the core of classical genetics and eventually enabled putting evolution to be understood in a molecular context (3). The integration of genetics with the theory of natural selection by Ronald A. Fisher, Sewall G. Wright and John B. S. Haldane formed the basis for population genetics and the modern evolutionary synthesis (4-7).

The concept of natural selection in genetics has evolved since then, and different types of selection were recognised, depending on whether an allele is advantageous or deleterious and on the fitness conferred by the genotype (8). In

this thesis we will use terminology proposed by Akey and Nielsen to describe simplified modes of natural selection (8, 9). Imagine an initial (ancestral) single allele A_1 and a new (derived) allele A_2 introduced by mutation into the population. Their possible genotype combinations are A_1A_1, A_1A_2 and A_2A_2, each with its own genotype fitness. We can represent the fitness of the new genotypes relative to the fitness of the initial ancestral genotype (equal to 1) as $1 + hs$ and $1 + s$ for A_1A_2 and A_2A_2 respectively, where h is the heterozygote effect and s is the selection coefficient (8). If $s = 0$ there is no difference in fitness between genotypes and allele frequencies are assumed to evolve naturally. Directional selection occurs if their fitnesses are not all equal (8, 9). In the case of incomplete dominance ($0 < h < 1$) and $s > 0$, the new mutation is advantageous and will rise in frequency in the population until fixation, as A_2 carriers are better adapted and favoured by the positive selection (8). If $s < 0$, then the newly occurred mutation is deleterious and will be purged from the population by purifying (or negative) selection as A_2 carriers are less fit (8). Random new mutations are more likely to be deleterious than beneficial and are constantly selected against and removed from the gene pool before achieving appreciable frequencies (a phenomenon called background selection) resulting in conserved genomic regions with little or no variation (10-12). Finally, in case of over-dominant selection acting on an advantageous allele ($s > 0$ and $h > 1$) the heterozygote has the highest relative fitness (so-called heterozygote advantage) (8). Selection of this kind is called balancing selection and multiple alleles are maintained within the gene pool (9, 10, 13). There are other types of selection defined by the phenotypic outcome rather than the underlying pattern of variability that are commonly used in the population-genetic literature e.g. diversifying and stabilizing selection. Diversifying (or disruptive) selection is described as a trend where extreme phenotypes are favoured over intermediate phenotypes; while stabilizing selection favours intermediate phenotypic values (10). Furthermore, another type of selection proposed by Darwin is referred to as sexual selection, driven by competition for mates, which explains sexually dimorphic features or increased prevalence of sexually attractive traits (1). It is important to realise that allele frequencies in a population (especially those of selectively neutral alleles that do not affect the organism's fitness) are also subjected to random fluctuation known as genetic drift (9, 14, 15). New mutations that arise in the population may increase

in prevalence due to genetic drift, even though they do not confer any selective advantage (9). Finally, selection efficiency depends critically on the effective population size, with small populations being more prone to genetic drift and thus experiencing less efficient selection (9).

1.1.2. Modes of positive selection

Here we define positive selection as any type of selection where new mutations or existing variants are advantageous (with positive selection coefficients) and there is no heterozygote advantage. There has been great interest in positive selection as it is the primary mechanism of adaptation and the evolution of novelty (9, 10). Selective episodes leave their signatures in the human genome and thus can be recognised from the pattern of nucleotide polymorphisms in a population sample due to genetic hitchhiking (16-18). The classical hard sweep model assumes that a new advantageous mutation rapidly spreading to fixation or high prevalence affects the pattern of linked variation (16, 18, 19). Its genetic characteristics include high-frequency long-range haplotypes with a concomitant reduced level of genetic variation, large allele frequency differences between populations, and changes to the allele frequency spectrum (e.g. increased fraction of derived common and rare alleles, depletion of intermediate-frequency variation) (Figure 1), although these features can also arise by genetic drift or purifying selection and are confounded by population demography (9, 16, 18-23). The size of the genomic region that is subjected to hitchhiking depends mainly on the local recombination rate and the selection coefficient, and its signature decreases with increasing distance from the selected allele (24). Ongoing, or incomplete sweep refers to any stage of selective sweep before reaching fixation, while fixed sweep is said to be complete (19).

However, it has been argued that hard sweeps were rather rare in recent human evolution and it is unusual for a new mutation to be rapidly driven to fixation (18, 20, 21). Just the opposite, it seems that most of the variants increase in frequency rather slowly and steadily, without reaching fixation and creating extensive LD patterns, because of limited dispersion over large geographic areas and low selection coefficients (18, 20, 21). Furthermore, the waiting time for new mutations can be extremely long and hard sweeps may be an inefficient response to a rapidly changing environment (18). Therefore, selection may more often operate on pre-existing variation that has evolved neutrally in the population until it becomes advantageous under certain conditions (‘selection on standing

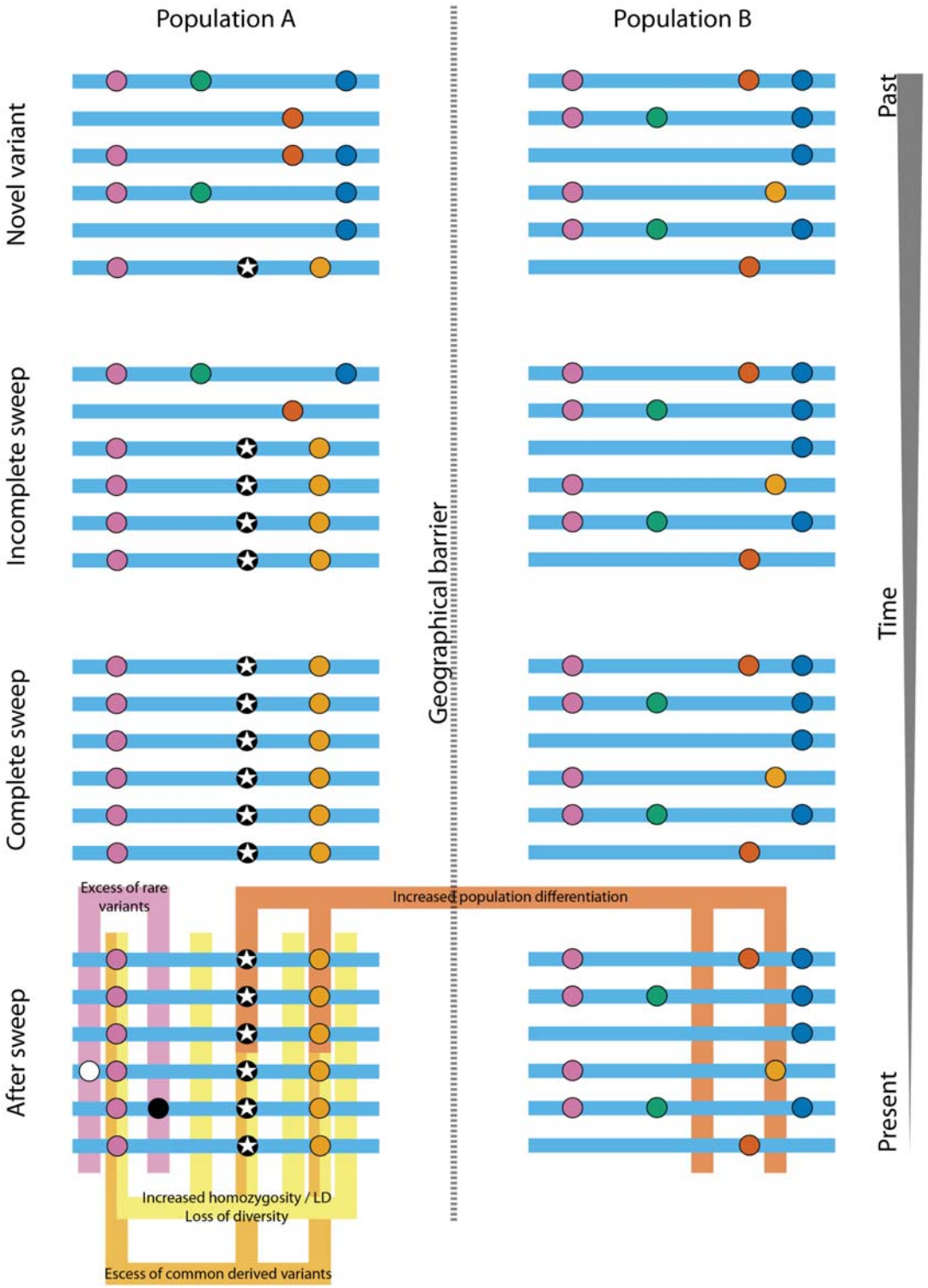

Figure 1. Hard sweep model. The blue lines indicate individual haplotypes, and derived SNP alleles are represented as circles. Population A and B with identical haplotype structure are geographically isolated with no gene flow between them. A new advantageous mutation (indicated by a star) appears *de novo* on one haplotype in population A and rapidly spreads to fixation bringing nearby linked derived alleles to high frequency. This creates a region of extended homozygosity (high LD) as there were not enough time for recombination to break it down. It also causes a population-wide reduction in genetic diversity around selected mutation as SNP-alleles that do not occur on the selected haplotype will be lost. After the sweep is complete, new mutations appear against a homogenous background creating an excess of rare alleles. Finally, differences in allele frequencies between population A and B reflects the population-specific adaptation.

variation') (18, 20, 21). Selection from standing variation is difficult to detect using most standard approaches, because the selected variant often exists on multiple haplotype backgrounds (so called 'soft sweep') and has weaker effects on closely-linked sites, so does not produce the classical selective sweep signatures of strong linkage disequilibrium (LD) and site frequency spectrum (SFS) changes, although it might exhibit an increased proportion of alleles at intermediate frequency (10, 17, 18, 20, 21, 25, 26). Similarly, the decrease in diversity around the standing variant is subtle (10, 25). Another alternative model of positive selection on standing variation is polygenic adaptation, defined as selection at many loci simultaneously affecting quantitative traits composed of hundreds of alleles of small individual effect sizes (17, 18, 21) e.g. selection on standing variants associated with greater height in Northern Europe (27-29). Polygenic selection could allow rapid adaptation (18). Signatures of selection on a complex trait are even more problematic to detect as they are composed of subtle shifts in allele frequencies at multiple loci while not producing classical sweep signatures (17, 18, 21). Finally, adaptive variation might have been acquired from archaic hominins in the process of admixture (so-called adaptive introgression), as modern humans were shown to have had limited interbreeding with archaic Eurasian hominins after out of Africa migration (30). The signature of adaptive introgression is the presence of a high frequency haplotype characterised by strong LD in a particular population that is also found in the archaic source population but is absent from populations depleted of archaic admixture (30).

It is difficult to estimate the proportions of hard sweeps, soft sweeps and polygenic adaptation, as well as adaptive introgression, in human evolution, but it seems that much of human adaptation may not have produced classical signatures of selective sweeps (18, 20).

1.1.3. Human population diversification

The out-of-Africa expansion ~60,000 years ago exposed humans to a diverse range of new environments and selective pressures including new pathogens, climatic conditions and diets (18, 23, 31). Genetic drift and local adaptations in spatially distant populations consequently led to geographically-structured genetic and phenotypic diversification, illustrated by the inter-population variation observed for numerous morphological and physiological traits, such as skin pigmentation (18, 21, 23). As gene flow between groups decreases with increasing distance, members of the same local group are usually more closely related to each other than to members of groups living in distant geographical areas (32). Traits that show extreme differentiation between populations are thus candidates for local adaptations (10). Pigmentation is not the only trait whose phenotypic values strongly associate with geography. Similar trends were proposed for hair shape (Figure 2) and body shape (e.g. larger, stockier body shape in cold climates due to thermal efficiency or the 'pygmy' phenotype in tropical rainforests) (18, 33, 34). Apart from the latitude pattern, altitude-associated adaptation has also been reported, i.e. physiological adaptations to low oxygen at high-altitudes (35, 36). We do not know, however, to what extent these phenotypic differences between populations are driven by selection. It is important to realise that genetic and morphological variation is often gradual, and phenotypic boundaries are not discrete but often show continuous clines correlated with geography (32). In addition, populations with similar physical characteristics can be genetically very different, partially due to convergent evolution (32). Finally, the genetic diversity of humans is relatively low compared with many other species (37-39) and the relationships between ethnicity, patterns of human genetic variation, and ancestry, are complicated (32, 40).

Not only are the genetic variants underlying differences between populations crucial for understanding recent human evolution and present-day human diversity, but they may also be clinically relevant, as the prevalence of some common diseases and disease susceptibilities and drug responses varies across regions e.g. the higher odds of developing hypertension in African Americans

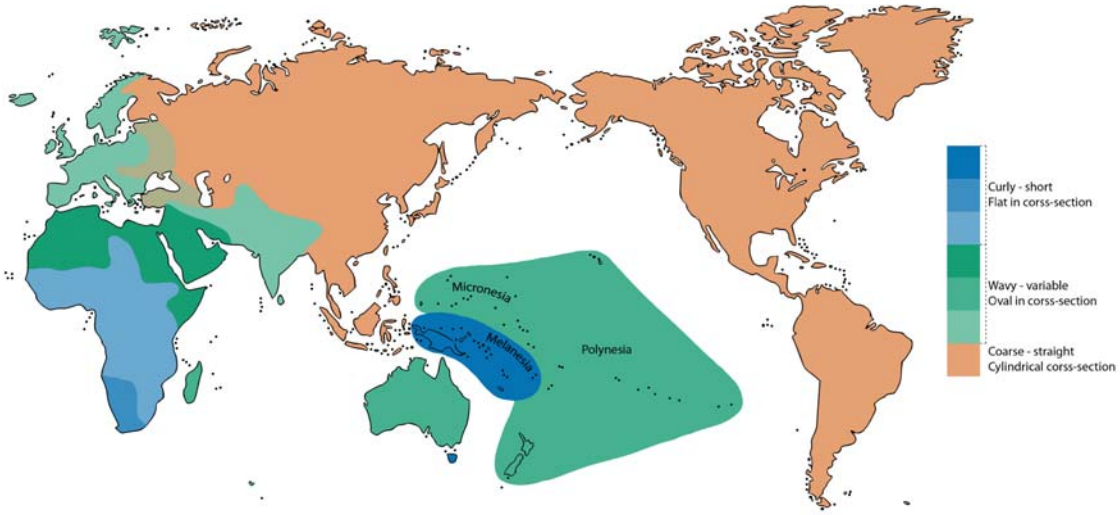

Figure 2. Global human hair texture distribution. Reproduced after (41).

compared with European Americans as a result of past positive selection favoring salt and liquid retention in hot African climate (21, 42, 43). Similarly, according to the thrifty genotype hypothesis, the high prevalence of obesity and type 2 diabetes might be a result of an adaptation to a harsh past environment and food shortages: an attractive hypothesis that nevertheless has some limited genetic support (42, 44-47). The field of evolutionary medicine (also called Darwinian medicine) argues that evolutionary biology and studies of natural selection could improve our understanding of the origins and causes of complex diseases (42, 48, 49). Medical implications of adaptive variation arise because natural selection can only act in a direct way on functionally-important variants driving phenotypic variation (14, 22); selected alleles usually confer protective effects (e.g. pathogen resistance *CASP12* (50), *CCR5* (51), *FUT2* (52) deficiency alleles), but paradoxically, may turn harmful in non-traditional environments (42, 44, 45, 53, 54) e.g. *CPT1A* (55, 56) and *APOL1* (57, 58), or in a homozygous state, e.g. sickle-cell alleles, Tay-Sachs disease, cystic fibrosis and Phenylketonuria (59-62). Regions targeted by positive selection might be disease-causing not only due to alleles that lost their advantage or balancing selection, but also through the effects of genetic hitch-hiking of moderately deleterious variants (19, 22, 42, 63). Some have argued that hitchhiking, rather than genetic drift, might be the primary force shaping the pattern of neutral variation (so-called genetic draft) (64, 65). All of the above might contribute to disease-causing mutations that segregate at relatively high frequencies (22, 66). Identification of the genetic variants that underlie regional adaptations and proving their functionality might thus sometimes facilitate disease-related research and shed more light on the diversification of modern humans and refine the human genotype-phenotype map.

1.2. Methods to detect positive selection using genomic data

Advances in genotyping and sequencing technologies laid the foundation for population genomics and enabled moving from hypothesis-driven candidate gene studies toward hypothesis-generating genome-wide screens for selection (8). Whole genome analyses provide a less biased way of searching for selection signatures, free from a priori assumptions regarding putatively selected loci (8). The common approaches applied in population genomics to distinguish neutral variation affected by genetic drift from variation subjected to selection involve:

1. i) Calculation of the summary statistic informative about selection in empirical data, ii) comparison of the results against data generated by model-based simulations of genetic drift, iii) rejection or acceptance of the null hypothesis of neutrality (10). A common problem with such an approach is that population-genetic models often make unrealistic assumptions about the demography of the populations, such as a constant population size and no population structure, and sometimes uniform distribution of recombination and mutation rates across the genome (9). Many neutrality tests have been shown to be highly sensitive to such unrealistic demographic assumptions (9, 67).
2. i) Calculation of the summary statistic informative about selection across the whole genome, ii) construction of the empirical distribution of this genome-wide statistic, iii) investigation of the top outliers in the extreme tail of the empirical distribution as selection candidates (8). Such a nonparametric outlier approach is based on the assumption that demographic history and stochastic processes affect the whole genome equally, while selection acts in a locus-specific manner placing selected targets in the extreme tails of the genome-wide distribution (8, 68). However, presence in the extreme tail of empirical distribution alone does not prove that a candidate was indeed targeted by selection, but rather that it shows unusual characteristics relative to the rest of the genome consistent with the hypothesis of selection

(8). In particular, we do not know how prevalent selection has been and what percentage of the genome should be considered an extreme outlier.

We will present the rationale behind the most commonly used approaches to detect positive selection in the following sections. An overview of the main classes of methods recovering selection events of different time-scales and modes is given in Table 1.

Table 1. Comparison of time-scale and modes of positive selection detected by different methods. 'Relative-rate' refers to comparative methods based on inter-species comparisons explained in the next section. *Diff* – population differentiation based methods; *SFS* – site frequency spectrum based methods; *LD* – linkage disequilibrium based methods; *Comp* – composite methods. + indicates that a method is sensitive to given type of selection; - indicates lack of power; (+) indicates low power. Old selection stands for species-wide adaptation that occurred during the divergence of species. Recent selection stands for recent or ongoing selection after out-of-Africa population split.

Selection time/mode		Relative rate	*Diff*	*SFS*	*LD*	*Comp*
Old selection (~200 kya – 6 Mya)		+	-	-	-	-
Recent selection (~5 – 100 kya)	Hard sweep	-	+	+	+	+
	Soft sweep	-	+	-	(+)	(+)
	Polygenic selection	-	-	-	-	-
	Adaptive introgression	-	+	(+)	(+)	(+)

1.2.1. Macroevolutionary relative-rate approaches

Relative-rate comparative methods are based on comparisons between different species and their relative rates of genetic substitution (10). Although they can in theory be applied to within-species comparisons, they are usually used to identify selective events that happened in the deep past at the macroevolutionary level, as their effect is stronger in divergence data than in polymorphism data and they are not suitable for recent selection (9, 10). Detecting positive selection using such methods involves comparison of homologous sequences across related taxa and searching for acceleration in the rate of evolution indicated by an excess of substitutions relative to the baseline mutation rate (10). Relative-rate methods have been widely used to identify genomic regions showing a significantly accelerated rate of substitution in the human lineage (69-72).

Probably the most common relative-rate method is the d_N/d_S ration, sometimes referred to as ω or K_a/K_s. This method compares the ratio of nonsynonymous mutations per nonsynonymous site to the number of synonymous mutations per nonsynonymous site in a multiple species alignment, and can be applied either to a region of interest or a single codon (9). Assuming that selection acts on nonsynonymous mutations, negative selection will reduce the number of nonsynonymous mutations, while continued positive selection will increase the number of nonsynonymous mutations, relative to the number of functionally neutral synonymous mutations that serve as a baseline substitution rate (9). In case of neutrality, synonymous and nonsynonymous substitutions should occur at the same relative rate, and therefore $d_N/d_S = 1$. If negative selection operates, $d_N/d_S < 1$, indicating a relative depletion of nonsynonymous substitutions. If region is under positive selection, then $d_N/d_S > 1$, indicating a relative excess of nonsynonymous substitutions. This method detects repeated selective fixations that occurred in the same gene or at the same site across taxa over long evolutionary time periods (9). One of its strengths is that it indicates directionality of selection (positive vs negative), but it is restricted to coding regions and nonsynonymous sites (9). Further improvements to this method were also proposed (73-77) and a similar

comparative method was designed for non-coding regions (78). The rate of substitution in noncoding regions relative to the rate of synonymous substitution in coding regions is estimated by a parameter ζ. When a site in a noncoding region is evolving neutrally ζ = 1, whereas ζ > 1 indicates positive selection, and ζ < 1 suggests negative selection (78). However, in practice this method only picks up highly variable or highly conserved noncoding sites, that may or may not be targets of selection (78).

Other relative-rate methods are based on the principle that selection modifies the levels of variability within and between-species (9). The MacDonald-Kreitman Test (*MKT*) employs between species variation ('divergence') and within-species variability ('diversity') (79). It calculates and compares two d_N/d_S values, one between species and one within species, which should be equal under neutrality (assuming constant mutation and substitution rates) (10). If one exceeds the other, then the null hypothesis can be rejected (79). Larger between-species values suggests positive selection between species, while a greater within-species ratio indicates balancing or weak negative selection within the species (10).

Similarly, the Hudson-Kreitman-Aguade (*HKA*) test compares the rate of divergence to polymorphisms for multiple genes (80). *HKA* calculates the ratios of fixed interspecific differences (*D*) to within-species polymorphisms (*P*) (10). The test assumes that under neutrality *D* and *P* should be proportional (given a constant mutation rate) and the deviation from the neutral *D/P* value allows rejection of the null hypothesis (80). The expected neutral *D/P* ratio for a lineage can be estimated by examining multiple sites (10). Relatively large *D/P* values indicate either directional selection between (accelerated speciation) or within (reduced diversity within the species) species (10). Relatively small values suggest balancing selection between species (10). In contrast to d_N/d_S based methods, it can be applied to both coding and non-coding regions (10).

1.2.2. Microevolutionary population genomic approaches

Population genomic approaches aim to detect microevolutionary selective events using within-species polymorphisms (9). They have been widely applied to uncover recent and ongoing selection underlying local adaptations in humans following the out-of-Africa migration (10). Most methods of this class rely upon the classical hard sweep model assumptions, and its effects on patterns of linked neutral variation (8, 18-20). Here, we present the basic strategies and some of their derivatives commonly used in the field of human population genomics that can be classified based on the type of selective signature they detect, as proposed by Vitti *et al*. (10).

1.2.2.1. Population differentiation-based methods

Local adaptation is manifested by a geographic gradient in the frequency of the selected allele within a geographical region (21). Any selection event, regardless of its mode, will eventually produce an excess of allele frequency differentiation between populations as long as (i) it has taken place in one population but not in another (and the allele was at low frequency when first favored) or/and (ii) there is variation in selection coefficient over space, (iii) migration and gene flow between the populations have been restricted, (iv) and there has been enough time for selection to act (20, 21). Even if an allele is equally advantageous in all environments, but its selection happened in a regionally-restricted manner, the selected variant will be concentrated around its geographic origin due to limited dispersal (21, 22). Therefore, larger than average allele frequency differences between populations may indicate local adaptation. This measure of selection is sensitive to many types of selection including classic sweeps, selection from standing variation and negative selection (17, 20, 21, 25, 26).

Using population differentiation as an indicator of geographically restricted positive selection was originally proposed by Cavalli-Sforza in 1966 (68). The first attempt to implement a population differentiation-based statistical test for positive selection was made by Lewontin and Krakauer, who used the variance of the F_{ST} parameter (Wright's fixation index) and proposed rejection of the neutral model for loci with F_{ST} values larger than expected by chance (81). F_{ST} is a common measure of population differentiation, defined as variance in the allele frequency between different subpopulations (weighted by the sizes of the subpopulations) divided by the variance of allele frequency in the total population (subpopulations combined) (82). Its values range from 0 to 1 (82). A zero value implies no population structure (a panmictic population), while a value of one implies no gene-flow between two populations (a fixed difference) (82). In practice, various F_{ST} estimators have been proposed, e.g. calculated using nucleotide diversity (π) or heterozygosity (H) (Equation 1).

Many F_{ST} derivatives have also been proposed (83-90), including the locus-specific branch length metric (*LSBL*) and the population branch statistic (*PBS*) which use pair-wise calculations of F_{ST} from three or more populations to isolate population-specific changes in allele frequency relative to a broader genetic context (91, 92), and the cross-population composite likelihood ration (*XP-CLR*) of allele frequency differentiation that extends F_{ST} to many loci (93). Some statistics explore differentiation of haplotypes instead of individual alleles (94). Another common summary of population differentiation is difference in derived allele frequency between populations *(ΔDAF*) (95). Finally, one can directly compare allele frequencies in ancient human genomes with modern samples, although the availability of well-preserved ancient DNA is limited (96, 97).

Equation 1. Wright's fixation index (F_{ST}). π_T - average number of pairwise differences between two individuals sampled from different sub-populations (nucleotide diversity within the total population). π_S - average number of pairwise differences between two individuals sampled from the same sub-population (nucleotide diversity within subpopulations). H_S - mean expected heterozygosity within subpopulations. H_T - expected heterozygosity in the entire population ($2pq$).

$$F_{ST} = \frac{\pi_T - \pi_S}{\pi_T} \qquad or \qquad F_{ST} = \frac{H_T - H_S}{H_T}$$

1.2.2.2. Site frequency spectrum-based methods

Selection is known to distort the site frequency spectrum (SFS) within a population (9). The SFS distribution of alleles in the population sample can be defined as a count of the number of mutations of a given frequency class, for each class (9). Negative selection increases the proportion of variants segregating at low frequencies in the sample (9, 10). Positive directional selection tends to increase the proportion of high frequency variants (selected alleles and linked neutral sites), but also the proportion of low frequency variants in the case of a hard selective sweep causing a population-wide reduction in the genetic diversity around the selected allele (Figure 3) (9, 10). Balancing selection increases the proportion of intermediate frequency variants (9). Such distortions can persist for thousands of generations (10).

The most common neutrality test summarizing the SFS is Tajima's *D*, comparing the average number of nucleotide differences between pairs of sequences with the total number of segregating sites in a population sample (98). If the difference between these two measures of variability is larger that expected under neutrality, then the null hypothesis is rejected. The rationale behind this method is that the low-frequency alleles contribute less to the average number of pairwise nucleotide differences (as most haplotypes in the selected region are the same or very similar, therefore there are few differences between them on average), but they do contribute to the total number of segregating sites. As a result, the excess of rare alleles drives smaller/more negative values of *D*, which might be indicative of selection (both positive or negative) or population expansion (98). Variations on this theme have been proposed with further extensions (99-101).

Another commonly used test is Fay & Wu's *H*, which compares the number of pair-wise differences between individuals to the number of individuals homozygous for the derived allele (102). As selective sweeps increase the frequency of derived alleles near the causal allele hitchhiking to high frequencies, small values of *H* indicate an excess of high-frequency derived alleles and possibly positive selection (10). Similarly, Kim and Stephan's composite likelihood ratio (CLR) test detects an excess of derived alleles across multiple sites (103).

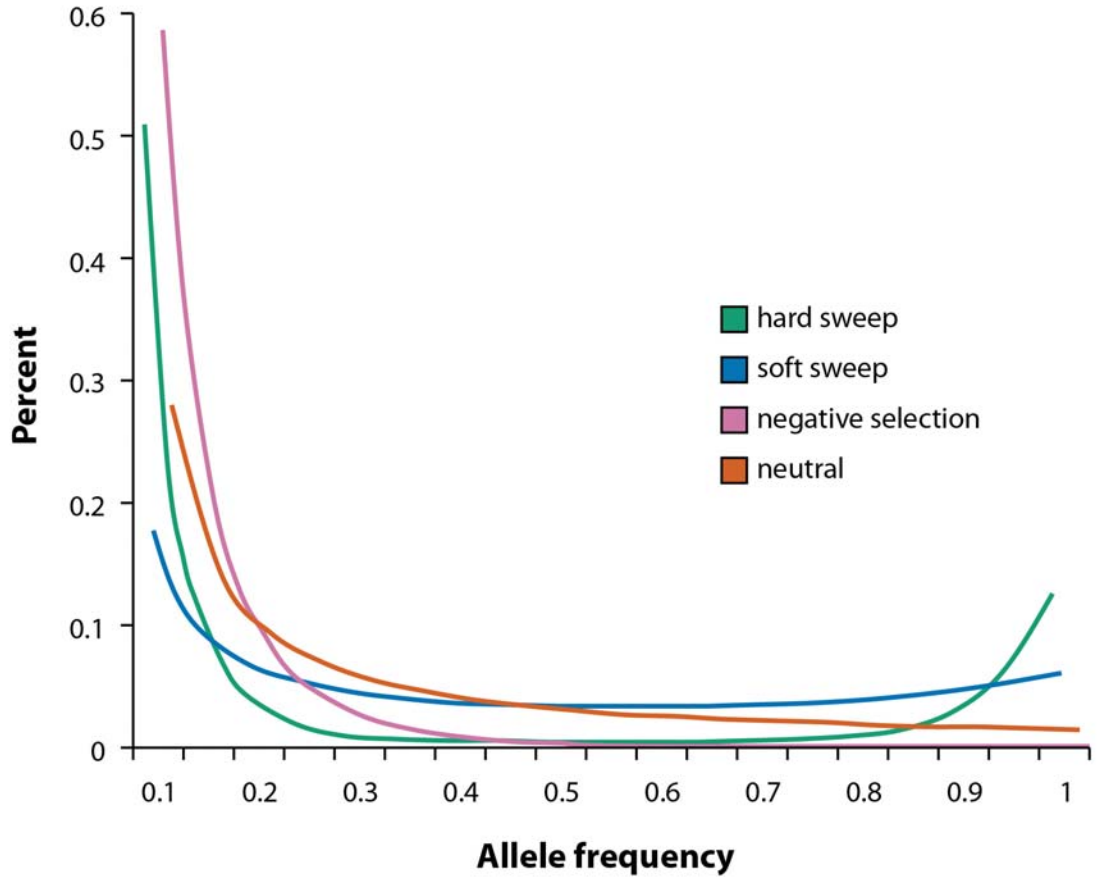

Figure 3. Site frequency spectrum under selection and neutrality.

1.2.2.3. Linkage disequilibrium-based methods

The level of linkage disequilibrium (LD) is defined as the correlation among alleles from different loci (9). A beneficial mutation spreading rapidly to a high frequency in the population brings nearby linked hitchhiker variants along, creating a region of high LD (or, equivalently, long haplotype) quickly enough that recombination has not had time to break it down (10, 16, 18, 19). Selection occurring less than 400 generations ago should leave a clear LD pattern (104). Many statistical methods for detecting regions of strong LD relative to their prevalence within a population, or a consequent reduction in haplotype diversity, have been proposed (104-107). This approach is commonly used to detect recent incomplete sweeps, but also has the potential to detect soft sweeps at lower power (10).

One class of such methods is based on the extended haplotype homozygosity (*EHH*) statistic that measures LD at a distance *x* from the core haplotype (a given haplotype at a locus of interest) (104). *EHH* is defined as the probability that two randomly chosen chromosomes carrying the core haplotype are identical by descent (homozygous at all SNPs) for the entire genomic interval from the core region to the point *x* (104). In fact, *EHH* detects the transmission of an extended haplotype without recombination (104). *EHH* ranges from 0 (no homozygosity, all extended haplotypes are different) to 1 (complete homozygosity, all extended haplotypes are the same) and decreases with increasing distance from the core region (104). Relative *EHH* is the ratio of the *EHH* on the tested core haplotype compared with the *EHH* on all other core haplotypes at the region, and ranges from 0 to infinity (104). The long-range haplotype (*LRH*) test compares a haplotype's frequency to its relative *EHH* at various distances, looking for core haplotypes that are extended as well as common, compared with other core haplotypes at the locus (104). Modified versions of this test have been proposed (108, 109), including integrated haplotype score (*iHS*) capturing extreme *EHH* for short distances and moderate *EHH* for longer distances, increasing power to detect incomplete sweeps (110), and cross-population extended haplotype homozygosity (*XP-EHH*) as well as *EHHS* comparing haplotype lengths between populations (111, 112).

LD decay (*LDD*) test is an alternative LD-based method that does not need phased data, as it operates on homozygous sites and looks for large differences in LD around the ancestral and derived alleles of a given SNP (assuming the derived allele arose on a single haplotype) (8). The fraction of inferred recombinant chromosomes (*FRC*) at polymorphisms surrounding the homozygous site (*S*) is then computed within a certain physical distance (8). Neighbouring sites are binned according to the separation distance from the site *S*. The calculated *FRC* associated with the distance from the *S* is informative about the LD decay at various distances and is compared to the genome average (8). Strong local LD around the new high-frequency allele in comparison with the alternative allele indicates a selective sweep (8).

Alternative methods detecting long identical-by-descent DNA stretches and reduced haplotype diversity or reduce heterozygosity with increased proximity to a selected mutation in a population have also been proposed (113-116).

1.2.2.4. Composite methods

Methods combining multiple complementary metrics based on distinctive signatures into one composite test might provide greater power and/or resolution in pinpointing drivers of selection (10). For instance, methods combining information from different SFS tests assessing the distribution of different frequency classes of variation (117, 118) or methods merging SFS inferences with the Ewens-Watterson homozygosity test of neutrality (*DH* test) were proposed (119-121). Nielsen *et al*. combined population differentiation-based signatures with site frequency distortion measurements (excesses of high-frequency derived alleles and low-frequency alleles) (122). The composite of multiple signals (*CMS*) combines multiple signatures of selective sweeps taking into account three features: (i) haplotype length (measured by *iHS*, *XP-EHH* and *ΔiHH*), (ii) population differentiation (F_{ST}) and (iii) differences in derived allele frequencies between populations (*ΔDAF*) (123). Finally, Pybus *et al*. applied a machine-learning hierarchical classification framework (*boosting* algorithm) that exploits scores of

11 different selection tests to classify genomic regions into specific adaptive scenarios considering selective sweeps' completeness and time-frame (124).

1.2.2.5. Methods to detect adaptive introgression

Detecting adaptive introgression from archaic humans is a two-step process comprised of detecting a signature of selection and a signature of introgression, as currently there is no method detecting both signatures jointly (30). Any previously introduced population genomic approaches could be used to detect selection on introgressed DNA (30). However, as introgression alone (not necessarily adaptive) changes the pattern of LD and distribution of allele frequencies, methods relying on LD and the SFS, as well as composite methods, can lead to false inferences of selection (30). Therefore, it seems that population differentiation methods detecting the high frequency of archaic haplotype in a specific population relative to other populations are rather robust in this scenario (assuming a low level of introgression and low starting frequency of introgressed alleles) (30).

Detection of introgressed DNA tracts requires whole-genome sequences of modern and archaic humans and is usually based on the number of uniquely shared sites (sites containing high-frequency derived alleles in a particular population, which are also present in a distantly related population but absent in other more closely related populations) (30). Such methods are based on the assumption that the gene flow from archaic to modern humans happened after the out-of-Africa migration and that non-African haplotypes shared with archaic hominins but not with present day Africans might indicate introgression (125). The most commonly used method to identify overall introgression from genome-wide data is Patterson's *D* statistic (or the so-called *ABBA-BABA* statistic) based on differential sharing of derived alleles among different pairs of individuals or populations (125-127). *D* measures the excess of shared derived alleles between each of two populations in a pair (in-group populations) and an out-group population. In a scenario of a strict population phylogenetic tree with no admixture or migration, either of the in-group populations have had any gene flow from the out-group population and each of the

two in-group populations should share approximately the same number of derived alleles with the out-group population (30). The significant deviations from the symmetrical pattern suggest introgression (30).

A challenge is to distinguish introgression from shared ancestral genetic variation (30). However, recent introgression should increase long-range LD, therefore introgressed tracts should be longer than shared ancestral variation or incomplete lineage sorting (30). The *S** statistic is a summary statistic based on patterns of LD and divergence which can be used locally to identify highly divergent haplotypes harbouring variants in strong LD shared with archaic hominins (128, 129). The expected length of the archaic tracts can be estimated using three parameters: the recombination rate, the number of migrants and the time since the admixture (130, 131).

Finally, an introgressed haplotype should have low sequence divergence from the putative archaic source population, but high sequence divergence from other present-day human individuals (30). This can be estimated by comparison of the time of the most recent common ancestor (TMRCA) of the test haplotype and the archaic haplotype with the TMRCA of the test haplotype and another modern human haplotype (132). A test human haplotype that has a recent TMRCA with an archaic population, but an ancient TMRCA with other human haplotypes is a candidate for introgression (30).

Probabilistic models based on the principles described above have also been employed to identify introgressed DNA fragments (130, 133, 134). Candidate adaptive introgressed segments are thus those that overlap between these two steps.

1.3. Thesis structure

This thesis presents results of a meta-analysis of previous selection scans, our computational approach to fine-map and prioritize candidate positively-selected variants, as well as results of functional follow-up of several candidates *in vitro* and *in vivo*. The general introduction to positive selection presented here is further extended in the following sections, each of which contains its own introduction and more specific relevant background information.

2. Fine-Mapping of adaptive variation *in silico*

2.1. Introduction

Previous surveys have reported vast lists of putatively selected genes/loci and variants, which contrasts sharply with the handful of functionally-validated examples of genetic adaptations with both a strong population selection signal and a compelling explanation for the reasons of selection linked to a relevant phenotype in humans (18, 20, 42, 135). This is partially because population-genetic based methods are often imprecise, identifying large genomic regions harboring many genes and a myriad of SNPs that could potentially drive the selection signal, but which are mostly neutral (10). Even if a selection statistic operates at the individual variant level, such as population differentiation-based statistics (e.g. F_{ST}; difference in derived allele frequency – *ΔDAF* (95)) or composite likelihood approaches (e.g. Composite of Multiple Signals – *CMS* (123)), the highest scoring variant is not necessarily causal. High LD around the selected SNP often results in a stretch of highly-differentiated variants with the same allele frequencies, further complicating the identification of the most likely causal variant. Similarly, for each potentially causal variant identified by *CMS*, there are on average 20 neutral proxies, all indistinguishable from the functional mutation (123). As a result, the false discovery rate of genome-wide selection scans is potentially high, which is reflected by the low concordance between such studies (8, 18, 20, 22, 54, 135-137).

The focus of this field now needs to move from locus discovery to fine mapping of the signals of selection and biological understanding of their adaptive significance. However, population genetics alone is usually not sufficient to narrow down the signal of selection to a single causative SNP and the only way to distinguish true positives from artifacts or neutral passenger variation is functional validation (18, 138). Yet very few variants have been validated in this way, as current technology does not allow modeling in a high-throughput fashion (138).

Therefore, a useful step is to subject candidate variants to rigorous evaluation and narrow down extensive lists to a manageable subset of the strongest candidates for functional studies.

Nevertheless, there are a few well-supported cases of local genetic adaptation that conform to a classical sweep model (20). One example is the A allele at rs1426654 (within *SLC24A5*), which is nearly fixed in European populations, causing an amino acid (Thr to Ala) change and contributing to lighter skin pigmentation (139). Melanosomal differences between ancestral and derived alleles of *SLC24A5* were successfully assayed using a zebrafish model (139). Such examples are not restricted to amino acid changes, and have also been reported for cis-regulatory variants, such as the A allele at rs4988235, an intronic regulatory variant in *MCM6* which has been shown to increase the expression of the downstream lactase (*LCT*) gene *in vitro* enabling digestion of the milk sugar, lactose, as an adult in West Asian and European populations that traditionally practice pastoralism (140, 141).

Here, we develop a new *in silico* framework to shortlist candidate selected variants for further functional follow-up (Figure 4). In order to prioritise candidate variants, we need a starting list of variants, a protocol for prioritization, and a way of assessing whether or not the prioritization is effective. Since there is a large literature on positive selection in humans, we first performed a meta-analysis of previous studies at the gene level to obtain a summary of the field, and then extended this with a new analysis of the 1000 Genomes Project Phase 3 genetic variation (142) to produce a refined list of candidate variants for functional follow up. To do so, we introduced an integrative method that overlays population signatures of selection with functional annotation, and call it *FineMAV* (Fine-Mapping of Adaptive Variation). We assessed *FineMAV* results using 'gold standard' examples (where the evidence for positive selection acting on a particular variant is convincing) and the results of the meta-analysis. After calibration and assessment of our method's performance, we applied it to diverse populations and further explored some of the novel variants in our lists.

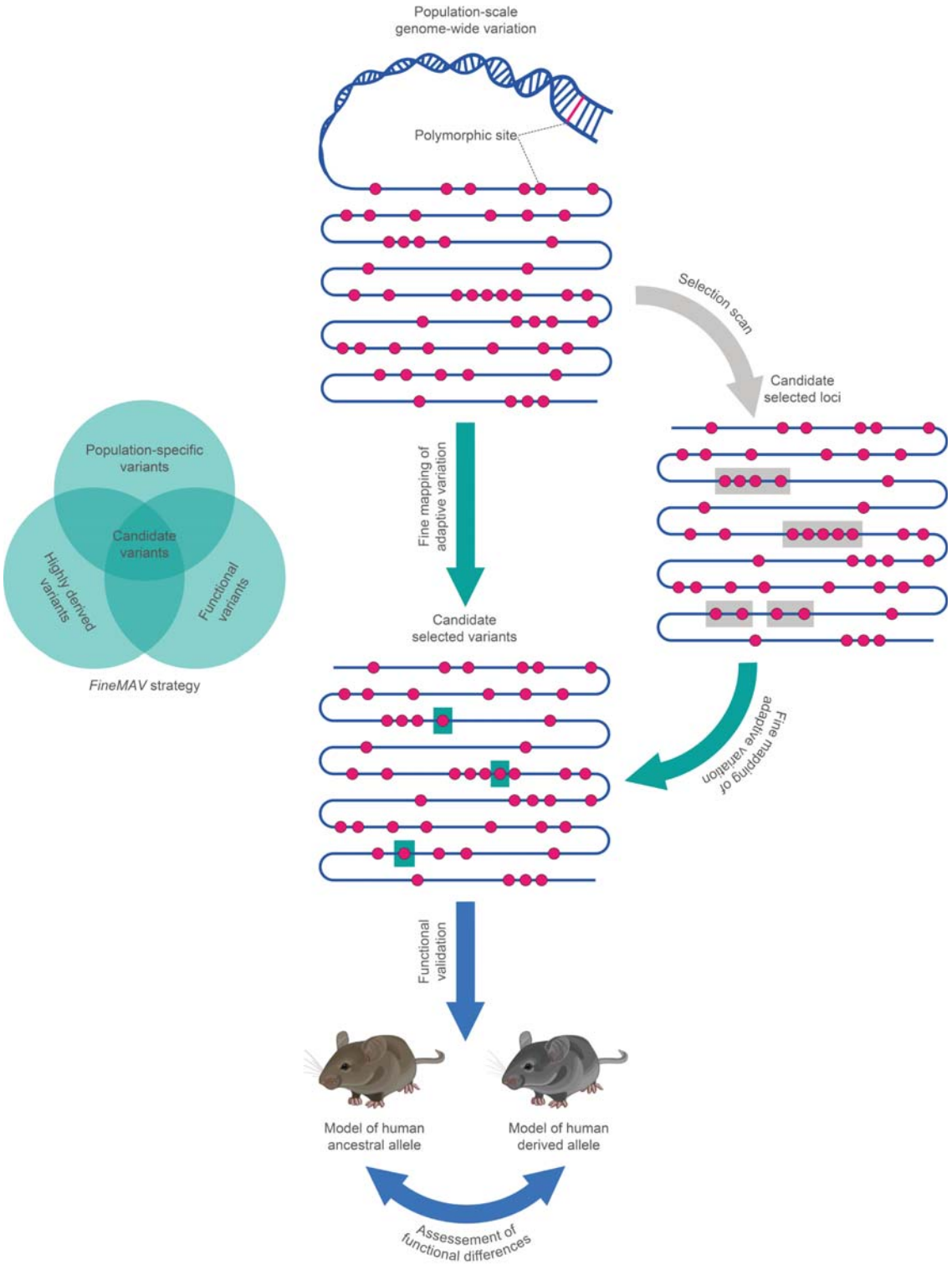

Figure 4. Workflow for prioritization of candidate variants for functional studies. The DNA molecule is represented as a blue line, with variants being red dots. Identification of the candidate causal variants from the genome-wide variation data, or the refinement of the known signal of selection to a causative SNP, is achieved by overlapping the statistical support from genetic analyses with functional annotation (implemented in *FineMAV*). A detailed follow-up functional study can then be performed (*in vitro* or *in vivo* experiments using model systems) to validate the implicated variant, quantify its phenotypic consequences and clarify its relationship with reproductive fitness, e.g. by assessment of phenotypic differences between mouse models carrying the human selected and non-selected alleles.

2.2. Meta-analysis of previous selection scans

2.2.1. Materials and Methods

We examined the concordance of all available genome-wide screens for positive selection published until September 2014, focusing on recent or ongoing positive selection, i.e. adaptations following the 'Out of Africa' dispersal that have not swept to fixation yet (incomplete sweeps or so-called microevolution). It is important to carefully curate the input data by selecting studies investigating the same mode of selection (identifying selective events of the same age and stage of selective sweep) from comparable genome-wide datasets in such an analysis (8). Therefore, we searched the PubMed publication database ('positive selection' enquiry) for studies using (i) tests based on intra-species polymorphism (excluding cross-species comparisons) and (ii) genome-wide sequencing or genotyping data (iii) across at least three main continental groups (Africans [AFR], East Asians [EAS] and Europeans [EUR]). This search yielded 26 genome-wide selection scans (83, 93, 95, 108, 110-112, 114, 123, 136, 143-158) complemented with an unpublished SFS analysis of 1000 Genomes Project Phase 1 (159). These were grouped into four methodological categories: (i) population differentiation (*Diff*), (ii) long haplotypes (*LD*), (iii) site frequency spectra (*SFS*) and (iv) composite likelihood methods (*Comp*). All reported findings were translated into gene-level nomenclature using Ensembl annotation (160). Genes reported only by a single study were excluded at this stage.

Since one particular method of looking for evidence of selection might be more abundant in the published literature than others, its results might outweigh other methods in a simple summation of the evidence and inappropriately dominate the meta-analysis. To avoid this bias and obtain a balanced view based on all four methods, we developed a correction to control for the proportion of studies that are not independent. We first calculated a per-gene selection confidence level within each methodological category (ranging from 0 for genes not reported by any

study within that category, to 1 for genes supported by all selection scans employing that detection method). We then calculated a Selection Support Index (*SSI*) by first obtaining the mean of the squares of the selection confidence levels on a per-gene basis. This would penalize genes moderately supported by several methods and promote genes strongly supported by a single approach (Equation 2). The *SSI* value was then corrected for the gene length where this strongly departed from the mean (gene length was retrieved from Ensembl (160)). The theoretical maximal *SSI* for an average-sized gene reported by all studies analysed is 1, while genes reported by all studies within one methodological category would score 0.25 (Table 2). Thus, *SSI* weighs, combines and evaluates signals of selection on a per-gene basis, starting from the results of published genome-wide selection scans of autosomal loci.

Equation 2. Selection Support Index. To compute a Selection Support Index (*SSI*) for each gene *i* with length len_i, suppose $i \in \{1, 2, \ldots, n\}$, and let $Diff_i$, LD_i, SFS_i and $Comp_i$ be its selection supports within each methodological category across all compiled genome-wide selection scans. Gene length is measured in base pairs.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} len_i$$

$$SSI_i = \frac{Diff_i^2 + LD_i^2 + SFS_i^2 + Comp_i^2}{4} \times \sqrt[10]{\frac{\mu}{len_i}}$$

Table 2. Selection support index values calculated for different scenarios. $gene_1$ – gene maximally supported by all methods; $gene_2$ – gene supported strongly by population differentiation methods only; $gene_3$ – gene moderately supported by all methods.

	Diff	*LD*	*SFS*	*Comp*	*SSI*
$gene_1$	1	1	1	1	1
$gene_2$	1	0	0	0	0.25
$gene_3$	0.25	0.25	0.25	0.25	0.0625
…					
$gene_i$	$Diff_i \in [0,1]$	$LD_i \in [0,1]$	$SFS_i \in [0,1]$	$Comp_i \in [0,1]$	$SSI_i \in [0,1]$
…					
$gene_n$					

2.2.2. Results

We assessed the confidence in selection on genes by an *in silico* quantification of the strength of the signal and its reproducibility across 27 genome-wide screens for positive selection ((83, 93, 95, 108, 110-112, 114, 123, 136, 143-158) and unpublished SFS analysis of 1000 Genomes Project Phase 1 (159)). The rationale behind integrating data from multiple sources is that the most extreme selection events should leave the strongest signals, detectable by different methods, and thus be characterised by high reproducibility across independent studies: a strong hard sweep should leave multiple signatures of selection (8). Although the ultimate goal of our analysis is to narrow down the signal of selection to a single causative variant, many selection scans identify large genomic regions and do not pinpoint a single causative SNP (10). Moreover, such scans often report outlier genes exhibiting the most extreme hallmarks of selection, instead of the precise genomic location of the signal itself. To nevertheless benefit from the rich data resource accumulated in the literature, we unified the selection-scan results by bringing them to the gene level. However, taking a simple overlap of loci reported as selected by different studies might introduce biases because the studies are not all independent. Thus, we applied a per-gene 'selection support index' (*SSI* – Equation 2) that weighs, combines and evaluates signals of selection from genome-wide selection scans focusing on recent human adaptations (adaptations that arose after the out-of-Africa population expansion) that have not swept to fixation in the species yet (incomplete sweeps or so-called microevolution).

If classic hard sweeps were frequent in human evolution, we would find many candidate genes showing multiple signatures of selection and thus scoring highly in the meta-analysis. Instead, in agreement with previous meta-analyses (8, 18, 20, 22, 54, 135-137), we found many candidate genes that were reported by only one or few studies, to which our index assigned low confidence in their selection (Figure 5.A). In contrast, some widely-accepted cases of adaptations with compelling functional evidence were found among our top-scoring candidates, such as *EDAR* (138, 161), *SLC24A5* (139), *LCT*/*MCM6* (140, 141), *HERC2* and *OCA2* (162-164). Nevertheless, even when a candidate gene has strong support from our index,

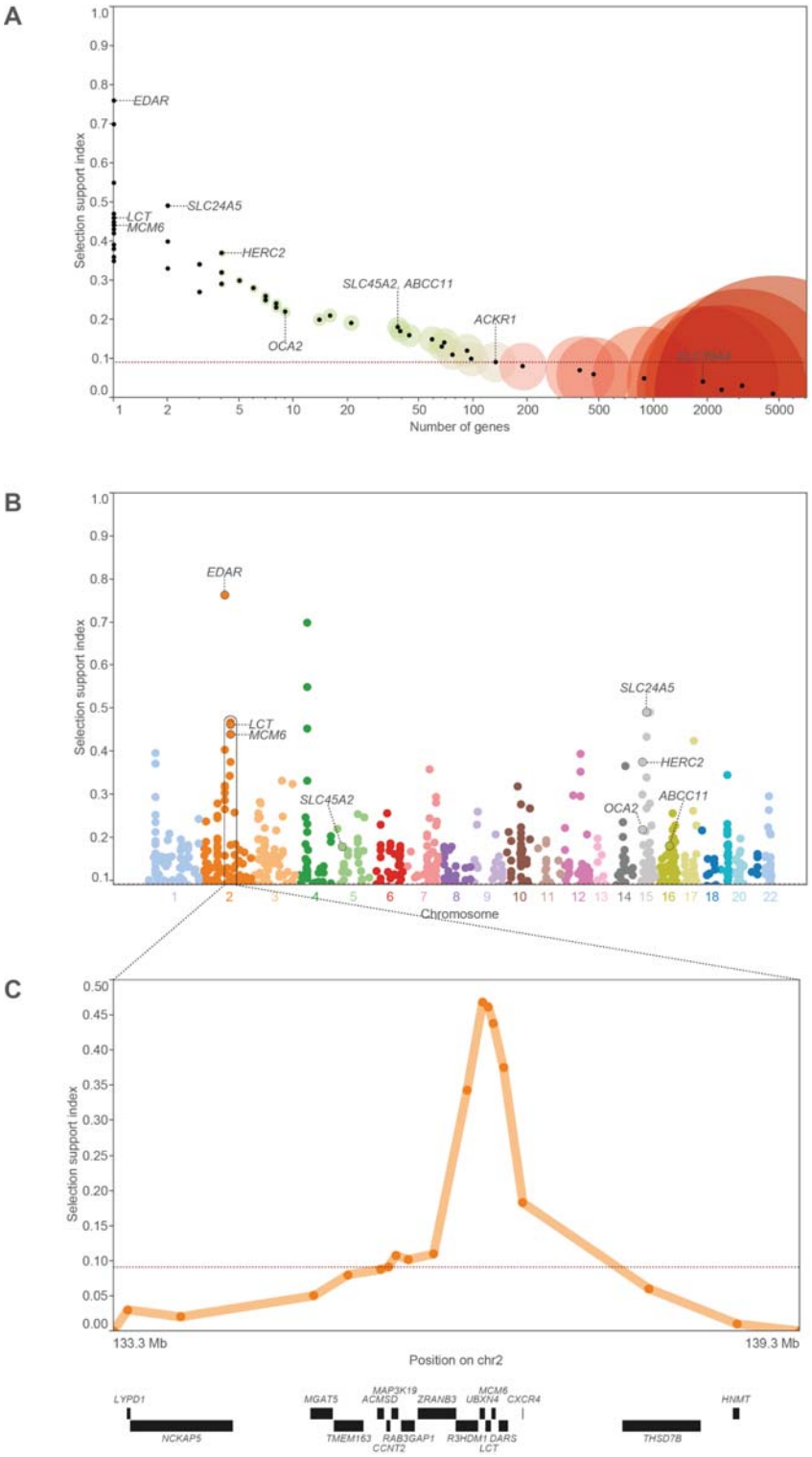

Figure 5. Meta-analysis of published genome-wide selection scans. (A) – Plot of Selection Support Index (*SSI*) scores for the positively selected genes in the published literature against the number of genes with this score; *SSI* score is also illustrated by the circle colour, and gene count by the circle size. (B) – Manhattan plot of the top ~6% putatively selected loci meeting the threshold of *SSI* score ≥ 0.09; each dot represents a gene midpoint; the cluster of genes underlying lactose tolerance is boxed. (C) – An expanded view of the lactase persistence signal showing the strong signature of positive selection that extends over a large genomic region; each dot represents the midpoint of a gene surrounding *LCT*; genes are shown as rectangular boxes in the gene track plotted below the x-axis displaying their chromosomal positions in GRCh37.

rapid hard sweeps can result in a cluster of adjacent genes scoring highly (Figure 5.B) representing a single selection event spanning up to 1 Mb (e.g. the selection signal underlying lactose tolerance in Europeans which is detectable within a 1.3 Mb window as lactase (*LCT*)-surrounding genes are often reported as extreme outliers in selection studies (Figure 5.C)). The proportion of clustered candidate genes whose selection footprint could be explained by selection acting on a nearby gene depends on the *SSI* cutoff and varies from 50% up to 70% for top candidate selected genes (meeting the threshold of ≥ 0.17 (top ~1.5%) and ≥ 0.09 (top ~6%) respectively). However, we cannot exclude the possibility that in some cases selection truly acted on more than one gene within a contiguous cluster. The list of top 7% protein coding genes and their *SSI* values can be found in Appendix A.

2.2.3. Discussion

There are many diverse approaches to search for positive selection footprints, most based on a single characteristic left by a hard sweep, although emerging composite likelihood methods combine multiple lines of evidence (8). Each method picks up on a slightly different signal and has its own strengths and weaknesses (10), thus combining several complementary methods should increase the chance of finding truly selected loci, as selected loci reported by multiple studies are more likely to be real (8).

However, previous reports should not be regarded as definitive as there are many caveats contributing to the observed low concordance between studies and clustering of candidates. Factors potentially contributing to this result include genetic hitchhiking, imprecise methods identifying large genomic chunks, the incomplete nature of the chip-genotype input data, and inconsistent criteria for reporting the most extreme outlier loci (8). Furthermore, selection studies often do not report footprints in intergenic regions, so meta-analysis is biased toward genic regions. Low overlap between previous selection studies may also indicate both differences between various methods (also recovering different selective events) and the overall high false positive rate of such scans (136).

New whole-genome sequencing datasets coupled with novel methods to detect selection can outperform previous research and detect unreported candidates (as full-sequence data ensure that all potential candidate variants are evaluated). For example, the zinc uptake transporter ZIP4, known for its striking selection signature, did not show up among the top candidate genes in the meta-analysis of the published literature (Figure 5.A). ZIP4, encoded by *SLC39A4* is characterised by an extreme difference in the frequency of leucine-to-valine substitution (Leu372Val) between West Africans and Eurasians (165). The functionality of this variant was verified through *in vitro* functional experiments demonstrating differences between the human derived and ancestral alleles in surface protein expression, intracellular levels of zinc and zinc uptake (165). However, genomic scans for selection based on extended long haplotypes or deviations in the allele frequency spectrum had failed to identify ZIP4 as a candidate

gene for positive selection. Such an extreme pattern of population differentiation and the absence of additional accompanying classic sweep signatures can be explained by the effect of a local recombination hotspot (165). In this scenario, *SLC39A4* should have obtained moderate support in our meta-analysis, but was missed in many studies employing population differentiation methods, as the selected SNP (or any SNP tagging it) was not included in the commonly-used Affymetrix and Illumina SNP arrays and consequently it was absent from the HGDP and Perlegen datasets (166, 167). As a result, *SLC39A4* was very weakly supported in our meta-analysis (Figure 5.A).

Nonetheless, even though cases that do not confirm to a classical hard sweep model could be overlooked in such gene-level overlap analysis for technical reasons, most extreme adaptive events would remain the same across different studies. However, even the strongest signals highlighted in the combined scans need to be functionally validated to be considered real. To do so, the signature of selection needs to be narrowed down to one or a few candidate SNPs.

2.3. Fine-Mapping of Adaptive Variation

2.3.1. Materials and methods

2.3.1.1. *FineMAV*

Fine-Mapping of Adaptive Variation (*FineMAV*) is designed to refine a signal of selection to a single most likely selected variant and thus to differentiate between selection-driving and passenger variants for functional follow-up studies. *FineMAV* is most relevant for targets of recent or ongoing local positive selection (within the last ~60,000 years) and can be applied to a region of prior interest, or to the whole genome for discovering novel selected variants.

A *FineMAV* score was calculated for the derived allele of each SNP by combining its Derived Allele Purity (*DAP*), continental Derived Allele Frequency (*DAF*) and functional prediction (the *CADD* PHRED-scaled C-score (168)) (Equation 3). The rationale behind doing so is that variants predicted to be non-functional are likely to be neutral, since natural selection can only act directly on variants that confer phenotypic effect. If an allele is predicted to be highly functional and rare, it is likely to be deleterious; but it cannot be harmful if it is both functional and common, and may potentially be adaptive. Importantly, all three metrics are allele-specific (rather than site- or gene-specific) and consequently allow direct evaluation of individual alleles. We simply scaled and combined the metrics to obtain a single measure giving high values to derived alleles that are common, population-specific and functional. In other words, we generate a high score for a derived allele that is common, population-specific and has a strong predicted functional effect. Individual components are introduced in the following sections.

Equation 3. Fine-Mapping of Adaptive Variation. To compute *FineMAV* per derived allele across n populations, suppose $i \in \{1, 2, \ldots, n\}$, and let DAF_i be derived allele frequency in population i.

$$FineMAV_i = DAP \times DAF_i \times CADD$$

2.3.1.2. Measure of population differentiation

We used an allele frequency differentiation method as a signature of local selection in *FineMAV*. We chose a measure of population structure differing somewhat from other methods, as: it (i) operates at the variant level, (ii) does not rely on the hard sweep assumptions of strong LD and SFS signatures (which might be erased by recombination), (iii) is sensitive to many types of selection including classic sweeps and selection from standing variation and (iv) detects recent human adaptations (17, 20, 21, 25, 26).

We proposed and applied a new measure of population differentiation called Derived Allele Purity (*DAP*). *DAP* is related to differences in derived allele frequencies (*ΔDAF* (95)) and other pairwise comparison-based methods, but able to summarise population differentiation (spatial pattern of the derived allele) across many populations in a single measure for each variant. *DAP* is a measure of derived allele entropy based on Gini impurity (169) and describes how unequally the derived allele is distributed among diverse populations. *DAP* operates on derived allele counts in a population sample when distinct groups are equally represented and is calculated according to Equation 4. When population groups are not equally represented, derived allele count can be estimated from derived allele frequency. *DAP* counts derived allele occurrences across populations and describes their spatial distribution, reaching its maximum of 1 when all cases (derived alleles) fall into a single population category, and penalizes allele sharing between different populations. The magnitude of the penalty can be controlled by the *x* parameter ('penalty parameter') depending on the user's purposes and the number of

Equation 4. Derived allele purity. To compute derived allele purity per site (*DAP*) across *n* equally represented populations, suppose $i \in \{1, 2, \ldots, n\}$, and let d_i be derived allele count in population *i*.

$$d_N = \sum_{i=1}^{n} d_i$$

$$f_i = \frac{d_i}{d_N}$$

$$DAP = \sum_{i=1}^{n} f_i^x$$

populations being compared (n). For maximally differentiated derived alleles (observed in one population only) *DAP* is constant (DAP_{max} = 1) and insensitive to n, while for the other extreme, minimally differentiated derived alleles (with the same frequency in all populations), *DAP* depends on n and $DAP_n > DAP_{n+1}$. To adjust for this, the x parameter for lower n needs to be higher. We calibrated x using a subset of our gold standards (see the following section).

2.3.1.3. Measure of allele prevalence

We estimated allele abundance using two alternative approaches: (i) global derived allele frequency and (ii) continental derived allele frequency. In both cases *DAF* ranges from 0 to 1. We obtained the continental *DAF* by averaging *DAF* across all populations within each continent, and calculated global *DAF* for each variant by averaging continental *DAF*s. Both approaches yield similar results (almost identical lists of top 100 extreme outliers). The main difference between these two measures of allele prevalence is that incorporation of global DAF results in a single *FineMAV* score for each derived allele (which is then assigned to a single population based on the difference in derived allele frequency between examined populations), while application of continental *DAF* leads to calculation of *FineMAV* scores for each population separately. Global *DAF* is n-dependent, while continental *DAF* remains constant regardless of n, thereby making *FineMAV* values comparable across different values of n. Here, we report results incorporating continental *DAF*.

2.3.1.4. Measure of functionality

It is crucial that variant-level functional inferences are based on whole-genome level measures to ensure that all potentially selected variants are treated equally. We wanted a measure of functionality to be allele-specific and applicable to all variation, both coding and non-coding, since many signals of selection localise in regulatory elements or intergenic regions (17, 123). As proteins are usually

involved in many processes through complicated interaction pathways with other proteins, amino acid change in one protein may affect many diverse traits i.e. pleiotropic phenotypes (138). In general, pleiotropic changes are thought to be disadvantageous (170), thus it is believed that a great deal of human phenotypic variation is based in regulatory variation (17, 140, 170-172). However, having a different set of annotations for coding and noncoding variation makes it challenging to compare these distinct variant categories. Thus consensus methods combining multiple annotations, each with its own weaknesses, are especially needed here for functional prioritization of variants across many functional categories (168). In our analysis we used the Combined Annotation-Dependent Depletion (*CADD* v1.2 PHRED-scaled C-score), which integrates 63 diverse genome annotations into a single measure for each variant and in theory takes a value between 0 and 99 (168).

2.3.1.5. *FineMAV* calibration

We compiled a gold standard panel of the eight best examples of experimentally-validated causal variants underlying signals of positive selection which are linked to specific phenotypic consequences (Table 3), and calibrated our method using population-scale sequence data (1000 Genomes Project (142)) of genomic windows spanning randomly chosen half of the gold standards. In the calibration stage, we needed to find the value of the *x* penalty parameter that assigns

Table 3. List of 'gold standard' selected variants used for *FineMAV* calibration and replication. 'Pop.' – population with the reported selection signal: AFR – Africans; EAS – East Asians; EUR - Europeans. 'Dataset' indicates whether given gene was used in calibration (C) or replication (R) analysis. *Note that *ACKR1* is also known as *DARC* and the derived allele at rs2814778 is the Duffy O allele.

Gene	SNP	Pop.	Function	Dataset
*ACKR1**	rs2814778	AFR	Malaria resistance(173-176)	R
SLC39A4	rs1871534	AFR	Zinc level(165)	C
ABCC11	rs17822931	EAS	Earwax and sweat type(177, 178)	C
EDAR	rs3827760	EAS	Hair shape and thickness(138, 161)	R
HERC2	rs12913832	EUR	Eye pigmentation(162-164)	R
MCM6	rs4988235	EUR	Lactose tolerance(140, 141)	C
SLC24A5	rs1426654	EUR	Skin pigmentation(139, 179)	C
SLC45A2	rs16891982	EUR	Skin pigmentation(179-181)	R

the background neutral variation and highly functional derived alleles fixed on the human lineage in the window around the selected mutation low scores. Imagine two scenarios. In scenario 1: a maximally differentiated derived allele that is exclusively fixed in population *i* but absent elsewhere (DAP_{max} = 1), which implies a maximal frequency (DAF_i = 1), and is predicted to be functional (*CADD* = 20). In this scenario, *FineMAV* = 20 and would be constant regardless of *n* (the number of populations used in the analysis). Alternatively, in scenario 2, for a derived mutation that is fixed in all populations (DAF_i = 1) and is highly functional (*CADD* = 45) we need to penalize for allele sharing between populations to keep *DAP* (and consequently *FineMAV* value) at a low level relative to scenario 1. The calibration analysis revealed that penalty parameter *x* set according to Figure 6 is sufficient to keep highly functional fixed alleles at a low level (scenario 2: *DAP* ~0.064 and *FineMAV* ~2.88, which is at least 7 times lower than the gold standard calibration set), but higher penalties might also be applied. Note that *x* decreases with increasing *n* to keep *FineMAV* value insensitive to *n*.

2.3.1.6. *FineMAV* calculation in 1000 Genomes Project

DAF and *DAP* values were calculated from the 1000 Genomes Project, Phase 3 data release (142) using a custom script; *CADD* PHRED-scaled C-scores v1.2 (168) were obtained from http://cadd.gs.washington.edu/. We ran our analysis for both autosomes and sex chromosomes focusing on three continental populations: Africans (AFR), East Asians (EAS) and Europeans (EUR). We ran it in two contexts: (i) to re-discover continent-specific positive selection signals in Africa, East Asia and Europe (*n* = 3; *x* = 3.5), and (ii) to analyze selection that happened outside of Africa by pooling East Asians and Europeans together (*n* = 2; *x* = 4.96). Even though we ran our analysis with the above continental scale configuration, *FineMAV* could also be applied to study signals of selection within continents. *FineMAV* was calculated for derived alleles (annotated accordingly to Ensembl (160, 182)) using a custom script (SNPs only; indels were omitted). We applied a conservative *FineMAV* cut-off to include only the top 100 candidate variants in each continental

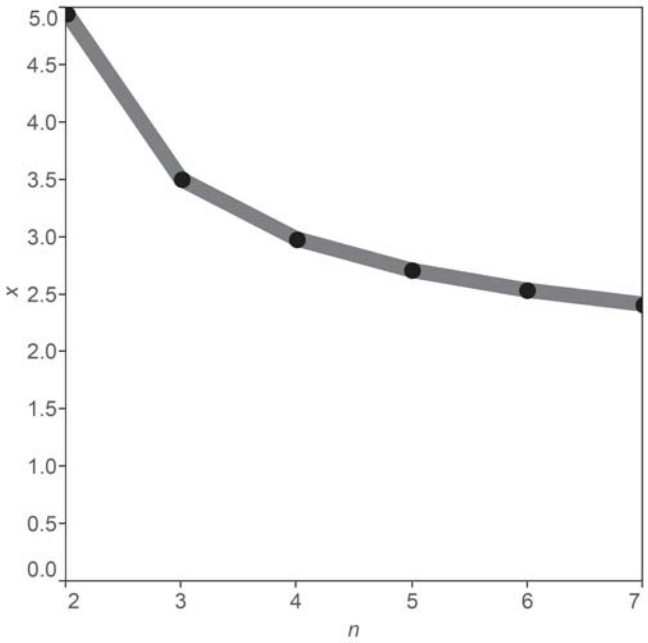

Figure 6. Recommended minimal values of x for given n. x – penalty parameter. n – number of populations being compared.

population (incorporating all gold standards and giving a total of 300 variants corresponding to the top ~0.0004% of the whole-genome distribution) for our downstream enrichment analysis.

2.3.1.7. Simulation analysis

Simulation analyses to assess *FineMAV*'s performance were limited by the unknown relationship between the prediction of functionality (*CADD* score) and the selection coefficient. Although the functional range of *CADD* scores has been estimated, its precise false discovery rate and sensitivity remain unknown, while *FineMAV*'s performance is closely tied to the accuracy of the functional annotation. Nevertheless, we performed simulation analysis using individual based forward-time simulation implemented in simuPOP v1.1.7 (183) to assess the power (True Positive Rate (TPR)) and False Discovery Rate (FDR) of the *FineMAV* algorithm. The simulation analysis was coded and run by Massimo Mezzavilla (Wellcome Trust Sanger Institute). We simulated three populations with a set of demographic parameters (starting effective population size, migration rate and time of divergence) similar to estimates in Europeans, African and East Asian populations accordingly to (184). We simulated a genomic window of 1,000 SNPs with only one SNP under selection per window in one population. The probability of recombination between two SNPs was set to increase with the increasing physical distance between sites. The starting derived allele frequency for the selected marker was set to 0.01, and the allele frequencies of the remaining neutral SNPs were drawn from a beta distribution. Each SNP was assigned a *CADD* score value as follows:

i) Neutral SNPs were randomly assigned a *CADD* score value drawn from the genome-wide *CADD* distribution of derived alleles seen at ≥2% frequency in the 1000 Genomes Project, Phase 3. Our simulation does not include a purifying selection against rare highly functional/pathogenic variants of high *CADD* prediction, therefore the derived allele frequency cutoff has been set to 2% (approximately minimal frequency at which derived allele could

be seen at least once in a homozygous state in a population of the Phase 3 size) to remove rare deleterious variants from the *CADD* distribution.

ii) We had to assume that the *CADD* distribution of selected variants is functional (which is supported by the *CADD* predictions of the gold standard panel). Based on this assumption, the *CADD* score for the selected SNP was drawn from the outlier distribution in the range of 10.78-47 (see Result section).

We then simulated 4 scenarios under the additive selection model with different selection coefficients: $s = 0.001$, $s = 0.007$, $s = 0.01$ and $s = 0$ (no selection) and a sample size of 500 individuals in each population. The populations were sampled after 1,000 generations of selection and drift. Each scenario was replicated 100 times. *FineMAV* was subsequently applied to each scenario. We then checked how often the selected variants fall outside of the neutral *FineMAV* distribution. To determine the upper end of the neutral distribution we bootstrapped 1,000 *FineMAV* values from the simulated neutral variation 100 times and took the maximum sampled value as our cut-off (set to *FineMAV* of 10.7).

2.3.2. Results

2.3.2.1. *FineMAV* power analyses using simulation

FineMAV's power to detect selected variants depends on the strength of the selection coefficient and is unable to distinguish weak selection (s = 0.001) from the neutral variation as it does not produce population differentiation (Figure 7). The medium and strong selection coefficients produce *FineMAV* distributions that are different from the neutral variation (Figure 7) and it is unlikely to find neutral variants in the extreme upper tail of the *FineMAV* distribution (assuming that *CADD* annotation is characterised by low false discovery rate). *FineMAV*'s false discovery rate in the extreme upper tail due to drift or hitchhiking is low: ~4%. The power to detect the selected variants that fall outside of the neutral *FineMAV* distribution is 46% and 77% for s = 0.007 and s = 0.01 respectively. Although the real power, which depends on the functional annotation accuracy, might be lower (as functional annotation might be incomplete), we do not attempt to pick up all selection in the genome (potentially high false negative rate), but rather to minimize the false discovery rate by using known functional annotation to identify a small number of truly selected variants for functional follow up studies.

2.3.2.2. *FineMAV* evaluation using 1000 Genomes Project

To calibrate *FineMAV* and evaluate its performance, we compiled a gold standard panel of the eight best examples of experimentally-validated causal variants underlying signals of positive selection that are linked to specific phenotypic consequences in 3 well characterised main continental populations (Table 3). We calibrated the method using genomic windows spanning half of the positive controls (randomly chosen from each population), applied it to genome-wide data from the 1000 Genomes Project (Phase 3) (142) to discover positive

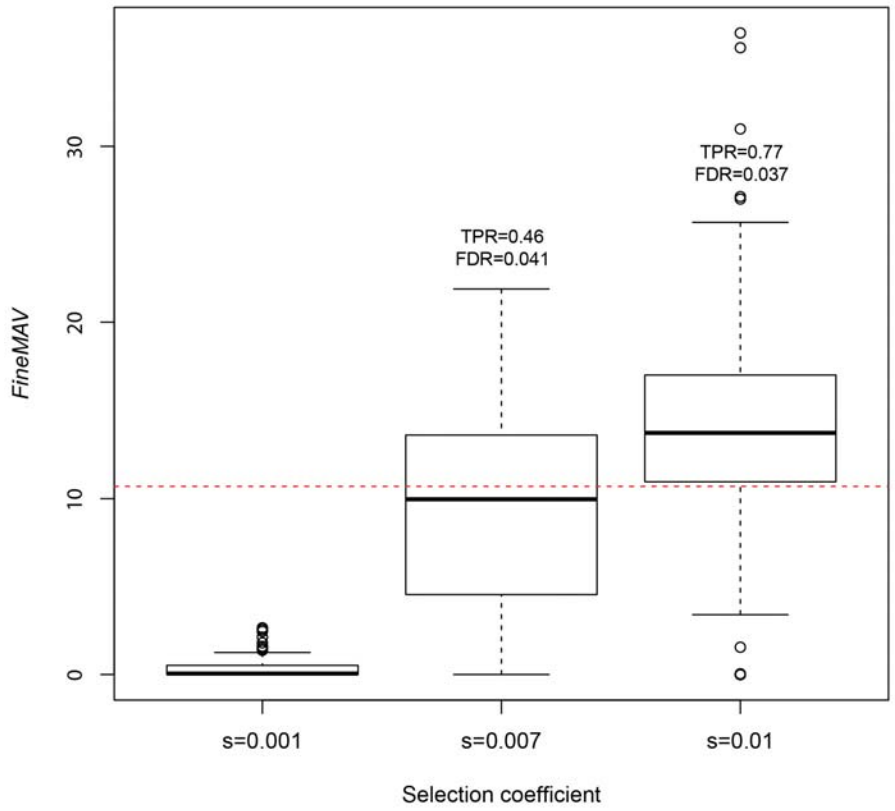

Figure 7. Simulated distribution of *FineMAV* scores for variants under selection. Three selection scenarios of varying selection strength were plotted: s = 0.001, s = 0.007 and s = 0.01. Distributions of *FineMAV* values for selected variants from each scenario are shown as box-plots. The red dotted line represents the upper end of the neutral distribution.

selection signals in Africa, East Asia and Europe, and tested the results by examining: (i) whether our method was able to separate the other half of the gold standard variants from the surrounding linked SNPs, (ii) whether the gold standards as a group were found among the extreme outliers of the genome-wide distribution, and (iii) whether *FineMAV* also enriched for genes identified in previous genome-wide selection scans with high Selection Support Index (*SSI*) values (Equation 2).

Results of the refinement of the signal of selection for the gold standard panel calibration set and replication sets are shown in Figure 8 and Figure 9 respectively, together with the performance of methods relying on population-genetic data alone (*ΔDAF* – a standard measure of population differentiation (95), and *CMS* – a composite method (123, 155)). Our integrative approach successfully distinguished the selected variants from the neutral background variation in all cases, whereas the standard methods were often unable to differentiate between the functional variant and its neutral proxies. Inclusion of functional data improved the fine mapping of truly selected variants remarkably.

We then ranked all variants based on their *FineMAV* value to identify extreme outliers in the upper tail of the empirical genome-wide distribution for each continent, and examined whether or not the gold standard variants fell in the extreme tail. We indeed found all the gold standards to be high scoring (Figure 10) (among the top 0.0004% of the whole-genome distribution (Figure 11 and Appendix B)) and set a conservative threshold to include the top 100 candidates per population (incorporating all gold standards and a total of 300 variants, out of more than 78 million derived alleles (Figure 11 and Appendix B)) for downstream analysis. Among those 300 *FineMAV* top-hits we saw variants with varying level of allele frequency (*DAF* range of ~0.25-1) and allele sharing between populations (*DAP* range of ~0.38-1), all characterised by a functional *CADD* score prediction (in the range of ~11 to 47 with a mean of ~19). It is worth noting that although *FineMAV* prioritises population-specific alleles, it also allows some degree of allele sharing between populations. The distribution of continental *DAF*, *DAP* and *CADD* in the top *FineMAV* outliers in each population are shown in Figure 12, Figure 13 and Figure 14 respectively.

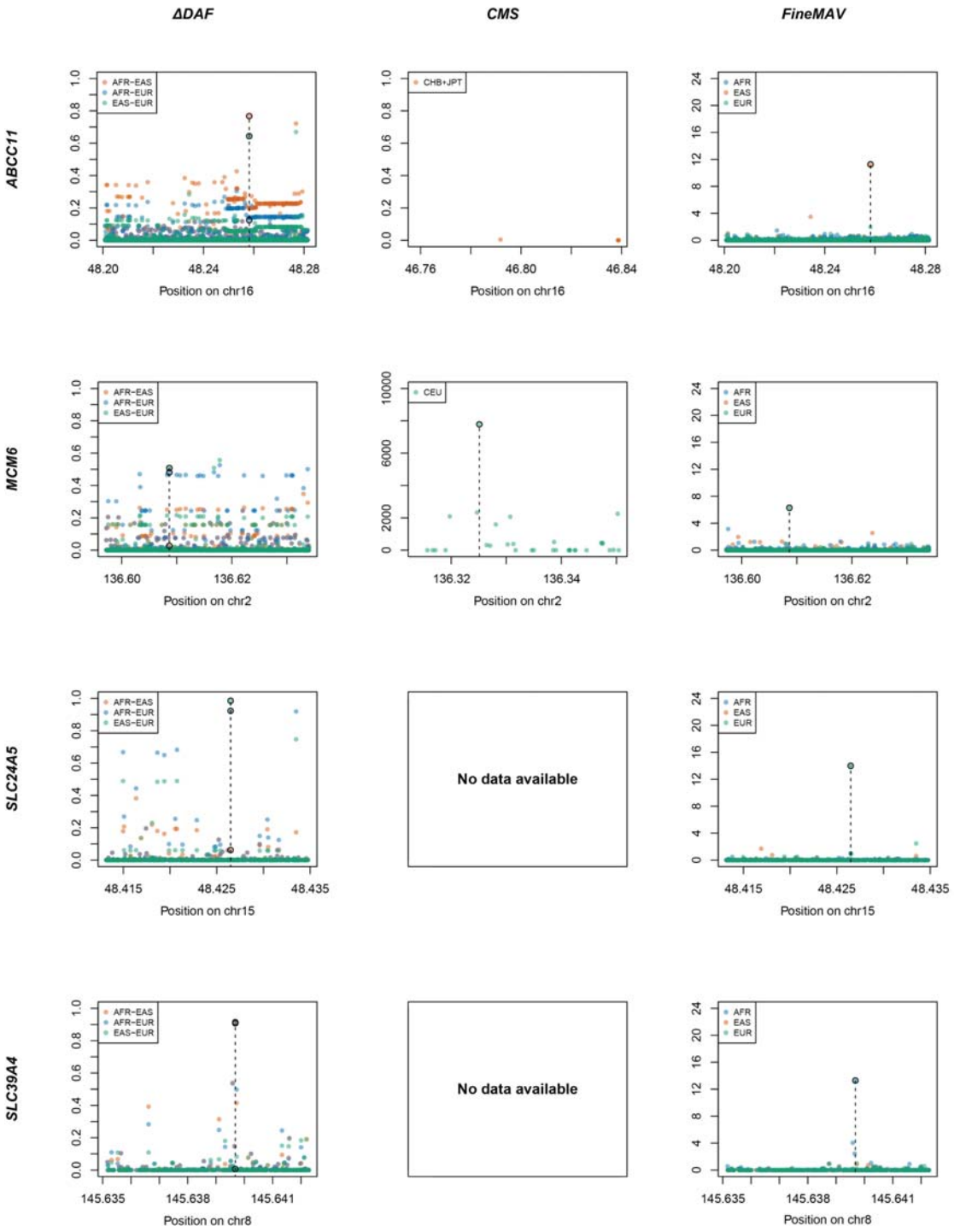

Figure 8. Comparison of three different approaches for pinpointing selected variants in the calibration set. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes from the gold standard calibration panel. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). *CMS* scores for localised regions (155) spanning genes of interest were calculated using the pilot phase of 1000 Genomes Project (185) and downloaded from http://www.broadinstitute.org/ (namely, region8new covering *MCM6*, and region152new for *ABCC11*). Variants with *CMS* value set to 'nan' were not plotted, thus there is missing variation in *CMS* plots. Genomic positions are given in Mb according to GRCh37 for *ΔDAF* and *FineMAV*, and build NCBI36 for *CMS*. The selected variant is marked with a dashed line. *FineMAV* notably reduced the noise of neutral background variation, so that the selected variant is always the highest scoring one in the given gene. Note that the y-axis scale in the *CMS* plots is not standardised.

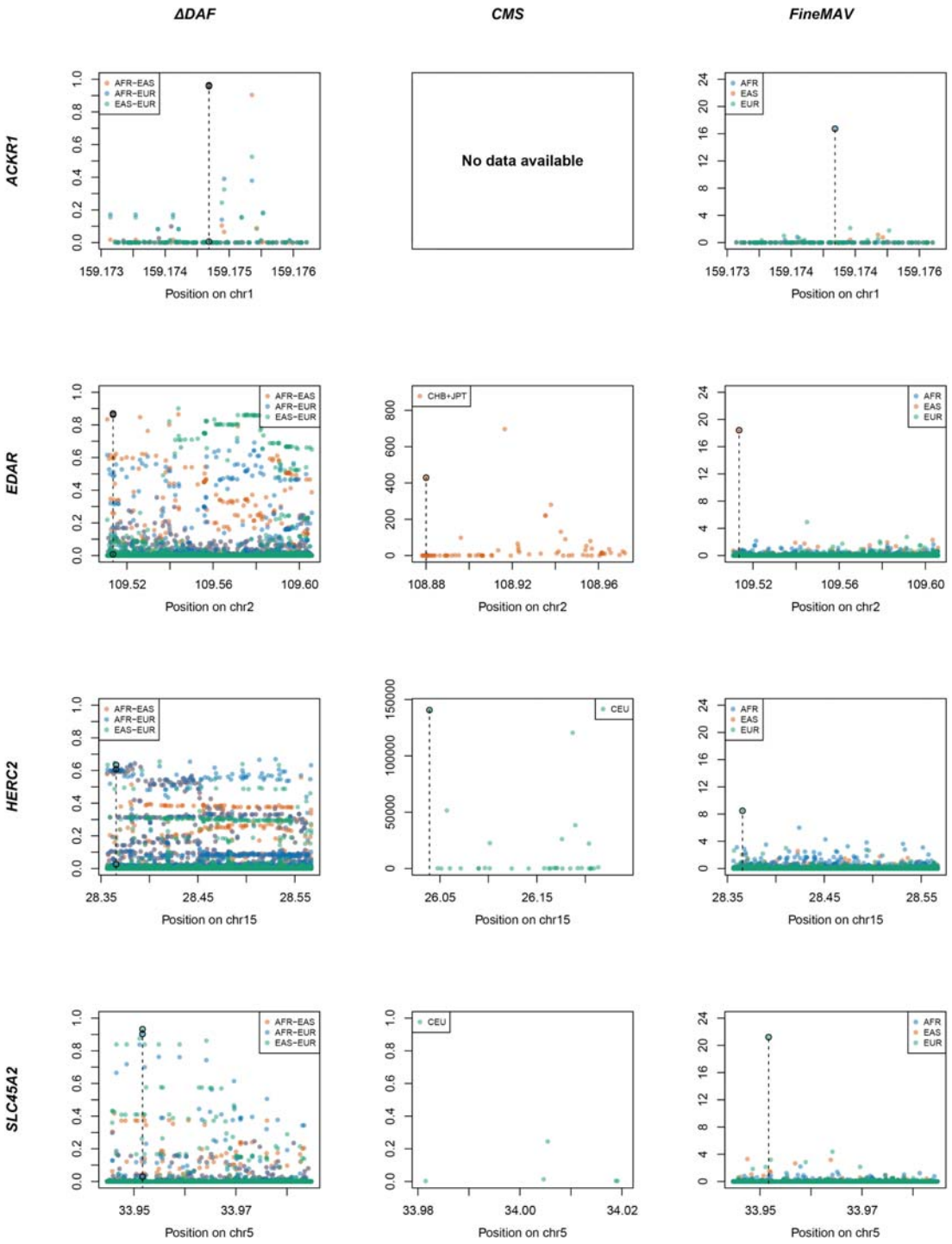

Figure 9. Comparison of three different approaches for pinpointing selected variants in the replication set. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes from the gold standard replication panel. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). *CMS* scores for localised regions (155) spanning genes of interest were calculated using the pilot phase of 1000 Genomes Project (185) and downloaded from http://www.broadinstitute.org/ (namely, region34new covering *HERC2*, region104new for *EDAR* and SLC45A2old for *SLC45A2*). Variants with *CMS* value set to 'nan' were not plotted, thus there is missing variation in *CMS* plots. Genomic positions are given in Mb according to GRCh37 for *ΔDAF* and *FineMAV*, and build NCBI36 for *CMS*. The selected variant is marked with a dashed line. *FineMAV* notably reduced the noise of neutral background variation, so that the selected variant is always the highest scoring one in the given gene. Note that the y-axis scale in the *CMS* plots is not standardised.

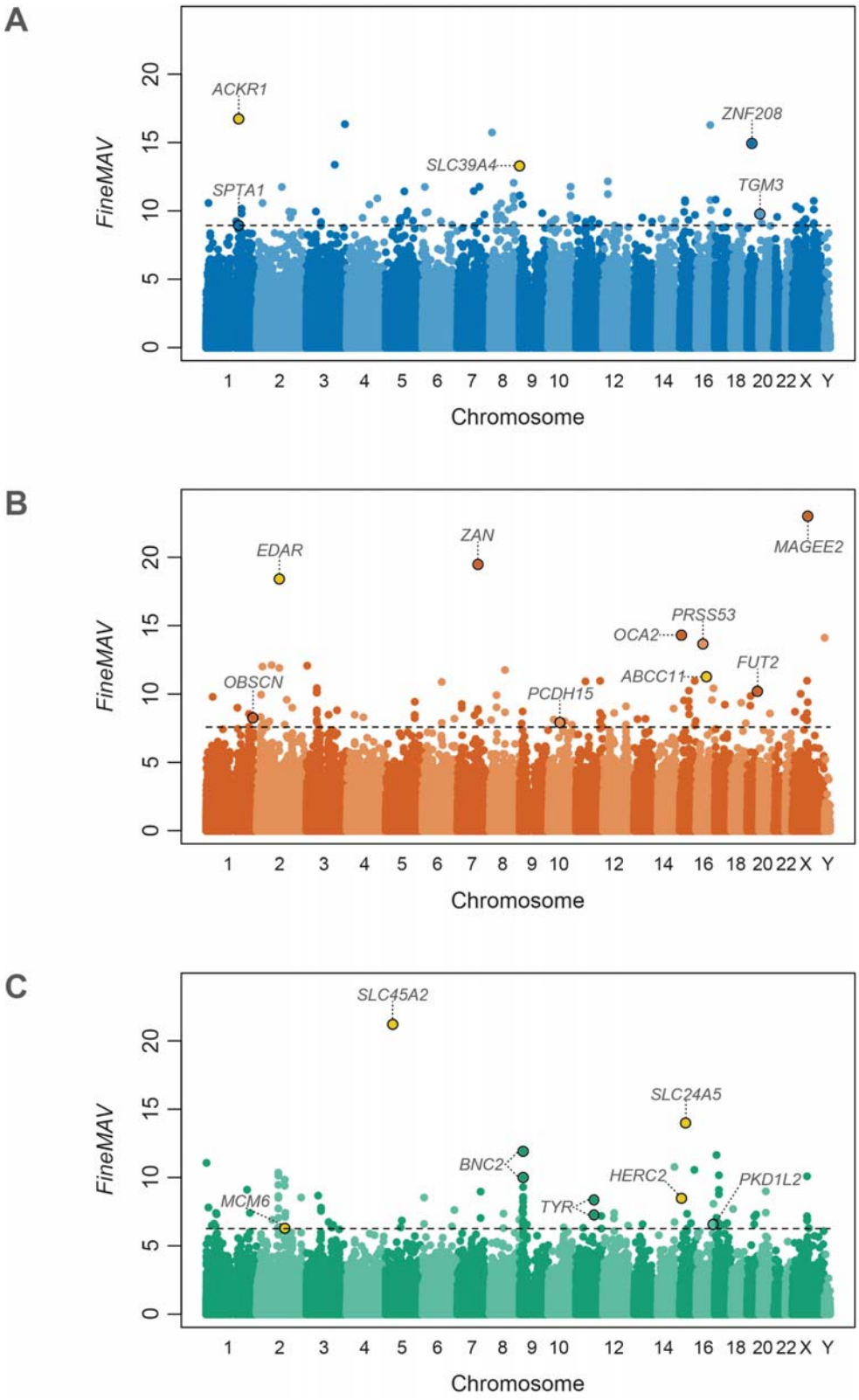

Figure 10. Manhattan plot of genome-wide *FineMAV* scores. *FineMAV* scores were calculated for genome-wide SNPs from 1000 Genomes Project Phase 3 (142) in three populations: (A) – Africans (AFR, blue); (B) – East Asians (EAS, orange); (C) – Europeans (EUR, green). Each dot in the Manhattan plots represents a single SNP plotted according to coordinates in GRCh37. The threshold (dashed lines) was set to include the top 100 variants (top ~0.0004% of the whole-genome distribution). All gold-standard SNPs (yellow dots found among the top outliers) and other interesting candidate variants are labeled with the name of the gene they fall into.

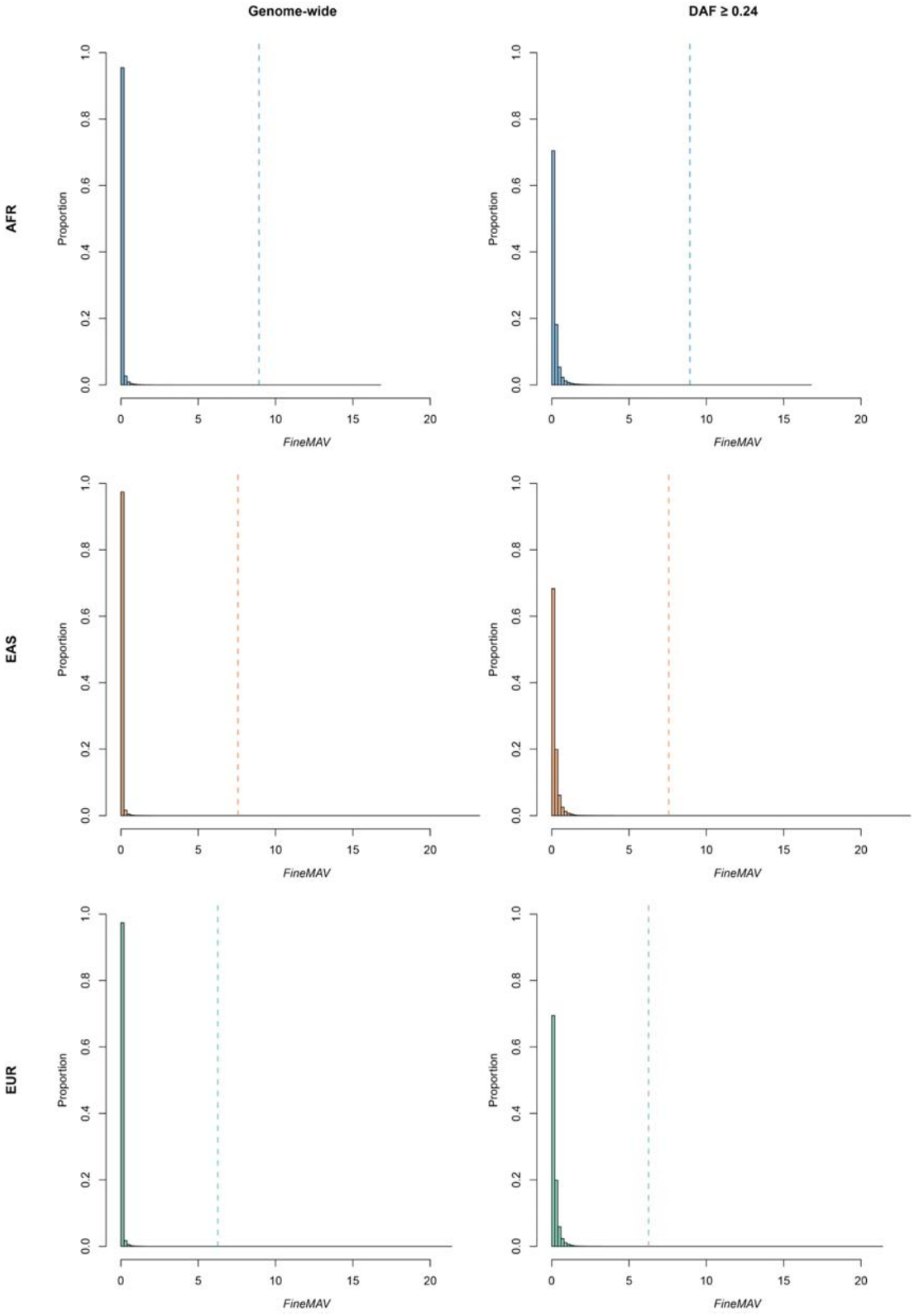

Figure 11. Distribution of *FineMAV* scores. Left: genome-wide distribution of *FineMAV* scores in each population. Right: *FineMAV* score distribution of variants matching continental derived allele frequency of our top outliers (*DAF* ≥ 0.24). Dashed vertical lines indicate *FineMAV* cutoffs to include the top 100 variants in each population. Even after accounting for *DAF*, *FineMAV* identifies extreme outliers.

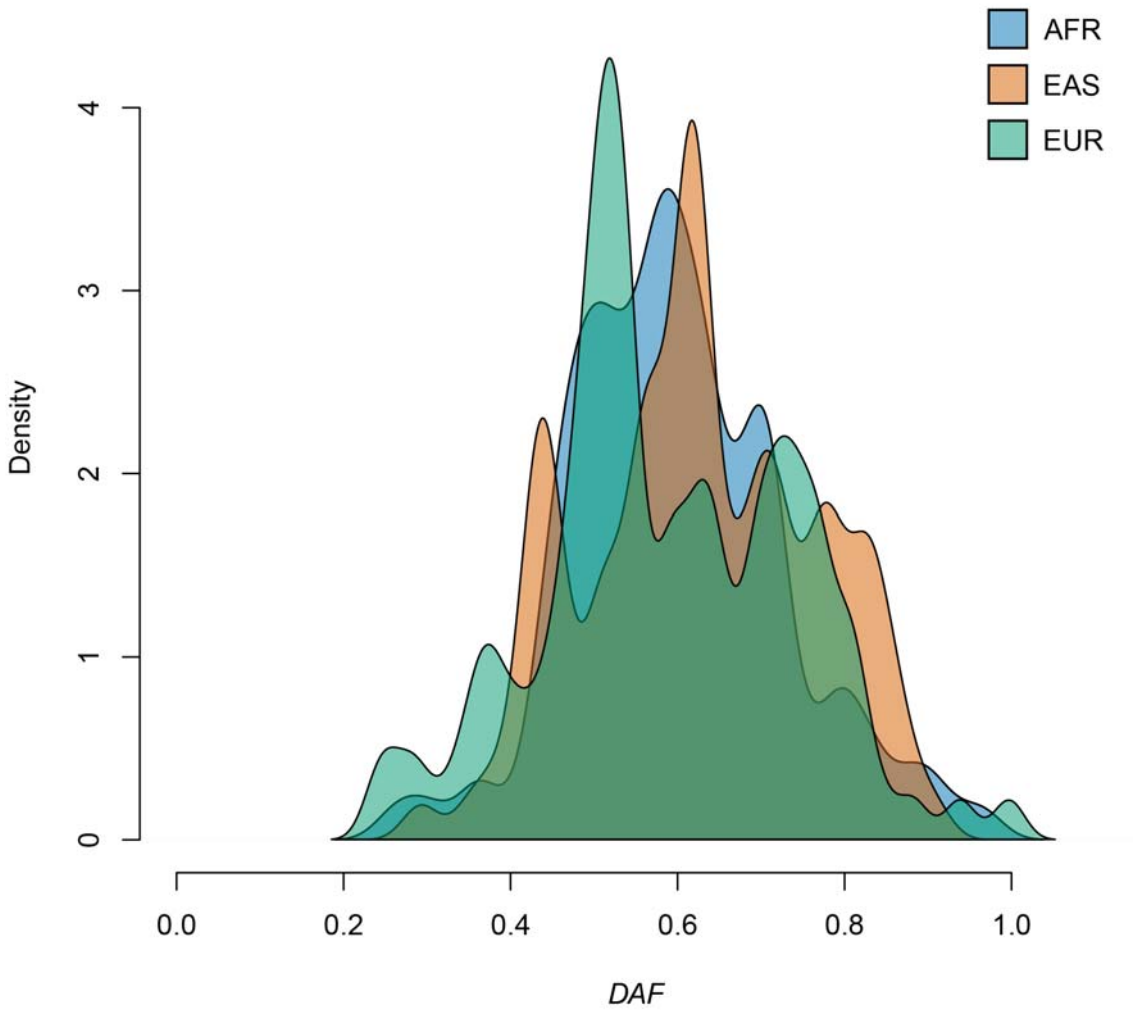

Figure 12. Derived allele frequency distribution among the top 100 *FineMAV* hits within each population.

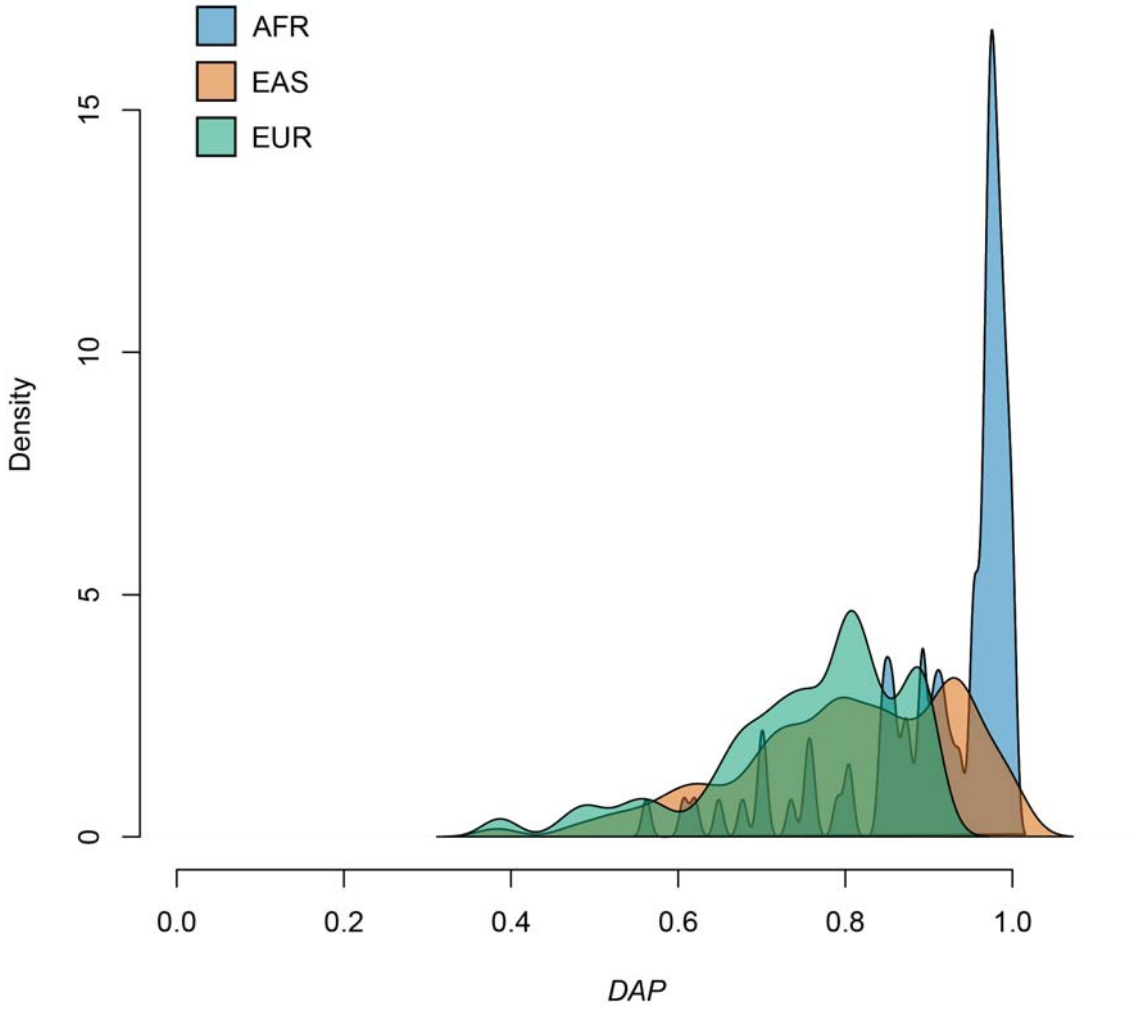

Figure 13. Derived allele purity distribution among the top 100 *FineMAV* hits within each population.

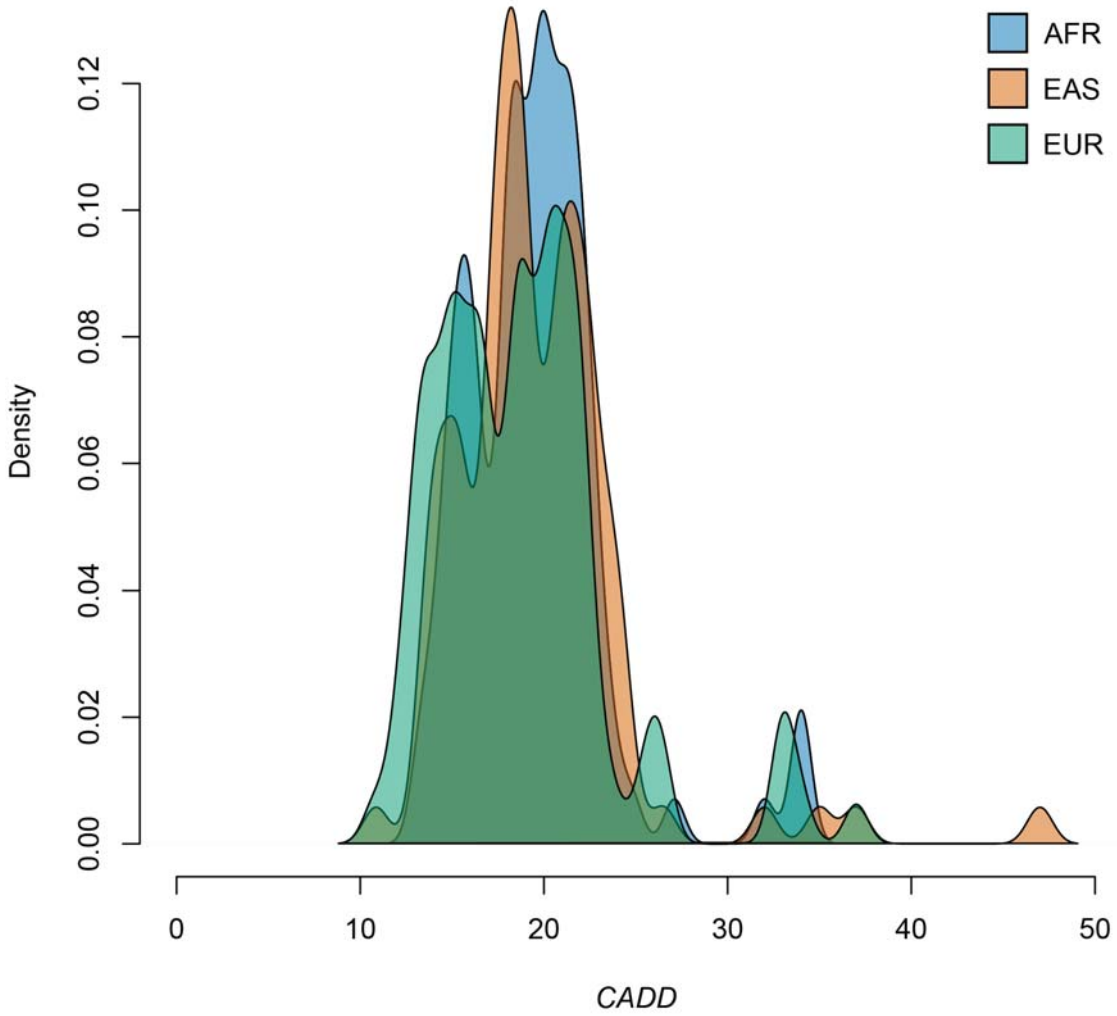

Figure 14. CADD score distribution among the top 100 *FineMAV* hits within each population.

2.3.2.2.1. Top *FineMAV* hits classification and enrichment analysis

Our list of the top 300 candidates was annotated using Ensembl (160) and we found it significantly enriched for variants of functional classes like missense mutations (p-value < 2.2×10^{-16}, Fisher's Exact Test) or regulatory region variants (p-value = 5.30×10^{-9}, Fisher's Exact Test) as compared to random expectation (list of random alleles matched for the global allele frequency) (Figure 15). This is expected because of the inclusion of the *CADD* value (168) in the *FineMAV* score.

We also used independent measures of functionality to test our results, and observed that our outliers have higher *fitCons* scores (probability that a point mutation will influence fitness) (186) (p-value < 2.2×10^{-16}, Wilcoxon rank sum test) than expected by chance. Furthermore, variants falling in broadly non-functional classes (noncoding variation) are also biased toward higher *GWAVA* scores (predicted functional impact of non-coding genetic variants) (187) as compared with random expectation (p-value < 2.2×10^{-16}, Wilcoxon rank sum test). These analyses were performed after excluding *FineMAV* hits on the sex chromosomes as *GWAVA* and *fitCons* scores are available for autosomes only (186, 187). Thus although we used one particular measure of functionality in our discovery process, we also see very strong enrichment in other available functional prediction scores, which illustrates the consistency of our results.

Finally, we used the results of the meta-analysis of previous selection scans to compare *FineMAV* top hits with previous work. Our outliers fell in or nearby genes (~200 distinct genes) significantly enriched for high *SSI* from the meta-analysis, as compared to random expectation (p-value = 6.59×10^{-10}, Wilcoxon rank sum test; after excluding gold standards: p-value = 9.20×10^{-9}). This illustrates significant concordance with previous studies, as we find our strongest signals enriched in regions that have been independently identified as being under selection, although this comparison was limited to variants falling in or near genic regions on autosomes, as previous selection scans often do not report intergenic signals and excluded the sex chromosomes. We also compared the distribution of *FineMAV* scores of top SNPs falling in *SSI* outlier genes with the null expectation. To

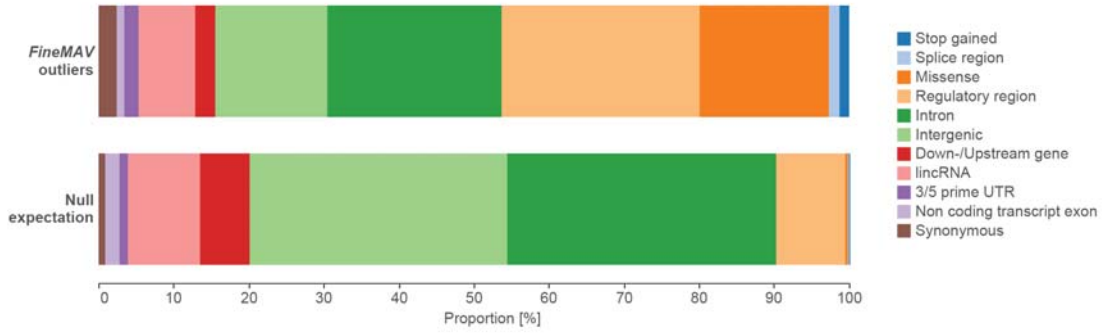

Figure 15. Functional consequences of *FineMAV* top outliers as compared to random expectation. The 100 top outliers from each population (AFR, EAS, EUR) were pooled together. The chart uses the most severe predicted consequence for each variant from Ensembl (160)

do so, we took the top ~1% of genes with the highest *SSI* scores (*SSI* ≥ 0.18), extended those genomic regions by 50 kb up- and downstream, extracted a top SNP falling in each window, and built a *FineMAV* distribution. We found this to be significantly different from the null expectation (p-value < 2.2×10^{-16}) (Figure 16).

2.3.2.2.2. Functional validation *in silico*

To further evaluate our *FineMAV* hits, we performed an *in silico* validation by searching available literature for relevant functional information about our shortlisted variants. *FineMAV*'s performance is supported by several lines of evidence. The first verification comes from the 'gold standard' replication set (the best examples of validated causal adaptive variants). Not only did *FineMAV* replicate a signal in well-know cases of strong selection, but also narrowed it down to a single functional SNP (often in high LD regions). The number of such positive controls extends to other variants that were not included in the 'gold standard' panel, but whose evidence of causality is also strong, providing additional support. *FineMAV* rediscovered many known variants with prior evidence for being causal of positive selection signals including several SNPs involved in eye, hair and skin pigmentation in non-Africans, such as rs1800414 in *OCA2* (skin lightening in East Asians) (188-190), rs1042602 and rs1126809 in *TYR* (pigmentation and freckling in Europeans) (191-193), rs12350739 in *BNC2* (freckling and colour saturation of human skin pigment in Europeans) (194) but also rs1047781 in *FUT2* (an enzyme-inactivating mutation conferring advantage in avoiding certain viral infections in East Asians) (52, 195).

Finally, *FineMAV* picked up a variant with no prior implication of functionality that was experimentally validated in parallel to our study, which provides another proof of its performance. We picked-up a missense rs11150606 as sixth top scoring variant in East Asians and falling in *PRSS53* whose function was largely unknown. *PRSS53* encodes one of the polyserine proteases called polyserase-3 (POL3S) which hydrolyses peptide bonds. During the preparation of this thesis Adhikari *et al*., showed that *PRSS53* is highly expressed in the hair follicle

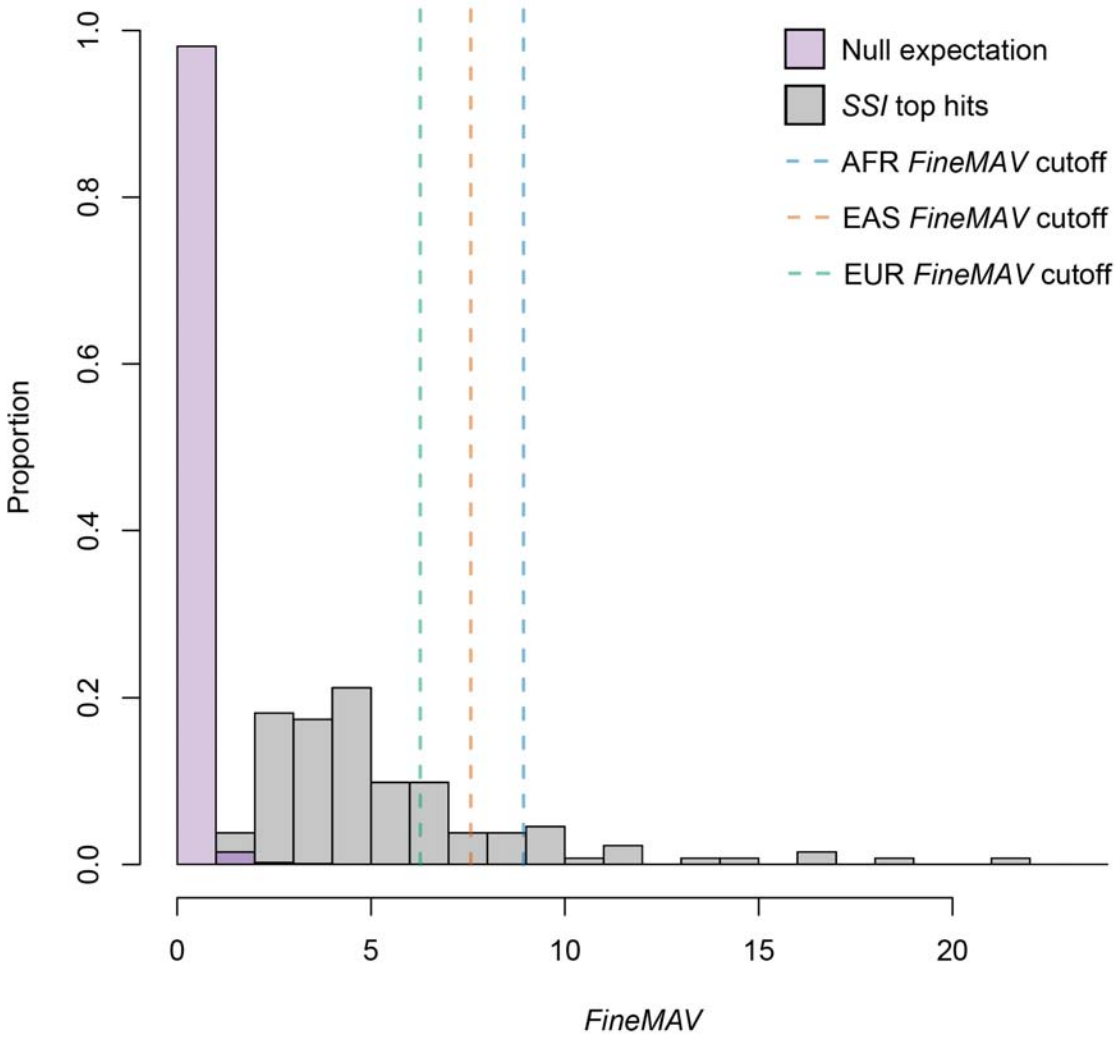

Figure 16. Distribution of *FineMAV* scores in *SSI* outlier genes. The null expectation is the distribution of *FineMAV* scores of variants matching the continental derived allele frequency of our top outliers (*DAF* ≥ 0.24) across all three populations pooled together. We then looked at the distribution of *FineMAV* scores of top SNPs falling in *SSI* outlier genes and their 50 kb surrounding regions (*SSI* ≥ 0.18 which corresponds to ~1% of top genes) and found that it is significantly different from the null expectation (p-value < 2.2×10^{-16}). Vertical lines indicate *FineMAV* cutoffs to include top 100 variants in each population.

and rs11150606 has been associated with hair shape in East Asians (196). The authors confirmed functionality of rs11150606 by *in vitro* assays showing that it affects processing and secretion of the gene product potentially contributing to the straight hair phenotype, similar to the well-established gold standard *EDAR* variant (196). They also showed that the genome regions associated with scalp hair features are enriched for signals of recent selection in humans (196). This can be considered as another example proving validity of our method in picking up truly functional variants.

2.3.2.3. Novel candidate variants across Africa, East Asia and Europe

We performed a new analysis of 1000 Genomes Project Phase 3 whole-genome sequence data (142) using *FineMAV* focusing on identifying individual putatively-selected SNPs driving recent local adaptations (adaptations that arose after the out-of-Africa population expansion). Our analysis overlays multiple lines of evidence for causality to prioritise the vast numbers of potential candidates in order to identify a small number for experimental follow up.

Although we have thus far highlighted known variants replicated in our analysis that serve as positive controls evaluating our method's performance, the vast majority of our outliers are novel and fall in non-coding regions (Figure 15); all of them are characterised by high functional prediction and derived allele patterns similar to the 'gold standards'. We also see potential signals of convergent or parallel evolution (31), i.e. selection on the same gene in geographically distant populations, but on a different SNP e.g. *BCOR*, *CDH13*, *FOXD1*, *FOXP1*, *HDAC8*, *MYH15* and *NFIB* all have a highly-scoring outlier SNP in two out of three populations analysed (as multiple mutations at the same loci can give rise to a similar phenotype (21)). Finally, our analysis picked up several novel potentially interesting candidates, including variants on the X and Y chromosomes which have been underrepresented in previous genomic scans, but further functional testing is needed to verify these findings.

Although our study focuses on local adaptation driving population differentiation at the continental scale, *FineMAV* might be also applied to study signals of selection within continents. It is also possible to investigate signal of selection shared between populations by relevant population grouping depending on user's purposes, e.g. we investigated selection that happened outside Africa by pooling East Asians and Europeans together (Appendix B).

In the following sections we discuss some some intriguing novel alleles, and speculate on plausible selection pressures. The functional significance of the novel candidate variants presented here needs to be experimentally validated, but narrowing their signal of selection to a single most likely candidate SNP is already a starting point in such efforts.

2.3.2.3.1. Nonsense variants

We observed some high-scoring nonsense variants among our top candidates, suggesting pseudogenization of *PKD1L2* (an endogenous fatty acid synthase in skeletal muscle (197)) in Europeans, *ZNF208* (zinc finger and SRY-interacting protein (198)) in Africans, as well as *ZAN*, *OBSCN* (sacromeric signaling protein involved in myofibrillogenesis (199)) and *MAGEE2* (melanoma-associated antigen expressed in the brain (200)) in East Asians. Mice homozygous for knockout alleles of *OBSCN* and *ZAN* are viable and fertile (201, 202); *ZAN* is particularly interesting as it encodes a zonadhesin protein located in the acrosome that mediates the species specificity of sperm binding to the extracellular coat of the egg (zona pellucida) (203). Sperm from zonadhesin-null mice exhibit dramatically higher levels of inter-species gamete adhesion without alteration in fertility (202). Zonadhesin is reported to be a rapidly-evolving protein with a high level of divergence between closely-related species, but is similar in species capable of interbreeding (204, 205). The adaptive advantage of species specificity conferred by zonadhesin might be the limitation of cross-species fertilization and avoidance of sterile hybrids (205). However, polymorphism data in humans reveal a signature of positive selection on haplotypes carrying a frameshift mutation (204). We find a signal of selection at a nonsense mutation (rs2293766) present at 51% frequency

in East Asians, but virtually absent elsewhere. An even higher frequency difference is observed for a stop allele at rs1343879 in *MAGEE2* on the X chromosome. Selection at this locus was previously reported by Yngvadottir *et al*., who observed lower diversity in haplotypes carrying the stop allele than in the others and concluded that, like *ZAN,* the truncated MAGEE2 conferred a selective advantage in East Asia (206).

2.3.2.3.2. Missense variants

FineMAV also highlighted rs6048066, a missense variant in *TGM3* in Africans. The *TGM3* gene product (TGase 3) is involved in the keratinization of the epidermis and hair follicle by crosslinking structural proteins, thereby contributing to hair structure, epidermal barrier functions and wound healing (207, 208). *Tgm3* knockout mice do not exhibit severe malformation apart from striking abnormalities of hair follicle function and hair development, manifested by rough-looking, curly or brittle hair (208-210). The missense variant we report here falls in the catalytic core of the protein, as does the mouse nonsynonymous we^{Bkr} allele causing the wavy coat and curly whiskers phenotype (210). The absence of TGase 3 seems to affect hair fiber morphogenesis, and could play a role in the maintenance of body heat in mammals (211). Similarly in humans, TGase 3 is likely to participate in human hair shaft keratinization and scaffolding (207), and its deficiency has been linked to Uncombable Hair Syndrome characterised by dry, frizzy and wiry hair, often with slower growth rate (212). SNPs in *TGM3* have been weakly associated with hair diameter in humans (213), and proteomic profiling of human hair shafts identified TGase 3 as a major component of the hair fiber and revealed considerable variation among samples of different ethnic origins, with the lowest levels in African Americans and Kenyans (214). We propose that this missense variant (rs6048066) might cause enzyme deficiency and contribute to African hair texture, hypothesised to have experienced strong positive selection in equatorial climates due to body-temperature-regulation (33, 215).

Another novel signal detected in African populations falls in *SPTA1*, encoding erythrocytic spectrin, alpha 1, a principal component of the erythrocyte membrane

skeleton, which is essential for the arrangement of transmembrane proteins, determining red cell membrane stability, cell shape and deformability (216-218). Variants in *SPTA1* have been associated with quantitative hematologic traits (219-221), and those causing its deficiency result in hemolytic anemias characterised by elliptically shaped erythrocytes (also seen in *Spta1*$^{-/-}$ null-mice) (222, 223). The high prevalence of such anemia in Africa (10 times higher in West Africa than in Europe or USA (224)) raised the question of a selective advantage, possibly contributing to protection against malaria (225, 226). It has been shown that decreased spectrin level inhibited malaria parasite growth *in vitro* (227) and in a mouse model (228). This evidence suggests that a functionally and structurally normal host membrane is necessary for parasite growth and development (225, 227). *FineMAV* pinpointed rs7547313 (Ile>Val) as a likely selected variant present at 0.37 frequency in Africans but absent elsewhere. Furthermore, this variant was reported to be an eQTL associated with lower expression of *ACKR1* [MIM: 613665] (also known as *DARC*); p-value = 0.000017 (200). It is worth saying that rs7547313 is not in LD with the known Duffy O allele (rs2814778); r^2=0.000228497. However, the functional effect of this missense variant on the protein level and malaria parasite growth remains uncertain.

2.3.2.3.3. Regulatory variants

Regulatory variants are particularly interesting as they form the most abundant functional category among *FineMAV* outliers (Figure 15) and are responsible for the bulk of human phenotypic variation (17, 140, 170-172). However, the functional effects of regulatory variants are currently difficult to predict and interpret. We find a signal of selection on rs2303893 - a splice region intronic regulatory variant that falls in a region flanking the *HADHB* promoter (160) and is associated with increased *HADHB* expression in adipose, arterial and brain tissue (Geuvadis and GTEx data (200, 229)). *HADHB* encodes the beta subunit of the mitochondrial trifunctional protein involved in the beta-oxidation of fatty acids, and its deficiency causes severe phenotypes (230-232), but the reason for selection in East Asians remains enigmatic.

Another interesting candidate selected in East Asians is rs2224442 falling in a promoter flanking region in the intron of *VRK1*. The region surrounding rs2224442, although non-coding, is characterised by high conservation across taxa and presence of DNaseI hypersensitivity. VRK1 is a protein kinase implicated in mitotic and meiotic cell cycles (233, 234) which plays an important role in gametogenesis in multiple species (235-238). VRK1-deficient organisms show abnormality of reproductive organs, followed by defects in germ cell development (235-238). Both sexes of VRK1-null mice have been reported to be infertile displaying defects in sex organs, oogenesis and spermatogenesis (239-242). It might be that this regulatory variant affects the expression level of *VRK1* and modulate maturation of gametes.

2.3.3. Discussion

The aim of this study was not to perform another selection scan, and it should not be interpreted in that way. Instead, it aims to refine a proportion of local adaptations to a single variant and prioritise candidates for further functional validation, as current methods often do not pinpoint causal SNPs. Therefore, this section provides a decision-making algorithm for elucidation of most likely causal variants that precedes laborious experimental work as it is impractical to assay thousands of variants in a high-throughput fashion. To do so, we introduced the *FineMAV* statistic which combines measures of population differentiation, derived allele frequency and molecular functionality. As it is difficult to distinguish true biological signals from false positives using population genetic variation data alone, incorporation of diverse functional annotations (such as predictors of deleteriousness) should improve the pinpointing of likely causal variants, as it has in the detection of disease-causing variants (243). It is worth noting that variants classified as damaging alter the level or biochemical function of a gene product, but do not necessarily decrease the reproductive fitness of carriers (168, 244). The functional consequence of the 'damaging' change for a person depends on many factors and can be either negative or positive (as deficiency alleles might be either beneficial or detrimental) depending on the environmental context. For instance, variants disadvantageous in one environment can be favored under different conditions e.g. sickle cell (62), *CPT1A* (55, 56).

FineMAV was calibrated and tested using a gold standard panel of the eight best examples of experimentally-validated causal variants underlying signals of positive selection in humans, and was able to identify the known functional candidate in all cases (Figure 8 and Figure 9). Using the complete 1000 Genomes Project dataset (142), we then ranked all genome-wide SNPs based on their *FineMAV* value and identified extreme outliers in the upper tail of the empirical genome-wide distribution in Africa, Europe and East Asia. *FineMAV* rediscovered other known variants with strong prior evidence for being causal of positive selection signals, but which were not part of the positive control set which provides additional support for our method. We also identified potential functional variants

in other genes reported to be under strong positive selection in the literature (with strong *SSI* score) where the causal mutation has not been confirmed yet, including *LPP*, *PCDH15* and *PRSS53*. The selection signal in *PCDH15* and *PRSS53* was attributed to a single missense variant (rs4935502 and rs11150606 respectively), replicating the results obtained by *CMS* (155, 196).

The signal in *BNC2* was particularly strong in Europeans, as reflected by a cluster of 12 SNPs found among the top 100 hits in the *FineMAV* distribution (Figure 10.C). The hypothesised casual SNP (the intergenic rs12350739) was the second highest-scoring *BNC2* variant in our analysis and has been reported to be a functional eQTL as it falls in a highly-conserved melanocyte-specific enhancer and regulates *BNC2* transcription (194). The highest-scoring *BCN2* variant (rs10962600) might also contribute to the differential expression of *BNC2* isoforms, as several regions inside and outside of the *BNC2* gene contain enhancer features (194). Interestingly, *BNC2* has been highlighted as one of the genes present in a region of the human genome that shows increased levels of Neanderthal ancestry (Figure 17), suggesting that Neanderthal introgression might have provided modern humans with adaptive variation for skin phenotypes involving *BNC2* (30, 129, 134, 194). Furthermore, a cluster of high-scoring SNPs in *FineMAV* analysis might be indicative of introgression as a source of adaptive variation as opposed to advantageous *de novo* mutations that usually arise individually. We also found other candidate SNPs falling in regions proposed to be adaptively introgressed from an archaic source (27 SNPs in total) in *GNAI2*, *GPATCH1*, *IRF6*, *POU2F3*, *RASSF1*, *SEMA3F* and *SLC38A3* (Figure 17) (30, 129, 134, 245) suggesting that some of the candidates might be of archaic rather than *de novo* origin. However, the origin of the adaptive mutations is not the focus of this study and has been carefully analysed elsewhere (30, 129, 134, 245).

Finally, *FineMAV* picked up variants with modest to high derived allele frequency ranging from ~0.25 to ~1 within continental populations (Figure 12). Most classical methods detect only extreme allele frequency differences between populations, which are less likely to arise by chance (20). On the other hand, highly functional alleles are less likely to be subjected to random changes in their frequency, thus it seems that filtering out neutral variation by applying functional information might allow more examples of weaker sweeps (potentially including

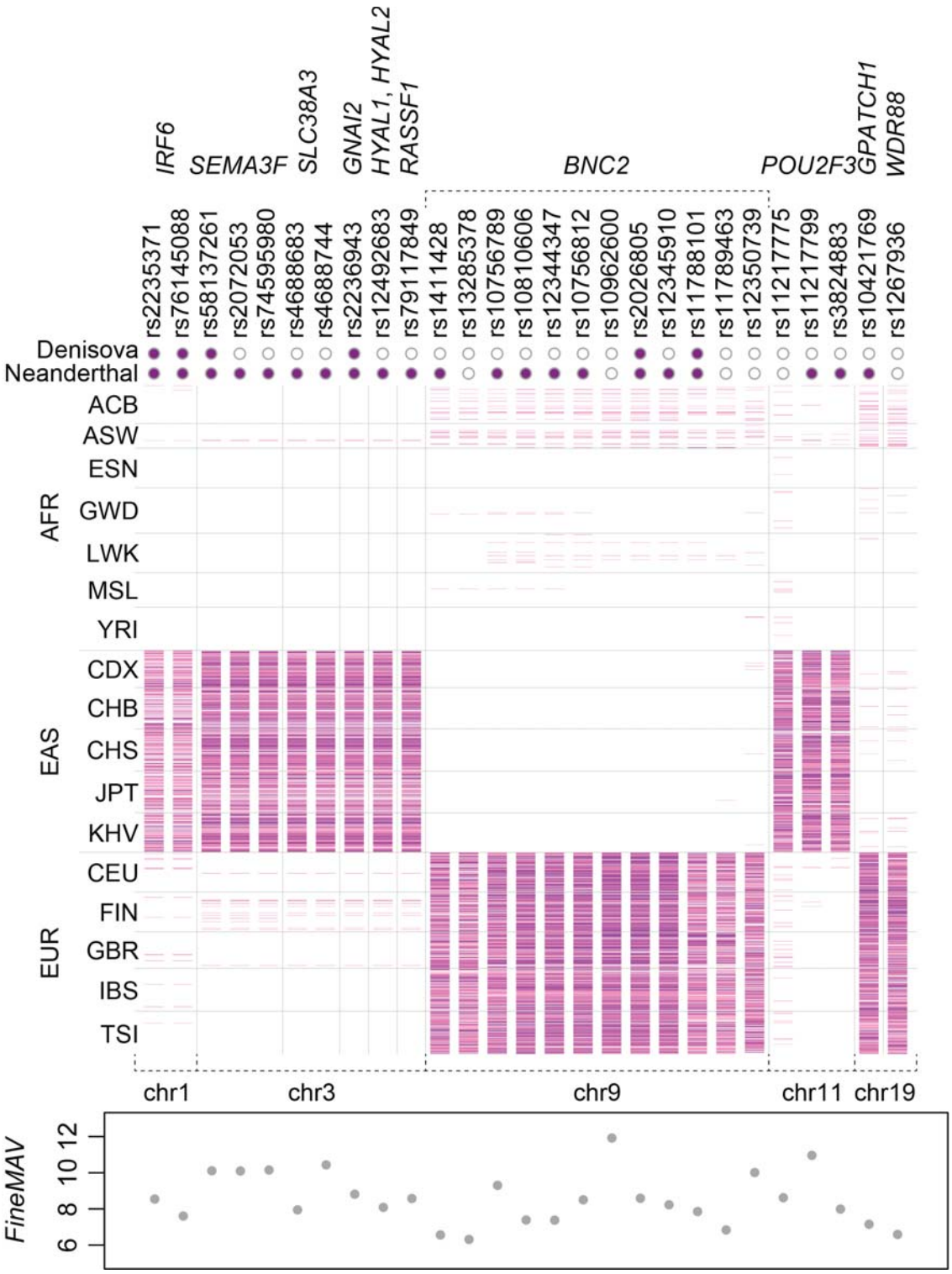

Figure 17. Genotypes of putatively introgressed SNPs identified by *FineMAV*. Rows represent individuals from Phase 3 of the 1000 Genomes Project (142) grouped by population, columns indicate variant sites picked up by *FineMAV* and falling in regions putatively introgressed from archaic hominins (30, 129, 134) ordered by genomic location. The associated gene name is given for each SNP. The first two rows specify the Neanderthal (133) and Denisova (246) genotypes coloured according to genotype: white dot – homozygote for ancestral allele; violet dot – homozygote for derived allele; pink – heterozygote. Human genotypes are denoted by lines (using the same colour coding). The bottom panel specifies the *FineMAV* score for each variant.

selection on standing variation) to be discovered, which are characterised by more modest allele frequency shifts (20, 21), although our method has no power to detect low selection coefficients that do not produce a population differentiation pattern. It is worth noting that the lack of *FineMAV* hits on the Y chromosome (only one in the top 300) shows strong dependence on the CADD score prediction.

2.4. *FineMAV* application to various populations

2.4.1. Materials and methods

After the calibration of our method and assessment of its performance in African, East Asian and European populations in the 1000 Genomes Project dataset, we applied it to investigate population-specific local adaptations in Egyptians, Ethiopians, Greeks, Lebanese, Native Americans and South Asians as described below.

2.4.1.1. Admixed Americans and South Asians

We ran *FineMAV* analysis in Admixed Americans (AMR) and South Asians (SAS) from the 1000 Genomes Project, Phase 3 data release (142) together with the three main continental populations (described in the previous section) as follows: AFR, AMR, EAS, EUR; n = 4; x = 2.98 and AFR, EAS, EUR, SAS; n = 4; x = 2.98. *DAF*, *DAP* and *FineMAV* values were calculated as described earlier.

2.4.1.2. Non-admixed Native Americans

We searched for local adaptations in non-admixed Native Americans (nAMR) using a dataset comprised of unpublished low coverage whole-genome sequences from 24 Quechua from Peru generated at WTSI. In total, 29 Quechua were sequenced on either an Illumina Genome Analyzer II using 108 bp paired-end reads or HiSeq 2000 with 100 bp paired-end reads with insert size of 300-500 bp. Reads were aligned to GRCh37 (hg19/NCBI37) for general sequencing QC and yielded average coverage of 4-6x. The 29 BAMs were then merged with a subset of

the 1000 Genomes Phase 1 and 2 samples using the varpipe tool. Variants and genotypes were called in the merged dataset by Luca Pagani and Petr Danecek (Wellcome Trust Sanger Institute) using Samtools and the procedure described in (247). Samples showing more than 5% European ancestry in ADMIXTURE analysis using common variants were excluded from subsequent analysis leaving a total of 24 individuals. *FineMAV* analysis in nAMR were performed using 3 reference populations from the subset of 1000 Genomes Project: AFR (Americans of African Ancestry, Southwest USA [AWS], Luhya in Webuye, Kenya [LWK], Yoruba in Ibadan, Nigeria [YRI]), EAS (Han Chinese in Bejing, China [CHB], Southern Han Chinese [CHS]), EUR (Utah Residents with European Ancestry, USA [CEU], Iberian Population in Spain [IBS], Toscani in Italia [TSI]); $n = 4$; $x = 2.98$. *DAF*, *DAP* and *FineMAV* values were calculated as described earlier. Common variants failing Hardy–Weinberg equilibrium and not called in 1000 Genomes Project, Phase 3 data release (142) were excluded.

2.4.1.3. Greeks, Lebanese, Egyptians and Ethiopians (GLEE)

The GLEE dataset comprised the following individuals: 100 Egyptians (EGP) and 100 Ethiopians (ETP; 25 each from Amhara, Oromo, Wolayta and Gumuz) sequenced at 8x depth using Illumina HiSeq 2000 (247); 100 Greeks (GRK) from the HELIC TEENAGE (TEENs of Attica: Genes and Environment) cohort comprising young adults from Athens, Greece, that were sequenced at 30x depth using the Illumina HiSeq X10 platform, then downsampled to ~8x using the Samtools -s option to have a coverage comparable to other populations in the dataset; 100 Lebanese (LEB including 34 Christians, 28 Druze and 38 Muslims) sequenced to an average depth of 8x using Illumina HiSeq 2500. This dataset was merged with similar data generated by the 1000 Genomes Project including CEU, CHB and YRI (around 100 individuals each) and the genotypes were called jointly using Samtools and Bcftools. Calling and quality control analysis were performed by Petr Danecek, Marc Haber, and Javier Prado-Martinez (Wellcome Trust Sanger Institute).

Genotype calling accuracy was assessed by checking concordance with array data from the same samples and was found to have >99% concordance. Outlier samples (deviating >8 SD from the core variation of the population in the PCA performed using Eigensoft) and first and second degree relatives were excluded from further analysis leaving: 91 EGP, 25 Amhara, 25 Oromo, 24 Wolayta, 23 Gumuz, 98 GRK, 34 LEB Christians, 28 LEB Druz and 38 LEB Muslims. *DAF*, *DAP* and *FineMAV* values were calculated for derived and ambiguous alleles (annotated accordingly to Ensembl Compara (160, 182)) using a custom script (SNPs only; indels were omitted). The *FineMAV* analysis were performed in the following contexts: i) CEU, CHB, YRI; $n = 3$; $x = 3.5$; as a sanity check to compare the concordance of *FineMAV* results calculated using full 1000 Genomes Project, Phase 3 (142) and results calculated from a single continental populations; ii) CEU, CHB, EGP, YRI; $n = 4$; $x = 2.98$; to investigate Egyptian-specific signal; iii) Amhara, Oromo and Wolayta were pooled together as admixed Ethiopians (247) (ETP) and analysed in the following context: ETP, CEU, CHB, YRI; $n = 4$; $x = 2.98$; iv) Gumuz (non-admixed Ethiopian population (247)) was processed separately: CEU, CHB, Gumuz, YRI; $n = 4$; $x = 2.98$; v) CEU, CHB, GRK, YRI; $n = 4$; $x = 2.98$; to explore Greek-specific signal; vi) CHB, GRK, YRI; $n = 3$; $x = 3.5$; replacing CEU with GRK in the inter-continental comparison; vii) CEU, CHB, LEB, YRI; $n = 4$; $x = 2.98$; to investigate Lebanese-specific signal (all Lebanese pooled together); viii) LEB Christians, LEB Muslims; $n = 2$; $x = 4.96$; to explore differentiation between different Lebanese groups.

2.4.2. Results

2.4.2.1. *FineMAV* analysis in Native Americans and South Asians

2.4.2.1.1. AMR and SAS from 1000 Genomes Project

FineMAV analyses of the 1000 Genomes Project Admixed Americans (AMR) and South Asians (SAS) revealed little population-specific variation in these populations (Figure 18). Even though the signal there was lower due to population admixture, we nonetheless saw promising candidates for local adaptations found exclusively in those populations. Interestingly, the only clear outlier observed in SAS, found at 0.54 frequency but virtually absent elsewhere, was a missense rs201075024 falling in *PRSS53* (Figure 18.A). A different non-synonymous variant in *PRSS53* was picked-up in East Asians (see previous section: Functional validation), and has been recently shown to affect enzyme processing and secretion potentially contributing to the straight hair phenotype (196). Furthermore, East and South Asian alleles fall in close proximity, only 10 bp apart (Figure 19), which might indicate a similar functional consequence and convergent evolution of a hair-related phenotype.

The *FineMAV* signal in Admixed Americans was lower (Figure 18.B) as admixture decreases differentiation and population-specific derived allele frequency, with the top 3 scores being missense variants: rs148608573 in *MAP7D1*, rs142326775 in *ZNF438* and rs34890031 in *LRGUK* (mouse homologue is essential for multiple aspects of sperm assembly and function (248)). Even though admixture decreases the *FineMAV* signal, the one-directional admixture i.e. European gene flow to Americas affects the frequency of derived Native American alleles, but not their purity (as private American alleles would still be found exclusively in Americas at high *DAP* values). In the case of common derived alleles selected to high frequencies before an admixture event, their *FineMAV* signal should still be detectable after European gene flow to Americas (assuming their high functional

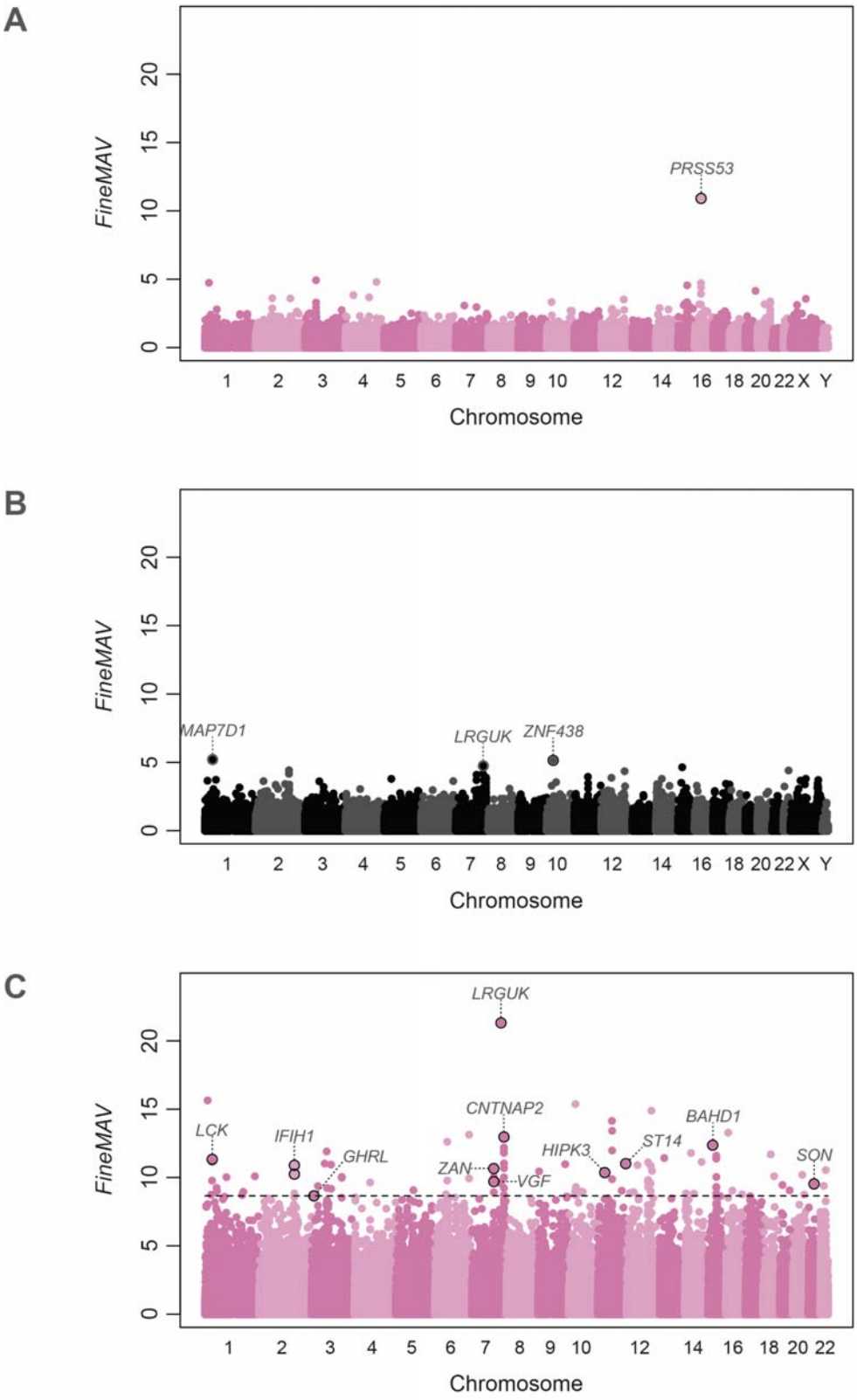

Figure 18. Manhattan plot of genome-wide *FineMAV* scores in Native Americans and South Asians. *FineMAV* scores calculated for genome-wide SNPs in: (A) – South Asians (SAS); (B) – Admixed Americans (AMR); (C) – Non-admixed Native Americans (nAMR; the threshold (dashed line) was set to include the top 100 variants). Each dot in the Manhattan plots represents a single SNP plotted according to coordinates in GRCh37. Interesting candidate variants are labeled with the name of the gene they fall into.

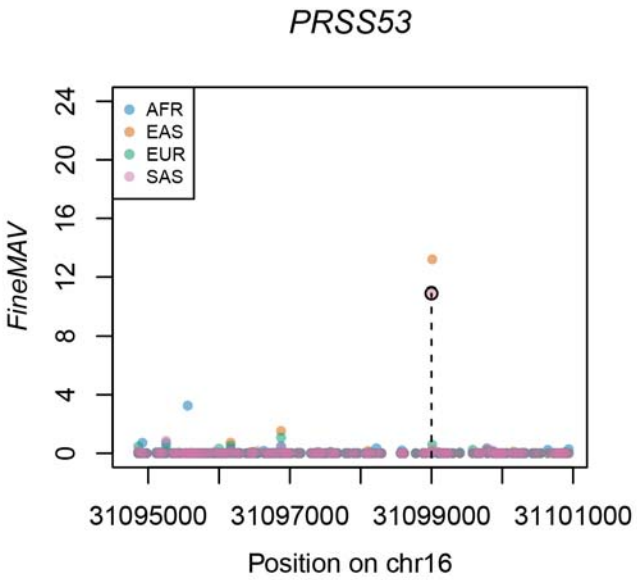

Figure 19. Signal of selection in the *PRSS53*. *FineMAV* scores of variants in the genomic window spanning *PRSS53* are plotted as dots. Genomic positions are given in bp according to GRCh37. The selected variant in South Asians (rs201075024) is marked with a dashed line with the variant selected in East Asians (rs11150606) just above it.

prediction (*CADD*) and *DAP* scores) even if their allele frequencies decreased substantially. Therefore, the strongest local adaptations should still fall in the tail of the *FineMAV* distribution even in cases of recent one-directional gene flow.

2.4.2.1.2. Non-admixed Native Americans

We found a strong signal of local adaptation in the non-admixed Native American population, with many potentially interesting candidates (Figure 18.C and Appendix B), although the allele frequency calculation was based on a small sample (n = 24). There was a substantial overlap between the top outliers found in admixed and non-admixed populations (reaching 50% among the top 50 hits). We also saw a moderate correlation (r = 0.58; p-value = 2.661×10^{-10}) between the *FineMAV* values of the top 100 non-admixed hits and their admixed equivalents. The highest scoring variant (similarly to results in admixed Americans) was a missense rs34890031 (found at 0.77 frequency) in *LRGUK*, a gene that plays a critical role in male fertility (248). All of the above suggest that *FineMAV* is indeed able to pick up the strongest selection signals even in admixed populations in cases of one-directional gene flow when the source population is used in the analysis.

Other interesting variants include missense rs62621285 in *ST14* and a stop gained rs2293766 in *ZAN*, present at 56% and 79% respectively. This nonsense mutation in *ZAN* (involved in sperm species specificity (202, 203)) has been introduced in the previous section as one of the top variants selected in East Asians, yet its frequency in Native Americans is even higher. *ST14* is known for playing an important role in hair development and growth and its deficiency in mice causes brittle, thin, uneven, and sparse hair, or even a complete absence of erupted pelage hairs and vibrissae in null animals (249-254), which is interesting considering the reduced body hair in Native American populations (255, 256). Furthermore, *ST14* is required for skin keratinization, formation and maintenance of the epithelial and epidermal barrier and integrity (250, 253, 254, 257-263). It seems that this gene has pleiotropic functions affecting the development of the epidermis, hair follicles, and cellular immune system (254) as it has been shown that the *ST14* protein product (matriptase) is also an influenza virus-activating protease supporting

multicycle viral replication in the human respiratory epithelium (264-266). The influenza genome does not encode any proteases and relies on host proteases for the cleavage activation of the surface receptor proteins in order to fuse with the host cell membrane (264-266). Knockdown of matriptase in human bronchial epithelial cells significantly blocked influenza virus H1 subtype replication (264-266).

We detected additional putatively causal mutations falling in genes linked to immunity including: i) rs4924468 in a promoter flanking region upstream of *BAHD1* (null mice exhibit decreased susceptibility to bacterial infection (267)); ii) rs12478730 and rs12474958 in the *IFIH1* enhancer (mediating the immune system's interferon response to RNA viruses including hepatitis B and C, influenza A, paramyxoviruses (mumps, measles, respiratory syncytial virus causing bronchiolitis and pneumonia), enteroviruses (including poliovirus), dengue, rotavirus and *Herpes simplex* virus among others (268-283); null mice were more susceptible to viral infection, experienced more severe symptoms and reduced survival (284-288)); iii) a missense/promoter flanking region mutation (rs145088108) in *LCK* (T-cell proliferation and activation gene whose deficiency causes severe immunodeficiency (289-297)) and iv) missense/TF binding site mutation (rs147302393) in *SON* (important for trafficking of influenza A virions to late endosomes during infection (298) and repressing transcription of hepatitis B virus (299)).

Furthermore, the *SON* protein product was shown to regulate ghrelin receptor (*GHSR*) transcription in the brain by repressing its promoter activity (300). Ghrelin (encoded by *GHRL* and acting via GHSR) is a pleiotropic hormone secreted by the stomach that promotes food intake, weight gain and fat storage by reducing fat utilization (beta-oxidation), but also decreased glucose tolerance and decreased insulin sensitivity in mice and rats (301-305). Knockout mice display increased utilization of fat as an energy source on a high fat diet, reduced food intake, weight gain and adiposity, increased energy expenditure and locomotor activity, decreased circulating glucose level, improved glucose tolerance, increased circulating insulin level and secretion (304, 306, 307). It seems that the absence of ghrelin protects from diet-induced obesity and type 2 diabetes (306, 307). On the other hand, the ghrelin circulating level was shown to increase during fasting and it

was suggested that it prolongs survival in starved humans but may also play a role in fetal adaptation to intrauterine malnutrition, while its absence impairs fasting tolerance (301, 302, 305, 308-310). It seems that ghrelin plays an important role in the metabolic adaptation to nutrient availability and determines the type of substrate (fat or carbohydrate) that is used for maintenance of energy balance (304, 306). Interestingly, one of the high-scoring variants in the Quechua population is a missense variant (rs4684677) falling in *GHRL*. *GHRL* encodes preproghrelin, which is a precursor of two peptides ghrelin and obestatin. Obestatin is ghrelin's antagonist involved in satiety and decreased appetite contributing to decreased body weight gain (311) and the variant we picked up (Gln to Leu substitution in position 90 of the ghrelin/obestatin prepropeptide; rs4684677) was shown to impact obestatin function. Gln90Leu was slightly more efficient than native obestatin in inhibiting ghrelin-induced food intake (312).

Highlighted example is not the only case of variants falling in genes regulating energy homeostasis, as we also picked up rs189645263 in a promoter of HIPK3 (a known regulator of insulin secretion whose deficiency impairs insulin secretion and glucose tolerance and may play a role in the pathogenesis of type 2 diabetes (313)), and rs116131136 missense/promoter flanking region in VGF (an energy homeostasis regulator). Processing of VGF generates multiple bioactive peptides and mouse homozygotes for the null allele are small, lean with reduced adiposity and increased fatty acid oxidation, hypermetabolic (with increased resting energy expenditure and oxygen consumption), hyperactive, cold intolerant and infertile (314, 315). Furthermore, VGF deficiency is characterised by decreased circulating glucose and insulin levels but increased insulin sensitivity and improved glucose tolerance, resistance to induced obesity and hyperglycemia which indicates that this gene may also play an important role in diabetes (316-320).

Finally, we detected a strong signal in the *CNTNAP2* gene, with a cluster of 9 SNPs in the top 100, which might indicate archaic introgression as a source of this haplotype (similarly to *BNC2* found in Europeans). Indeed, this derived haplotype is also found in the high-coverage Denisova genome, but in a heterozygous state which should be taken with caution as heterozygous haplotypes are rather uncommon in highly inbred archaic hominins and could arise from mapping and calling errors (246).

2.4.2.2. GLEE

The Near East, Southern Europe and East Africa form a region which is key for understanding the evolutionary history of modern humans. The region is at the centre of modern humans' expansion outside Africa and an established source of subsequent expansions such as that during the Neolithic into Europe, Central Asia and possibly back to Africa (247, 321, 322). Yet the genomics of the populations in this area have been little-studied, especially on the whole-genome level.

We first performed a sanity check to ensure that the results we are getting using a single reference population representing each continent (CEU for Europe, CHB for East Asia and YRI for Africa) are consistent with the results obtained for the full 1000 Genomes Project, Phase 3 (142) (reported in previous section). We found a very high concordance between the two runs with ~70% of the top 100 outliers being the same and a high correlation between *FineMAV* values of those 100 candidates (r = 0.85 in Africa, r = 0.83 in East Asia and r = 0.85 in Europe; all with p-value < 2.2×10^{-16}). All gold standards were successfully picked up as high-scoring in the sanity test. Furthermore, we detected two well-know adaptive variants among the top 100 hits that were missed in the full 1000 Genomes Project analysis: (i) rs3211938, a nonsense mutation in *CD36* selected in YRI and conferring protection against malaria and/or the metabolic syndrome (323-325), and (ii) a missense variant, rs1229984, falling in *ADH1B* selected in CHB possibly due to protection against alcohol dependence (326-329). The reason why rs3211938 was picked up in the test run is its high frequency in YRI (29%) compared to the frequency in general African population (12%) sampled by the 1000 Genomes Project (12% in the combined sample is too low to be detected by *FineMAV* at the continental scale analysis). On the other hand, rs1229984 was not picked up in the full 1000 Genomes Project survey as its evolutionary state (ancestral vs derived) could not been inferred and was subsequently excluded from the analysis, while this study was less stringent and ambiguous sites were retained.

We then replaced CEU with genetically close GRK population to see how it affects the analysis. The results for CHB and YRI remained virtually the same, while the most prominent difference between GRK and the general European population

sampled in 1000 Genomes Project Phase 3 was the loss of the selection signals underlying lactose tolerance (rs4988235 in *MCM6*; 0.51 *DAF* in EUR vs 0.13 *DAF* in GRK) and blue eyes (rs12913832 in *HERC2*; 0.64 *DAF* in EUR and 0.34 *DAF* in GRK) in Greeks (Figure 20.A and B). Conversely, the allele with the biggest difference in the *FineMAV* score between GRK and EUR that shows a signal of selection in Greeks but not EUR was an amino acid change (rs35392772) in *MOS*, a cell cycle-regulator essential for oocyte maturation in vertebrates (330-333) (0.24 *DAF* in GRK vs 0.16 *DAF* in EUR) (Figure 20.A and B). However, we did not pick up any convincing GRK-specific adaptation signal in a 4-population comparison (CEU + CHB + GRK + YRI) and the apparent moderate clusters seen in the Manhattan plot fall in repetitive elements or duplicated genes likely underlying mapping -> calling artifacts rather than true signals (Figure 21).

Similarly, we did not find any convincing population-specific signals in Egyptians, admixed Ethiopians, and Lebanese, which is consistent with their known admixture and/or extensive ancestry sharing with both Middle East, Europe, and Africa resulting in little population differentiation (247, 334) (Figure 21 and Figure 22). Finally, we did not detect selection-driven differentiation between Lebanese Christians and Muslims, which implies that the population structure seen in Lebanese is most likely due to population isolation followed by genetic drift rather than positive selection (334) (Figure 21.B and C). We did, however, see some signal of selection in the non-admixed Ethiopian population (Gumuz), although the results are based on allele frequencies calculated in a small sample size (n=23), with top 3 SNPs being: nonsense variant rs7904983 in *PKD2L1* (70% in Gumuz vs 19% in AFR), missense variant rs56683778 in *CCDC80* (48% in Gumuz vs 7% in AFR), and intronic variant rs9938729 in *MVP* (46% in Gumuz vs 2% in AFR) (Figure 22.C).

PKD2L1 is a sour taste and cellular pH sensor; mice lacking *Pkd2l1* showed no or decreased taste response to sour stimuli (335-338). Olfaction enables examination of food source properties including potential acidity manifested by sour taste, stimulating an aversive response (339). It is hard to speculate about the possible reasons for selection of *PKD2L1* loss of function, but variation in this gene was also associated with serum metabolite levels among African Americans (e.g. palmitoleic acid) (340-342).

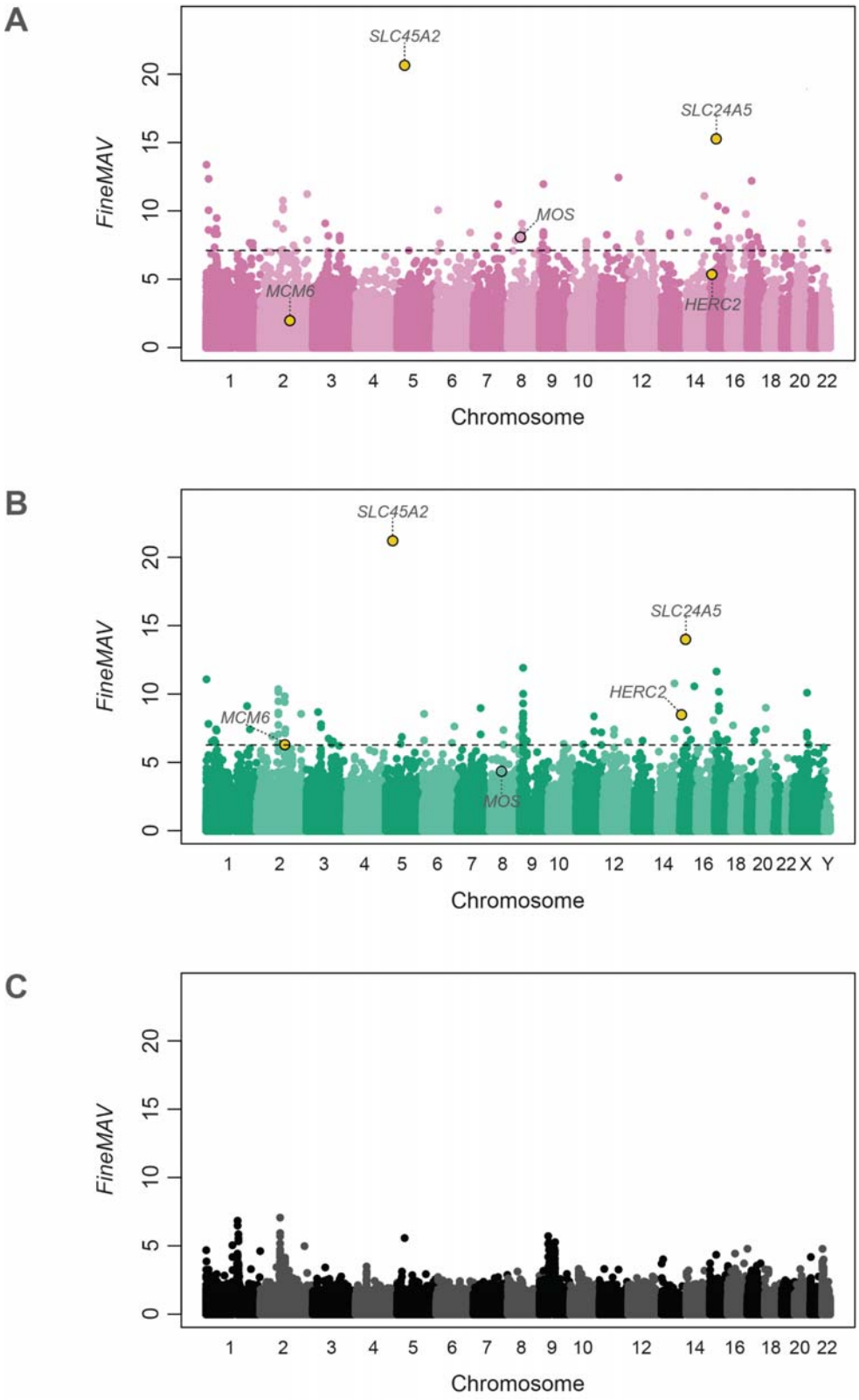

Figure 20. Manhattan plot of genome-wide *FineMAV* scores in Greeks. *FineMAV* scores calculated for genome-wide SNPs in: (A) – GRK (run together with CHB and YRI); (B) – EUR from the full 1000 Genomes Project Phase 3 calculated in the previous section; (C) – GRK (run together with CEU, CHB and YRI). Each dot in the Manhattan plots represents a single SNP plotted according to coordinates in GRCh37. The threshold (dashed lines) was set to include the top 100 variants. All gold-standard SNPs (yellow dots found among the top outliers) and other interesting candidate variants are labeled with the name of the gene they fall into.

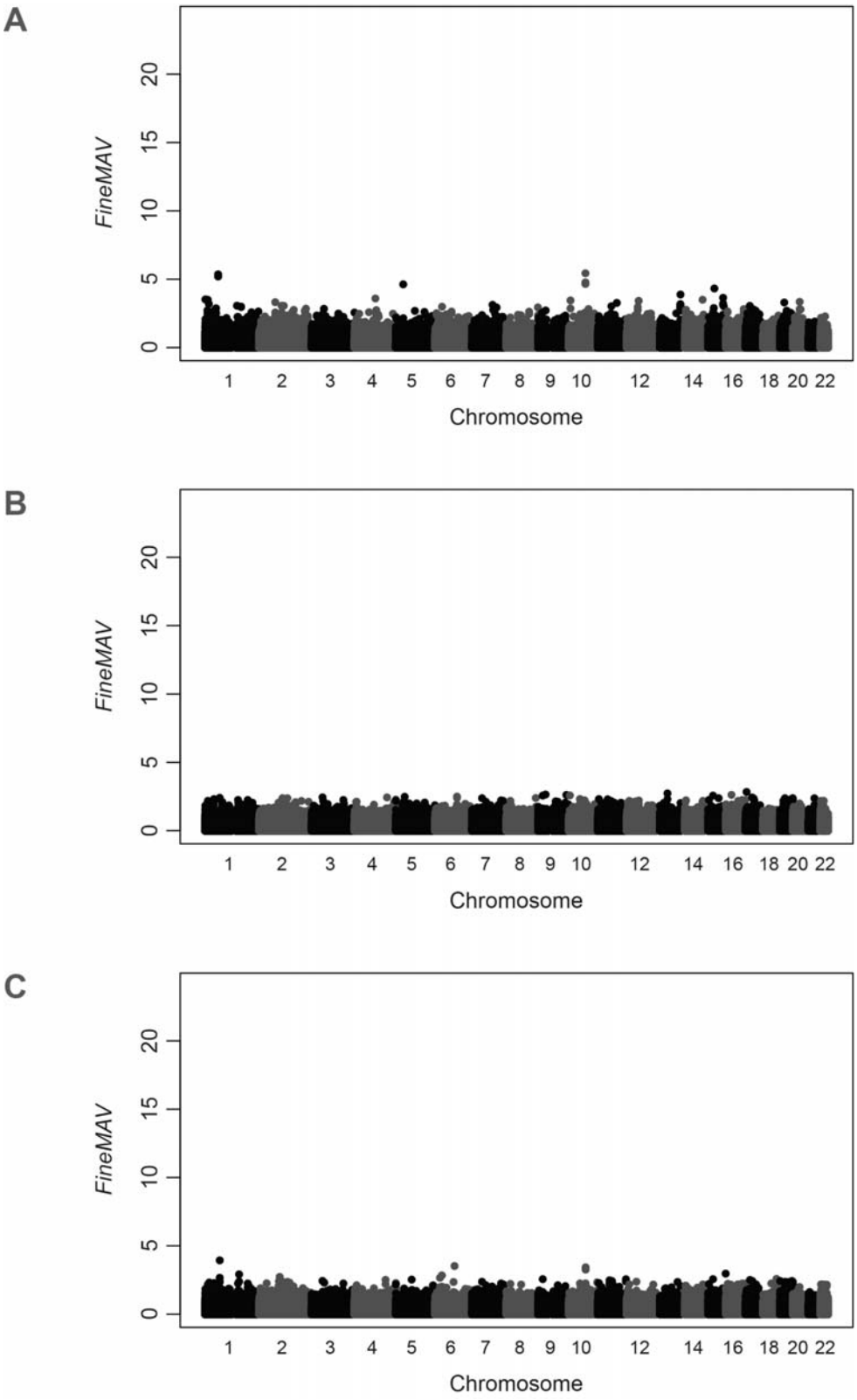

Figure 21. Manhattan plot of genome-wide *FineMAV* scores in Lebanese. *FineMAV* scores calculated for genome-wide SNPs in: (A) – LEB general population (run together with CEU, CHB and YRI); (B) – LEB Christians (run against LEB Muslims); (C) – LEB Muslims (run against LEB Christians). Each dot in the Manhattan plots represents a single SNP plotted according to coordinates in GRCh37.

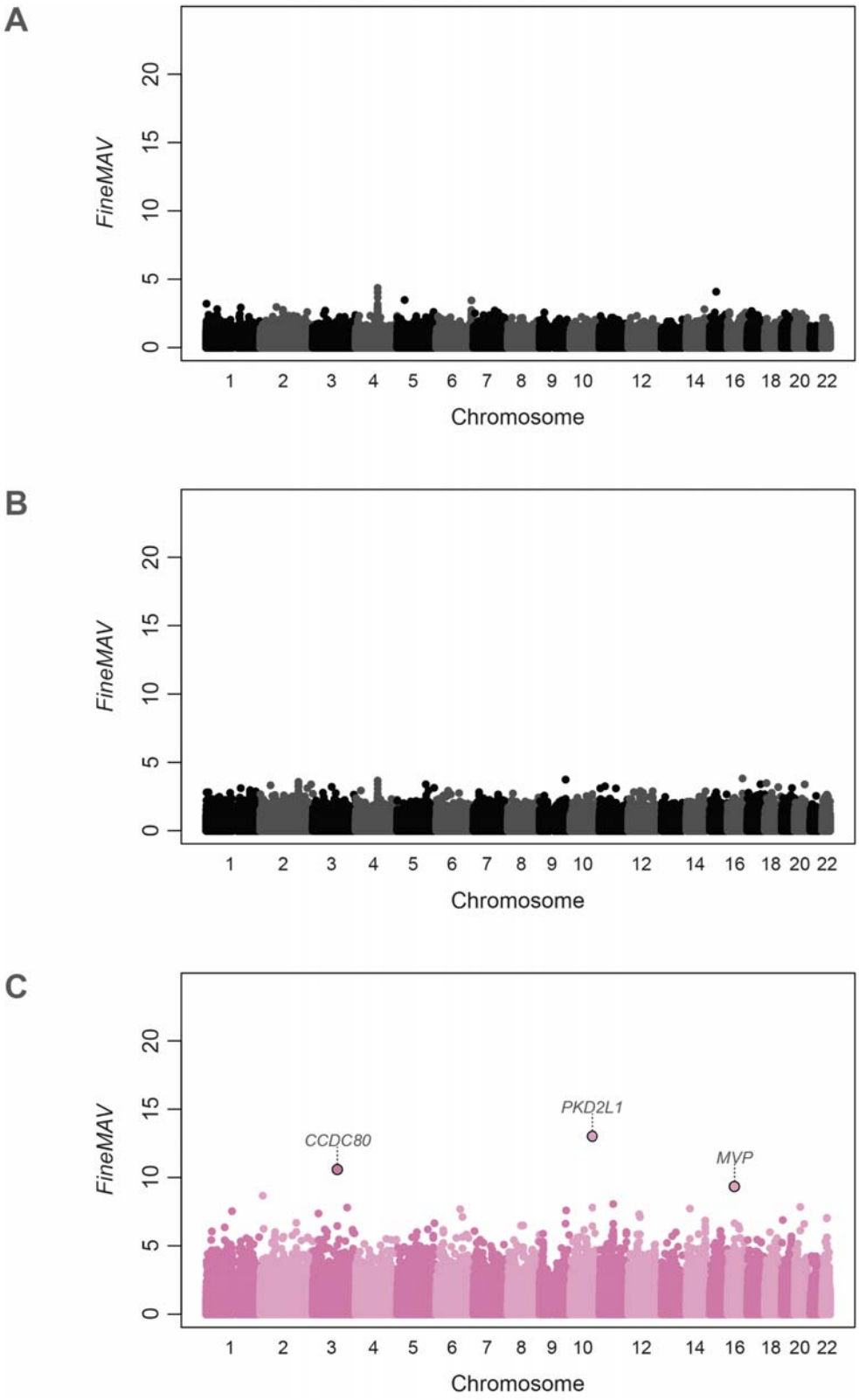

Figure 22. Manhattan plot of genome-wide *FineMAV* scores in Egyptians and Ethiopians. *FineMAV* scores calculated for genome-wide SNPs in: (A) – EGP (run together with CEU, CHB and YRI); (B) – ETP (run together with CEU, CHB and YRI); (C) – Gumuz (run together with CEU, CHB and YRI). Each dot in the Manhattan plots represents a single SNP plotted according to coordinates in GRCh37. Interesting candidate variants are labeled with the name of the gene they fall into.

CCDC80 has been shown to play an important role in adipocyte differentiation (343) and may be a key player in energy metabolism and body weight regulation (344, 345). The absence of *Ccdc80* in mice results in increased caloric intake, decreased energy expenditure, obesity, increased glucose level and enhanced lipolysis with decreased circulating insulin level and impaired glucose tolerance when fed a high fat diet (346). *CCDC80* has been flagged as having a protective role in obesity and diabetes (347).

MVP function has remained elusive. It has been shown to contribute to resistance against *Pseudomonas aeruginosa* lung infection (348) and confer response to an environmental toxin (349). On the other hand, some bacteria incorporate human *MVP* onto their surface in order to escape autophagy (350). Furthermore, *MVP* over-expression has been associated with tumor chemo- and radiotherapy resistance as it is involved in DNA double-strand break repair machineries and was shown to be upregulated in stress conditions (351). One study reported high *MVP* expression related to severe hypoxia in clinical tumors (352). This report highlights *MVP* as putative high-altitude adaptation gene, although such a claim is purely speculative and requires further functional investigation.

2.4.3. Discussion

FineMAV does not aim to detect all selection events, but rather to identify a small number of likely causal variants driving population diversification, therefore it is reassuring that we do not detect much signal in cases where population admixture and/or extensive ancestry sharing between populations has resulted in little differentiation. It has been shown that the recent back-flow of likely Near Eastern, and to a lesser extent European, ancestry to Africa has drastically influenced the genomes of present day Northeast African populations (247, 321, 322). Pagani *et al*. reported the average proportion of non-African ancestry in the EGP and ETP samples (excluding the Gumuz) to be around 80% and 50% respectively (247). Furthermore, the indigenous North African ancestry is closely related to populations outside of Africa as Northeast Africa was the last stop on the migration out of Africa (247, 321). On top of that, a significant signature of a sub-Saharan African component was also reported in North African populations (247, 321). Including proxies of source populations in *FineMAV* comparison cancels out admixed alleles described by low 'purity' score (*DAP*) as they are found across multiple populations, while the frequency of the indigenous-population-specific alleles drops below the detection level as a result of population mixing. Similarly, the South Asian population is made up of two main ancestry components called 'Ancestral North Indian' (ANI) and 'Ancestral South Indian' (ASI) (353). ANI was shown to be genetically close to Middle Easterners, Central Asians, and Europeans, and ranged from 39% to 71% in India with complex population stratification due to endogamy (353, 354). A complex population structure was also reported for the Lebanese population that falls into two main groups: one showing genetic affinity toward present-day Europeans and Central Asians, and the other more closely related to Middle Easterners and Africans due to a different admixture history with neighboring populations driven by culture and endogamy (334). We did not however detect any differentiation between these two groups that was driven by selection.

Nevertheless, *FineMAV* was able to pick up the strongest signals of local adaptation in admixed Native Americans, despite recent admixture (e.g.

rs34890031 in *LRGUK*). One-directional gene flow from Europeans to Americans (decreasing indigenous allele frequency (*DAF*), but not its purity (*DAP*)) is a much simpler scenario than the continuous population mixing at the edge of continents seen in Northeast Africa and Near East, with multiple components and a multi-layered history. The non-admixed Native Peruvians revealed a range of putatively selected SNPs falling in genes related to immunity, especially antiviral response. Historical record documented a massive bottleneck in the Inca Empire (and Americas in general) attributed to infectious diseases acquired upon European contact, mainly smallpox but also measles, influenza, mumps and pneumonia among others (355-358). The selective pressure (pathogen virulence) in immunologically naïve populations having no natural resistance against epidemic disease was very strong and is estimated to have wiped out over 90% of the Peruvian Inca population over only 50-100 years (356, 359). However, it is hard to tell if the signals we picked up were driven by recent strong selection ~500 years ago, or older events, or a combination of both. Similarly, Fumagalli *et al*. also detected local selective pressures acting on *IFIH1* (a sensor of viral RNA involved in antiviral host defense) favouring different alleles in distinct geographical regions (360). They reported directional positive selection in Europe and Asia as well as a long population-specific haplotype that swept to high frequency in South America (360). High F_{ST} between Asian and South American populations and the presence of an extended haplotype in America suggest a relatively recent selective sweep (360). This South American haplotype was defined by two SNPs only 3 bp apart (360). The same two SNPs (rs12478730 and rs12474958), falling in a conserved enhancer, were picked up in our *FineMAV* analysis and might increase *IFIH1* expression conferring stronger protection against viral infections. Notably, variation in *IFIH1* and its increased expression was also linked to increased risk of autoimmune diseases (type 1 diabetes, psoriasis and lupus among others) (360-362).

Finally, we found a signal of geographically restricted selection in energy metabolism genes in both Quechua and Gumuz. Widespread obesity and an elevated risk of developing type 2 diabetes and cardiovascular diseases have been reported for many indigenous communities including Native Americans (363-365). Such an observation has been linked to the so-called 'thrifty gene' hypothesis suggesting that decreased resting metabolic rate and increased energy storage was favoured

in populations historically facing feast-famine cycles (44, 45, 366, 367). Adaptation to food scarcity may predispose to metabolic syndrome in a non-traditional lifestyle with continuous food supply (44, 45, 366, 367). Furthermore, the traditional diet of aboriginal Americans and Ethiopians was estimated to be high in carbohydrates (~70-80%) and low in fat (8-12%), while adoption of a modern lifestyle resulted in much higher fat-intake (35% fat) (368-370). Urban Peruvians and rural-to-urban migrants showed a higher prevalence of obesity and cardiovascular diseases compared with the rural population (although environmental factors play an important role) (371-373), and a general high incidence of hypertension and obesity was reported in Peru among both cosmopolitan and Andean Peruvians (374-381) with nearly a quarter of the adult population at an increased risk of diabetes (382). Similar trends in the prevalence of cardiovascular diseases linked to urbanisation were reported in Ethiopia (370, 383-390). Furthermore, a previous selection scan in indigenous Ethiopian population of Wolaita has also reported a recent positive selection on genes involved in immunity and energy metabolism during prolonged food shortage that were linked to diabetes and obesity susceptibility (370). Apart from diet, high-altitude hypoxia that promotes lipid storage and carbohydrate oxidation might have contributed to metabolic adaptation (370, 391, 392). However, we did not replicate the high-altitude adaptation signals reported previously for Ethiopian highlanders and Andean Quechua (370, 393), although there is no information whether the populations analysed in this study were residing at high-altitude.

3. Functional follow-up of selected candidates

The endeavour to understand selective adaptation requires the in-depth functional validation of candidate causal variants. However, the scale of phenotyping measurements that can be easily and ethically assessed in humans is limited. Even if a variant is shown to associate with some trait(s) in humans, it remains uncertain whether it is a true causative variant driving the signal of selection and observed phenotype or a neutral linked mutation, as association studies discover correlation rather than causation (138). Pleiotropic effects create additional difficulties (138, 394); thus it seems crucial to isolate the phenotypic consequences of beneficial variants (often very subtle) from the genetic background that is variable between individuals (138). Non-human animal and cell-culture models coupled with genome editing (e.g. using clustered regularly interspaced short palindromic repeats (CRISPR) with CRISPR-associated protein-9 nuclease (Cas9)) as well as non-cellular experiments seem to be a suitable solution to overcome these difficulties, as they enable isolation of the variant and its direct testing. The obvious limitation of such approach is that variants may behave differently in humans compared with *in vitro* and *in vivo* model systems. Models need to closely replicate the predicted functional impact of the candidate variants. Thus, choosing the appropriate experimental methods is a critical step in such analyses, and will depend on many factors like the class of variant analysed and its biological context; organ, tissue type or process affected by the mutation; sequence and function conservation between human and the modelling system; availability of prior knowledge about variant functionality; predicted phenotype and its effect size; costs and efforts. This chapter presents first some *in vitro* studies, then *in vivo* studies. Contribution of internal and external collaborators is indicated in relevant Methods sections.

3.1. Functional studies *in vitro*

Using non-cellular assays of modified protein-protein interactions or human cell lines engineered to carry the proposed causal variant can be quicker and less laborious than generation of animal models. Furthermore, human cell lines are often the best approximation for modelling of human characteristics, as they guarantee sequence identity (which is often a problem for modelling of regulatory elements in different taxa) and likely functional similarity. However, function might vary from *in vivo* conditions. The limitation of this approach is that simple cultured cell models may be inappropriate for complex whole-organism level phenotypes but can be applied to study cellular phenotypes. A successful example of such validation is the derived G allele at rs12913832, associated with blue eye colour (162, 163). This variant is located upstream of the *OCA2* promoter in a highly conserved intronic sequence that represents a regulatory region controlling expression of *OCA2*, a major contributor to human eye colour variation (162, 163). The derived G allele was shown to decrease expression of *OCA2*, as it binds differently to nuclear extracts in *in vitro* assays in cell cultures (163).

The next two sections present the two examples of *in vitro* analyses performed as part of this work. Each is structured with individual introductory, methods, results and discussion sections.

3.1.1. Positive selection in the human olfactory receptor gene family

3.1.1.1. Introduction

The olfactory receptor (OR) proteins are members of a large family of G-protein-coupled receptors arising from mostly single coding-exon genes that often occur in large clusters in the human genome (395-397). ORs interact with odorant molecules in the nasal olfactory epithelium to initiate a neuronal response that triggers the perception of a smell (395, 398, 399), but might be also involved in other, non-olfactory-related, functions (400). It is also known that point mutations in OR genes contribute to olfactory phenotype diversity in humans and each person has a unique set of genetic variation that leads to enormous differences in olfactory perception between individuals (401-406).

It has been shown that the mammalian OR repertoire have been subjected to rapid evolution, presumably due to species-specific adaptation to the ecological niche (407, 408). Detection of chemical molecules in the proximate environment is informative about toxicity of food sources, habitat parameters and predators, but also helps in individual identification and mate selection and might play a crucial role in the organism's survival (400, 407, 409-411). However, it is commonly assumed that the primate lineage (especially humans) suffered significant gene loss in the OR repertoire and a decline in the importance of the olfactory system (399, 412-418). As much as 60% of the human OR genes are pseudogenes bearing one or more coding-region disruption likely resulting in a functional inactivation (414, 419). While Pierron *et al*. showed that negative selection is still relaxed in human ORs, suggesting that the olfactory capability might still be decreasing (420), others have reported positive selection acting on intact OR clusters and ethnographic variability (397, 419, 421).

Taken together, it seems that some OR genes might not be essential for human survival, but it appears that the general enhancement and diversification of the size of the OR repertoire may confer a selective advantage (397). On the other

hand, a recent study reported no evidence of positive selection on the olfactory receptor repertoire as a whole since the chimpanzee-human divergence (414), but did not rule out the possibility that a few intact OR genes could have experienced selective sweeps and the signature in the combined sample is undetectable (421). The emerging picture shows that whereas most human OR genes are under no or little evolutionary constraint, others might have important functions, and a subset have evolved under positive selection (421) e.g. *OR5I1* (422).

There has been an ongoing debate about whether the selection seen on human ORs was due to smell perception in olfactory epithelium, or to different recognition and signalling functions in other parts of the body. Such questions might be addressed by performing functional follow-up of selected human alleles. However, fast evolution between species makes it difficult to model human derived alleles *in vivo* using available model organisms, as a significant proportion of mammalian ORs are orphan receptors (407, 423). Furthermore, it appears that even with a clear 1:1 ortholog between closely-related species, functional equivalency is limited and sequence does not accurately predict the functional properties of ORs among orthologs in this multi-gene family (407). In such cases, *in vitro* approaches proved to be a reliable predictor of *in vivo* function and odour perception, and have provided insight into the functionality of ORs and their evolutionary history (407). We decided to functionally follow-up OR genes using an *in vitro* approach as we saw multiple signals of selection falling in olfactory clusters, and established collaborations with experts in olfaction biology.

3.1.1.2. Materials and methods

We explored previously-compiled lists of *CMS*, *ΔDAF* and *FineMAV* to search for putatively selected candidate variants falling in ORs. We then followed up experimentally on the strongest example in collaboration with Joel Mainland at the Monell Center (Philadelphia, United States). Mainland *et al*. looked at the activation (ligand specificity and activation strength) of the ancestral and derived version of the protein upon exposure to a chemically diverse odour library using a high-

throughput cyclic adenosine monophosphate (cAMP)-mediated luciferase *in vitro* assay for OR functional testing (407, 423-425). To do so, the coding sequences (full ORF) of the derived and ancestral haplotypes were cloned and expressed in a heterologous cell system (Hana3A cells, an HEK293T-derived cell line stably expressing accessory factors for OR expression (426, 427)). This system enables cell-surface expression of ORs and measuring their activation upon odour stimulation. If examined OR binds the ligand, binding will change the conformation of the ORs and initiate a cascade of signal transduction leading to OR activation, and production and accumulation of cAMP (which turns on the expression of a luciferase reporter gene that is readily quantifiable by luminometrical methods) (425). The ancestral and derived alleles of missense SNPs were then screened against a panel of 918 compounds to compare their dose-responses to individual ligands by testing each allele across a range of concentrations. A detailed description of the methods used in this study can be found elsewhere (401, 407, 423-425).

3.1.1.3. Results

In our database compiled from previous selection scans we found evidence of selection on 10 SNPs falling in 9 OR genes (Table 4). Not all of those signals need to be independent, as European and East Asian variants falling into clusters on chromosomes 1 and 11 respectively are in high LD. The strongest *CMS* signals pointed to two missense variants: rs2240227 in *OR10H3* selected in East Asians (CHB+JPT, HapMap data (123); Figure 23) and rs12273630 in *OR51B5* selected in Africans (YRI, 1000 Genomes Project data (155)) which falls in the OR gene cluster on chromosome 11. *FineMAV* analysis replicated the signal of selection on rs2240227 in *OR10H3* (Figure 23) as one of the strongest hits in East Asians (ranking as 28th in the whole-genome analysis), but picked up another SNP from the chromosome 11 cluster, rs331537 in *OR52K2* ranking 66th in Africans.

We chose to functionally follow up on rs2240227, as it seemed the strongest and most reproducible candidate, whose derived allele is seen at 61% frequency in

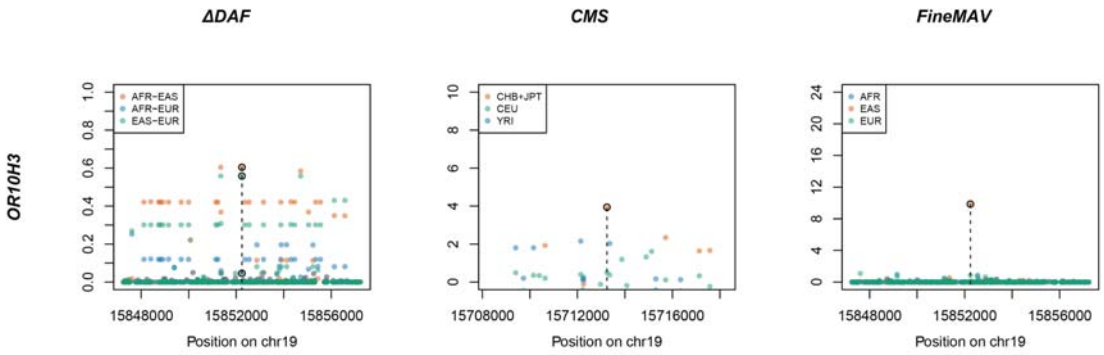

Figure 23. Signal of selection in *OR10H3* according to three different approaches. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic window spanning 10 kb around the variant of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). Normalised *CMS* scores (123) were calculated using the phase II of the International Haplotype Map Project (HapMapII) (146) and downloaded from http://www.broadinstitute.org/. Variants with *CMS* values set to 'nan' or below 0 are not shown. Genomic positions are given in bp according to GRCh37 for *ΔDAF* and *FineMAV*, and build NCBI36 for *CMS*. The selected variant is marked with a dashed line.

East Asia but is rare elsewhere (5% in Europe and 0.7% in Africa) and is predicted to be highly functional (*CADD* score of 22.3, *PolyPhen*: possibly damaging, *SIFT*: deleterious). To measure the functional consequence of rs2240227 polymorphism, we needed to pair *OR10H3* with odorant *in vitro*, as *OR10H3* does not have an identified odorant ligand. We therefore functionally assayed and compared the East Asian derived and reference ancestral haplotypes (different by only 1 amino acid residue at the Leu14Ile substitution; Figure 24) in collaboration with Joel Mainland at the Monell Center (Philadelphia, United States). The receptors showed no response to any stimuli tested. An example of such a negative dose-response is shown in Figure 25.

Table 4. Top-scoring candidates for positive selection in the olfactory receptor family. The 'Method' specifies the test that picked up the given variant. 'Pop.' – population exhibiting the signal of selection. 'Expression' provides information on ectopic expression based on (200, 428, 429); a hyphen indicates no reported ectopic expression.

Gene	SNP	Chr.	Method	Pop.	Consequence	Expression
OR2L2	rs6658141	1	*CMS*	EUR	missense (Val->Leu)	low ectopic
OR2L3	rs6658256	1	*CMS*	EUR	missense (Ser->Leu)	low ectopic
OR1B1	rs1476859	9	*ΔDAF*	AFR	missense (Ala->Thr)	-
OR52K2	rs331537	11	*FineMAV, ΔDAF*	AFR	missense (Arg->His), regulatory (promoter flanking region)	low ectopic
OR51B5	rs12273630	11	*CMS*	AFR	missense (Val->Ile), regulatory (enhancer)	medium ectopic
OR56B4	rs1462983	11	*CMS, ΔDAF*	EAS	missense (Pro->Ser)	low ectopic
OR52W1	rs11040760	11	*CMS, ΔDAF*	EAS	regulatory (enhancer, eQTL)	low ectopic
OR52W1	rs10839531	11	*ΔDAF*	AFR	missense (His->Arg)	low ectopic
OR10AD1	rs4760697	12	*CMS*	EUR	regulatory (promoter)	low ectopic
OR10H3	rs2240227	19	*FineMAV, CMS*	EAS	missense (Leu->Ile)	-

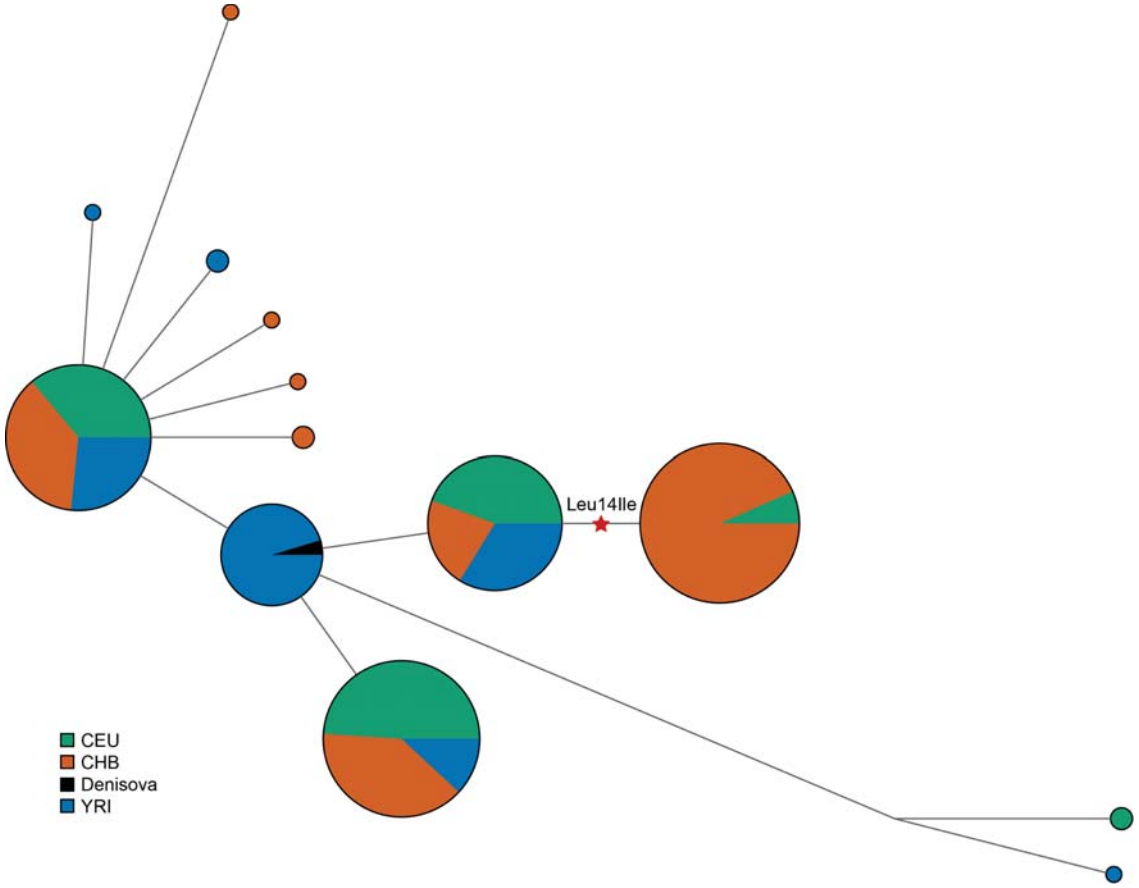

Figure 24. Haplotype network of *OR10H3*. The median-joining haplotype network was generated using Network 5.0.0.0 (430) and sequences of 270 individuals from the 1000 Genome Project, Phase 1 (85 CEU, 97 CHB, 88 YRI) (159) and the high-coverage Denisova genome (246). Each circle represents a distinct haplotype; circle area is proportional to haplotype frequency; the branch length shows number of mutational steps between haplotypes (shortest line equals one step); the selected rs2240227 mutation is marked with a star.

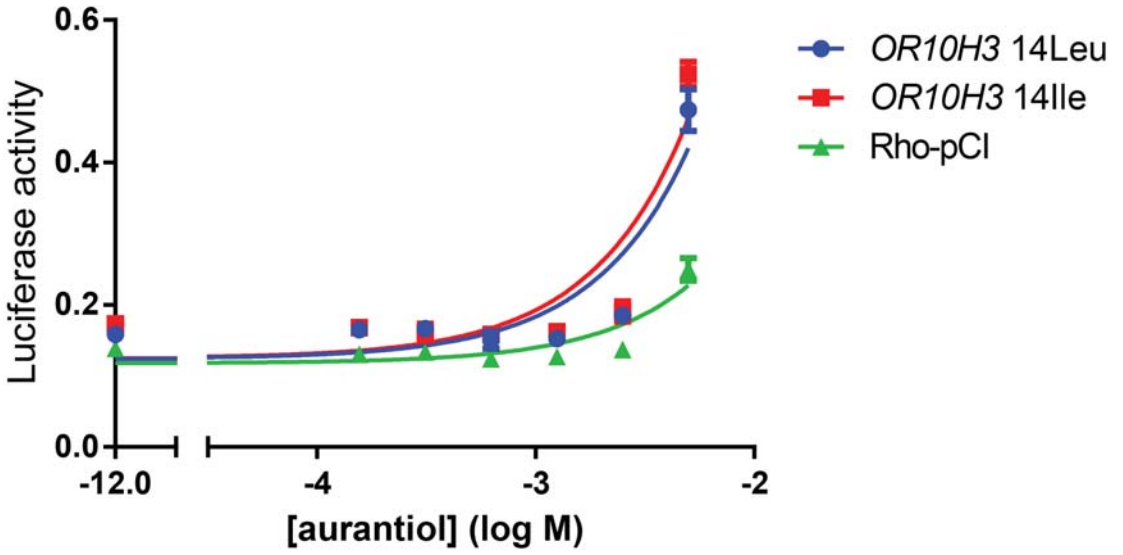

Figure 25. Dose-response curve of the derived (14Ile) and ancestral (14Leu) version of the *OR10H3* receptor to the aurantiol odorant. X-axis shows the concentration of aurantiol in Log Molar units. Y-axis shows the luciferase response (each concentration was tested in triplicate, error bars indicates ± S.E.M. over three replicates). Rho-pCI is a negative vector-only control (mammalian expression vector containing the first 20 amino acids of human rhodopsin (Rho-tag) that was used for OR cloning; the inclusion of the Rho-tag at the N-terminal end has been shown to promote the cell-surface expression of ORs (425)). Responses of cells transfected with a plasmid encoding *OR10H3* should fit the sigmoid curve upon activation with an empty vector showing no response. The odorant did not activate the receptor significantly more than the vector-only transfected control as described in (407).

3.1.1.4. Discussion

3.1.1.4.1. Lack of *OR10H3* activation

It is known that missense substitutions in ORs often alter receptor function and can have dramatic consequences on activation, ligand specificity and odour perception (401-407). We made an attempt to compare functional differences of the ancestral and derived receptor variants of *OR10H3 in vitro* with a diverse panel of odours, but we could not identify any ligands that activated either of the alleles (14Leu or 14Ile) preventing us from comparing their functional impact.

The fact that the receptors investigated did not respond to a panel of odours illustrates the limitations of the *in vitro* assay. Most likely, the relevant odour is simply not present in the panel, but the lack of receptor response might alternatively reflect a failure of the OR to function in the assay (402, 407). In addition, the difference between derived and ancestral allele might affect other aspects then ligand binding, e.g. differences in G-protein coupling or receptor recycling and could be investigated in the future (407).

It is also possible that *OR10H3* is non-functional, as not all ORs with an intact ORF are necessarily expressed and functional (407). Another possibility is that, according to a broader non-classical definition, ORs are small proteins responsible for transduction of a signal upon ligand recognition not necessarily linked to olfaction. It could thus be that OR10H3 does not have an olfactory function at all, but detects non-classical odorants, while our odour space panel is optimised for olfactory response in the human nose.

Such a hypothesis is supported by the observation that majority of the ORs (including *OR10H3*) picked up by the selection scans (Table 4) are poorly expressed in the human nasal olfactory epithelium (based on 3 human samples assayed for RNA expression using a custom NanoString CodeSet; personal communication, Darren Logan, 2015). There is no evidence that these ORs are particularly important for olfactory function, which is consistent with the hypothesis that they could have been under selection for reasons other than smell perception.

3.1.1.4.2. Ectopic expression of olfactory receptors

It has been demonstrated that OR expression is not restricted to the olfactory epithelium and about a quarter of ORs are expressed ectopically (i.e. in non-olfactory tissues), serving extra-olfactory functions, although some ectopic OR transcripts may not code for a functional protein (429, 431, 432). Functionality of ectopic ORs is supported by their strong evolutionary constraint compared to OR genes expressed exclusively in the olfactory epithelium (432). Furthermore, most of the key components of the olfactory signal transduction pathway were detected across many tissues, which suggests the existence of downstream signal transduction in non-olfactory tissues and might indicate the involvement of these gene products in other physiological processes (429) (but it is also possible that activation of ectopically expressed ORs targets other signalling pathways (433)).

Expression analysis of ectopically expressed ORs across multiple human tissues found that some ORs were broadly expressed in a variety of non-olfactory tissues, while others showed exclusive expression in one investigated tissue (such as *OR4N4* and *OR2H1* in testis) (429). OR genes expressed ectopically were more highly expressed in testis that in any other non-olfactory tissue examined, indicating a possible important functional role of ORs in testis (429, 434, 435) e.g. the expression and activation of *OR1D2* is believed to function in human sperm chemotaxis, influencing the swimming direction and speed of spermatozoa, which might be critical in the fertilization process (436-439). Furthermore, 40% of MHC-linked OR-genes were detected in the testis (spermatocytes) (429) and might participate in olfaction-guided mate choice, but also in MHC-dependent selection of the spermatozoa acting as surface chemoreceptors to favour the production of MHC-heterozygous offspring (440). Apart from involvement in chemotaxis, ORs expressed in testis were implied in the sperm development and competition or interaction between spermatozoa and oocytes (429, 440). Nonolfaction-associated OR function such as cell-cell recognition in human embryogenesis has also been suggested (441). Finally, ORs expressed in the human gut mucosa might control gut motility and secretion (433).

Most of the receptors picked out by our study were also reported to be expressed ectopically (Table 4) (200, 428), including *OR10AD1* and *OR51B5* as the

most highly expressed in the human tissues (429). *OR10AD1* and *OR2L2* showed the evidence of being expressed in human testes and could be involved in chemotaxis during fertilization (200, 442). *OR51B5* is a particularly interesting example as it was associated with the fetal haemoglobin (HbF) levels (443) and the observed signal of selection in this gene was reported in Africans. HbF partially compensates for the reduction or absence of normal HbA production in sickle cell anaemia and the β-thalassemias, and its increased level correlates with less severe complications, fewer pain crises and improved survival (443).

Solovieff *et al*. found an association between HbF concentration in sickle cell anaemia and a regulatory region in the olfactory receptor gene cluster (containing *OR51B5* and *OR51B6*) upstream of the β-globin gene cluster on chromosome 11 (443). The authors suggested that this region might play a role in controlling expression within the β-globin gene complex (containing the HbF gene) by altering chromatin structure (establishing and/or maintaining of an open chromatin domain) (443-446). It might be that the rs12273630 picked up by the *CMS* method that falls in the *OR51B5* (but also in introns of *HBE1* (embryonic haemoglobin subunit epsilon) and *HBG2* (fetal haemoglobin subunit gamma-2) and an enhancer) has been selected due to non-olfactory regulatory function. ORs from this cluster are transcribed at low levels in erythroid cells and are characterised by high evolutionary constraint (446, 447).

It was previously thought that the expression of β-globin genes is strictly dependent on a cis-acting element called the locus control region (LCR), that contains erythroid-specific DNase I hypersensitive sites necessary for establishing the open chromatin domain (446). However, it has been shown in mouse ES cells that the chromatin in the β-globin gene cluster remains in an open conformation, even after deletion of the LCR, if the olfactory receptor gene cluster remained intact (although the transcription of β-like globin genes was significantly reduced) (448). This suggests that the OR cluster, together with other elements scattered throughout the locus, might contribute to heterochromatinization independently from the LCR (448). Other GWAS studies have also reported association of the OR gene cluster on chromosome 11 with HbF level and thalassemia severity (449, 450). Furthermore, a DNase hypersensitive site was reported within the OR region (451),

but the functional impact of the olfactory gene locus on downstream globin genes remains uncertain and requires further experimental investigation (443).

3.1.2. Selection on fucosyltransferase 2 (*FUT2*)

3.1.2.1. Introduction

Pathogens have been a powerful selective force during human evolution and numerous host-cell surface molecules recognised as receptors by pathogens have experienced positive selection in humans (452, 453). One source of such strong pressures, historically responsible for high child mortality in developing countries, was rota- and noroviruses. These enteric viruses cause acute gastroenteritis characterised by vomiting, diarrhoea, dehydration and electrolyte imbalance, resulting in over 650,000 deaths per year (prior to the introduction of vaccination programmes) (454-456). Rota- and norovirus attachment to the host cell requires binding to carbohydrates (oligosaccharides) of the ABO(H)/Lewis histo-blood group antigens expressed in epithelial cells of the gut (454, 457-461). The synthesis of the H antigen (precursor of the ABO antigens) on epithelial cell surfaces and in body fluids is regulated by the human secretor locus (*Se*) *FUT2*, encoding alpha-(1,2)fucosyltransferase. Loss of function mutations in *FUT2* result in the non-secretor phenotype i.e. a lack of the FUT2 enzyme activity and a consequent absence of the α1,2-fucose antigen in the intestinal surface mucosa and body fluids, which is associated with resistance to virus attachment and infection (52, 454, 462). Non-secretors can still produce ABO(H) antigens in erythrocytes, as their precursor is encoded by *FUT1* (463).

It has been shown that many independent mutations are responsible for the nonsecretor phenotype around the world (452), and ~20-30% of the worldwide population fail to secrete H antigen (464, 465). The two most common mutations causing the nonsecretor phenotype are the stop-gained variant rs601338, also known as se^{428} (found at high frequencies in Africans (49%) and Europeans (44%) but absent in East Asians) (452, 465), and a missense variant, rs1047781 (se^{385}), found exclusively in East Asians at 44%. The latter results in an Ile140Phe amino acid substitution and was shown to reduce the FUT2 enzyme stability and activity to 2-3%, thus causing almost complete inactivation (195, 464, 466, 467). Therefore, homozygous carriers of the se^{385} missense mutation are sometimes considered

'weak secretors' expressing low levels of H-antigen, as opposed to 'nonsecretors' with the *se*^{428} nonsense mutation who do not secrete H-antigen at all (461, 467). It has been proposed that those two mutations may cause differences in susceptibility to specific viral strains, and that 'low secretor' status provides incomplete protection from noro-/rotaviruses (461).

Many studies have investigated the infection susceptibility of secretors vs non-secretors (468-470). A recent meta-analysis indicated that host genetic susceptibility to norovirus and rotavirus infection is strain-specific, and secretors are ~2-30 times more prone to infection (depending on the virus) compared with non-secretors (461, 471). Non-secretors showed strong although not absolute protection from infections depending on virus carbohydrate-binding profile, as different strains recognise slightly different glycan patterns (461, 471, 472). The strongest infection association with the secretor status was shown for GII.4 noroviruses and P[8] rotaviruses (461, 471). Other beneficial effects of *FUT2* null-alleles have been proposed, including avoidance of the carcinogenic bacteria *Helicobacter pylori* that colonise the stomach through binding to host gastric mucus layer containing H blood group structures (473-479), and a reduced risk of acquiring HIV-1 and a slowed progression of its infection in non-secretors (480-482). The latter could be linked to *FUT2* expression in the epithelial cells of the genitourinary tract (480, 481). In addition, FUT2 has also been shown to be expressed in the epithelial cells of the respiratory tract, and nonsecretor status was associated with a decreased risk of some respiratory viral diseases caused by influenza A and B viruses, rhinoviruses, respiratory syncytial virus, and echoviruses which enter the host via mucosal surfaces (483).

FUT2 activity has also been shown to affect the gut microbiota (species composition, diversity, absolute abundance and host-microbe interactions) and metabolite profiles in adults (484-487). The H antigen is an oligosaccharide that acts both as an attachment site and a carbon source for intestinal bacteria that protects from intestinal overcolonization by opportunistic pathogens and subsequent inflammatory diseases (485, 488-491). Non-secretors have an altered functional composition of mucosal microbiota which puts them at increased risk of developing inflammatory bowel disease (IBD) (492), including Crohn's disease

(485, 486, 493-496) and ulcerative colitis (497), but also primary sclerosing cholangitis (PSC) (498-500) and celiac disease (501).

Various studies have reported signatures of balancing and positive selection at the *FUT2* locus (452, 464, 502-504). The East Asian nonsecretor mutation (rs1047781) was proposed to have experienced a recent drastic increase in frequency likely due to strong positive selection in agreement with the reduction of genetic diversity and distortion of SFS (452). In this section we attempted to functionally follow up the selected *FUT2* variant picked up in our study.

3.1.2.2. Material and methods

We aimed to establish a stable cell line expressing exogenous ancestral and derived form of *FUT2* with no endogenous background expression. All molecular biology work was done by Carmen Diaz Soria (Paul Kellam's Viral Genomics group at the Wellcome Trust Sanger Institute). Cell culture work was performed by Carmen Diaz Soria and me. Western blot analyses were carried out by Elena Arciero (Wellcome Trust Sanger Institute) and me.

3.1.2.2.1. Construct design with GeneArt and Site-Directed Mutagenesis

Human DNA ancestral and derived *FUT2* sequences were synthesised by GeneArt. Constructs were made in duplicate carrying C-terminal HA or Myc tags. To mask an extra BamHI restriction site in the *FUT2* constructs, we carried out site-directed mutagenesis using the QuikChange II XL site-directed mutagenesis kit (Agilent) and primers (Metabion) shown in Table 5 under conditions shown in Table 6. These GeneArt plasmids were then transformed into NEB Turbo competent cells (New England Biolabs) according to the manufacturer's guidelines. Cells were spread onto ampicillin LB agar plates and incubated overnight at 37^{o}C. Single ampicillin-resistant colonies were picked from each LB agar plate and used to

inoculate 5 ml of LB medium in a 50 ml falcon tube. Cultures were left overnight in a shaking incubator at 37°C and 200 revolutions per minute (Innova44, New Brunswick Scientific). The culture was centrifuged (10,000 g, 5 min) and the DNA extracted (QiaPrep spin mini-prep kit, Qiagen) following the manufacturer's protocol. We checked the colonies using restriction enzyme digests with *Bam*HI and *Not*I (Promega, UK) followed by running products on an agarose gel. Digestion reactions were carried out in 20 μl and incubated at 37°C for 1 h under reaction conditions according to the Promega protocol.

Table 5. Primers used in this study.

Name	Sequence 5'-3'	Usage	Manufacturer
SFFV_F	TGCTTCTCGCTTCTGTTCG	Sequencing pHR SIN CSGW-PGK PURO	Sigma-Aldrich
WPRE_R	CCACATAGCGTAAAAGGAG	Sequencing pHR SIN CSGW-PGK PURO	Sigma-Aldrich
c425a_fut2_F	CCACGGCCAGCAGGATACCCTGGCAG	Site Directed Mutagenesis	Metabion
c425a_fut2_R	CTGCCAGGGTATCCTGCTGGCCGTGG	Site Directed Mutagenesis	Metabion

Table 6. PCR cycling conditions for Site Directed Mutagenesis.

Cycles	Temperature [°C]	Time
1	95	1 min
	95	50 secs
18	60	50 secs
	68	7 min + 10 secs
1	68	7 min

3.1.2.2.2. Construction of plasmids

The insert was removed from each GeneArt plasmid and ligated into a lentivirus expression vector, pHR-SIN CSGW PGK Puro. First, the GeneArt plasmids as well as the expression vector were digested using the restriction enzyme *Bam*HI and *Not*I (Promega, UK) as described above. The digestion products were run on an

agarose gel and the DNA fragment corresponding to the *FUT2* gene construct (~1.1 kb) and the expression vector (~9 kb) were extracted using a QIAquick Gel Extraction Kit (Qiagen) following the manufacturer's instructions. QIAgen-purified *FUT2* DNA fragments were cloned into pHR-SIN CSGW PGK Puro by ligation. All ligation reactions were carried out at a 3:1 molar insert:vector ratio in a 20 µl reaction volume and left overnight at 16^{o}C.

The ligation mix (pHR-SIN CSGW PGK Puro + *FUT2* gene) was transformed into NEB Turbo competent cells (New England Biolabs) as described in the previous section. The colonies were checked by colony PCR using the conditions described in Table 7. Sanger sequencing (GATC Biotech) was used to check the integrity of these sequences and ensure that the tag was in-frame with the rest of the protein sequence. The DNA sequence was amplified using SFFV_F and WPRE_R primers (Sigma-Aldrich) shown in Table 5.

Table 7. PCR cycling conditions for colony PCR and to generate DNA for Sanger sequencing.

Cycles	Temperature [^{o}C]	Time
1	98	30 secs
	98	10 secs
30	54	30 secs
	72	35 secs
1	68	10 min

3.1.2.2.3. Making lentivirus stocks

Lentivirus particles were constructed according to an in-house protocol using a gag-pol expressing vector (p8.91), a VSV-G expressing vector (pMDG) and the above vector expressing *FUT2* (pHR-SIN CSGW PGK Puro). Briefly, 10 µl Fugene-6 (Roche) was added to 200 µl Opti-MEM (ThermoFisher). A DNA mix carrying 1 µg gag-pol expresser (p8.91), 1 µg pMDG (VSV-G expresser) and 1.5 µg expression vector (pHR-SIN CSGW PGK Puro + FUT2) was made in 15 ul in TE (10 mM TRIS pH 8, 1 mM EDTA) and added to the Opti-MEM/Fugene-6 mixture. The mixture was left at room temperature for 15 minutes. This DNA mix was then added dropwise to HEK-293T cells that had been plated the day before in 10 cm plates. Cell were

returned to the incubator at 37°C in 5% carbon dioxide in air mixture. The next day, the medium was changed and 8 ml of fresh medium added. The supernatant containing the lentivirus particles was collected after 48 hrs, filtered with 0.45 μm filters and stored at -80 °C.

3.1.2.2.4. Production of stable cell lines and cell culture

We plated 6 x 10^5 A549 cells/ml in 6-well plates in duplicate. The next day, the cells were transfected with 500 ml of lentiviral particles, except for the control untransfected cells. A medium change was performed after 48 hrs. Cells were exposed to the selection antibiotic, Puromycin (1.4 mg/ml) at a 1/1000 dilution 4 days post-transfection. Medium containing the selection antibiotic was replaced every 3 days. Cell lines were grown in F12 (Invitrogen) supplemented with 10 % v/v foetal bovine serum (FBS, Biosera). Cells were passaged 1:6 or 1:10, twice a week.

3.1.2.2.5. Western blotting

Proteins were extracted from cell cultures using radioimmunoprecipitation assay buffer (RIPA Buffer; R0278 SIGMA) containing Halt™ Phosphatase Inhibitor Cocktail (78420B; Thermo Scientific) following the manufacturer's protocol. Protein samples were then mixed with Protein Loading Buffer Blue 2X (EC-886; National Diagnostics) and loaded into wells alongside the Precision Plus Protein™ Kaleidoscope™ Prestained Protein Standards (#1610375; BIO-RAD) to be separated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE). Electrophoresis was carried out using 4–20% Mini-PROTEAN® TGX™ Precast Protein Gels (#4561093; BIO-RAD) in a Mini-PROTEAN Tetra Cell system (BIO-RAD) containing electrophoresis buffer (1x PBS and 0.05% Tween 20 (Sigma)). Proteins were then transferred onto nitrocellulose membrane (Trans-Blot® Turbo™ Mini Nitrocellulose Transfer Packs; #1704158; BIO-RAD) using Trans-Blot® Turbo™

Transfer System (7 mins, 25V; BIO-RAD). Membranes were incubated with primary antibodies: goat polyclonal to *FUT2* (ab177239; Abcam) or Mouse monoclonal [AC-15] to beta Actin (HRP) (ab49900; Abcam) for 2 hours, followed by incubation with the following species-appropriate secondary antibodies for 1 hour: Donkey Anti-Goat IgG (6420-05; SouthernBiotech) or Polyclonal Goat Anti-Mouse IgG (Dako). All antibodies were diluted in 1x PBS (Sigma) containing 0.05% Tween 20 (Sigma) and 5% non-fat dried milk (Carnation). Proteins were then visualised using the ECL™ Prime Western Blotting Detection Reagent (RPN2236; GE Healthcare) following the producer's instructions, and the Celvin® S chemiluminescence imaging system (Biostep). Images were captured with SnapAndGO software.

3.1.2.3. Results

Our *FineMAV* analysis in 1000 Genomes Project, Phase 3 (142) picked up the known 'weak' secretor mutation rs1047781 as the 21st highest scoring variant in East Asians (Figure 26). The molecular functionality of this variant and its impact on noro- and rotaviruses susceptibility is well documented (195, 464, 466, 467), but the hypothesised selective advantage of the low-/inactive enzyme in the resistance to other viral infections has not been directly measured *in vitro*. We aimed to express the ancestral and low-activity derived forms of the FUT2 enzyme in A549 cells, a cell line that does not express the endogenous copy of this gene, and establish a stably-transfected cell lines. We intended to assess the cell fucosylation level associated with each allele using anti-fucose staining as described in (505) and their susceptibility to a range of viruses (2 strains of influenza A virus subtype H5 from Vietnam and Indonesia, and 5 Ebolaviruses: *Bundibugyo ebolavirus*, *Reston ebolavirus*, *Sudan ebolavirus* and *Zaire ebolavirus* including Mayinga and Mayinga M2 strains) using the luciferase pseudovirus infection assay that, knowing rs1047781 causality, seemed as a low-hanging fruit.

We successfully transfected cells with the lentiviral vector and detected transient *FUT2* expression of all four variants: ancestral HA-tagged, derived HA-tagged, ancestral Myc-tagged and derived Myc-tagged. We then isolated stably-

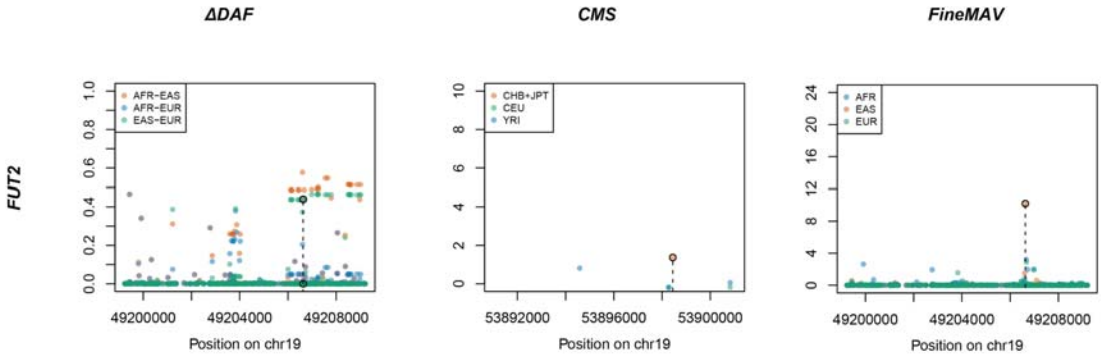

Figure 26. Signal of selection in *FUT2* according to three different approaches. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic window spanning the *FUT2* gene. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). Normalised *CMS* scores (123) were calculated using phase II of the International Haplotype Map Project (HapMapII) (146) and downloaded from http://www.broadinstitute.org/. Variants with *CMS* values set to 'nan' or below 0 are not shown. Genomic positions are given in bp according to GRCh37 for *ΔDAF* and *FineMAV*, and build NCBI36 for *CMS*. The selected variant is marked with a dashed line.

transfected cells (successful integration of the vector into the genome) using selection medium. Only Myc-tagged lines passed the antibiotic selection and were expanded and maintained under the selection regime. However, we failed to establish a stable cell line expressing the transgene, as the *FUT2* expression was lost over time (after ~10 passages corresponding to ~1.5 month) (Figure 27) and we could not test for susceptibility to a panel of viruses.

3.1.2.4. Discussion

The genomic integration of engineered transgenes is a standard genetic manipulation of mammalian cells that has been applied to investigate *FUT2* function in the past. For instance, overexpression of exogenous *FUT2* in the human HuH-7 cell line (hepatocarcinoma) was shown to enhance norovirus binding (506). Enzymatic activity of different FUT2 variants was evaluated in CHO-K1 (Chinese hamster ovary) or COS (Kidney from African green monkey) cells transfected with expression vectors carrying different version of the human gene (195, 467, 507). Both experiments were carried out in transiently-transfected cells. Rotavirus binding to fucosylated cells was, on the other hand, shown in a *FUT2* positive stably-transfected CHO cell line (460).

An isolated stably transfected cell clone should ideally express the transgene at constant level over prolonged period of time (508). However, a complete loss of the transgene expression over time is quite common, despite the successful integration of the expression vector into the genome and its integrity (508). Such a phenomenon is often attributed to epigenetic downregulation of transgene activity in cell cultures, which critically depends on the integration site and its chromatin environment (508). Expression of the antibiotic resistance marker does not guarantee persistent expression of the gene of interest even when placed in close proximity as they are independent transcription units (508). It has been shown that progressive transcriptional silencing of the gene of interest may occur over propagation in the presence of antibiotic selective pressure, as cells with an intact antibiotic resistance gene but a disrupted/silenced gene of interest have a slight

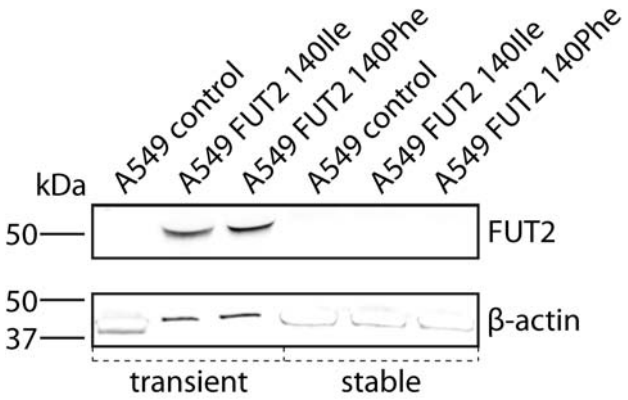

Figure 27. Western Blot analysis cropped to show regions of interest. Lysates of transient and stable transfectants of A549 cells carrying Myc-tagged FUT2, ancestral (140Ile) or derived (140Phe) vector, were separated by SDS-PAGE and transferred onto nitrocellulose membranes. Plain A549 served as a negative control. Predicted molecular mass under SDS-PAGE reducing conditions: FUT2 ~50 kDA; β-actin ~42 kDa.

competitive growth advantage over the protein-producing cells, and tend to dominate the culture (509). A way to overcome this problem could be using a transfection vector where *FUT2* and the antibiotic resistance gene are driven from the same promoter or applying more stringent antibiotic-based approaches for cell line selection or selection techniques based on expression of the gene of interest itself like the one described in (508). An alternative solution, although more laborious, would be editing of the endogenous copy of the *FUT2* via CRISPR/Cas9 in ES cells without introducing exogenous DNA prone to epigenetic modifications. Such an approach would also ensure the same *FUT2* copy number per cell and homogenous expression in the cell population.

3.2. Functional studies *in vivo*

3.2.1. Introduction

Model organisms have been extensively used in a variety of biological research to identify and characterise disease-gene associations, candidate gene function, their pathway involvement and expression patterns (510). This approach is based on the widely-accepted assumption that orthologous genes usually perform equivalent or identical functions across taxa, which allowed translation of animal research into human health applications (511-514). For instance, gene knock-outs in inbred mouse strains have been widely applied to study the function of human many genes (515, 516). Similarly, several human diseases and pathological variants have been successfully modelled in mice, which share 99% of their genes with humans, and are the only mammal whose genome can be efficiently manipulated on a large scale (138, 515, 517). Model organisms also provide an opportunity to assess phenotypic impact at the whole-organism level, thus enabling phenotyping of more complex traits like behaviour, hearing or cold resistance, which are difficult to measure at the cellular level.

Modelling of non-pathological human genetic variation, however, has received little attention to date. Nevertheless, mouse knock-outs have been successfully used to study human-specific evolutionary adaptations like fixed loss of function mutations in *MYH16* (resulting in the reduction of the masticatory apparatus), and *CMAH* (loss of the enzyme due to immune-related selection) (518-520). There are, however, only two reports of successful modelling of human adaptive alleles (humanized knock-in models) in mouse: the human-specific form of *FOXP2* (521) and a population-specific allele in *EDAR* (138). The derived G allele at rs3827760 in *EDAR* causes an amino acid change that is widespread in East Asian populations (up to 93%) and virtually absent in Africans and Europeans (1%) (138). A mouse model carrying the derived allele recapitulated the associated human phenotype of increased hair thickness and sweat gland number, proving causality of the point mutation (138). This study demonstrated the suitability of the

mouse model for isolation and characterization of subtle phenotypic effects of a human adaptive allele in a genetically homogeneous background when the conservation of protein and target organ function between the two species is high enough (138). Non-mammalian models have also been successfully applied to study human adaptations, e.g. melanosomal differences between the ancestral and derived alleles of *SLC24A5* were successfully assayed using a zebrafish model (139). Although non-mammal animals might offer a faster and cheaper way of characterizing the biological consequences of the putatively selected mutation *in vivo*, this approach can only be applied to study basic functions of conserved one-to-one orthologues.

There are, nevertheless limitations to model organism studies as the conclusions depend critically on the appropriateness of the model system for the human phenotype (138). It has been shown that specific groups of genes and regulatory elements have undergone more rapid evolution than others, e.g. a large fraction of enhancers have changed their activity in the human lineage (including human-specific activity gains and losses) (518, 522, 523). Lack of the sequence and/or target organ homology between mouse and human makes it difficult to model human-specific derived variants in genes/*cis*-elements subjected to rapid evolution (e.g. related to the immune or central nervous system and other human-specific aspects of biology). On the other hand, even human-specific traits have been successfully modelled in mice. Enard *et al*. introduced two nonsynonymous substitutions in the *FOXP2* gene that have been fixed specifically on the human lineage, probably due to effects on aspects of speech and language, into the endogenous mouse orthologue (521). As a result, subtle changes in the central nervous system that might be linked to some aspects of speech and language in humans have been found in the mouse model (521).

However, for both *FOXP2* and *EDAR* variants, the selected human phenotypes were reliably suggested by mouse knock-outs or naturally-occurring human loss-of-function mutations. An absence of such prior knowledge makes it difficult to predict and experimentally validate the selected advantageous phenotype associated with a variant in a gene whose function remains unclear, as it is impractical to assay every trait in every cell type (155). Therefore, formulating a prior hypothesis about the selected phenotype based on biological insights

extracted from publically-available sources (e.g. detailed single-gene studies) or conducting novel knock-out studies to improve our understanding of gene function seems to be fundamental to future phenotyping efforts. Models with disrupted gene/regulatory element might show a phenotype that helps to localise the affected function and physiological system that has been the target of recent positive selection, as the single point mutation would perhaps have a subtle effect, but probably one related to the knockdown phenotype. Comparison of animal knock-out phenotypes and naturally occurring human loss-of-function phenotypes might prove to be very informative for decision-making, e.g. modelling of a human adaptive variant by changing one nucleotide in a mouse orthologue that exhibits undetectable knock-out phenotype might seem a risky undertaking better avoided. If direct experimental data on the gene function is unavailable, making predictions about the selected phenotype might be facilitated by the expression pattern and function of orthologous or paralogous genes, protein interactions, pathway involvement and co-localization, expression patterns and disease associations.

3.2.2. Materials and methods

3.2.2.1. Candidate variant selection for *in vivo* studies

Candidate variants chosen for functional studies in mice had to meet several modelability criteria. They had to be strong candidates for positive selection characterised by high *FineMAV* scores, ideally, supported by *CMS* or *SSI*. They had to fall in genes with a 1-to-1 human-mouse ortholog of at least 70% reciprocal amino acid identity. The local alignment of the DNA sequence surrounding the putatively selected mutation had to be characterised by high conservation between human and mouse. They had to fall in genes of known functionality that allowed formulation of a prior hypothesis about the reasons for selection and predictions of the selected phenotype. The selected candidate variant's function was therefore extensively annotated and assessed using publicly-available high-throughput functional genomic and phenotypic databases (Table 8). We overlapped clinical features observed in naturally occurring loss-of-function mutations in humans,

Table 8. List of databases used for functional annotation of candidate variants and genes.

Database	URL
Ensembl	www.ensembl.org
Human Gene Mutation Database	www.hgmd.cf.ac.uk
Clinical Genomic Database	research.nhgri.nih.gov/CGD
Online Mendelian Inheritance in Man	www.omim.org
Catalog of Published Genome-Wide Association Studies	www.ebi.ac.uk/gwas
Expression Atlas	www.ebi.ac.uk/gxa
The Genotype-Tissue Expression (GTEx) Project	www.gtexportal.org
GENCODE	www.gencodegenes.org
Multiple Tissue Human Expression Resource	www.muther.ac.uk
GenCord Project	ega-archive.org/dacs/EGAC00001000105
GENe Expression VARiation	www.sanger.ac.uk/resources/software/genevar
Genetic European Variation in Health and Disease (GEUVADIS)	www.geuvadis.org www.ebi.ac.uk/Tools/geuvadis-das
Mouse Genome Informatics	www.informatics.jax.org
International Mouse Phenotyping Consortium	www.mousephenotype.org
WTSI Mouse Resources Portal	www.sanger.ac.uk/mouseportal
Zebrafish Mutation Project	www.sanger.ac.uk/resources/zebrafish/zmp

with mouse and zebrafish null phenotypes annotations and performed PubMed searches to get insights into the function of genes showing signatures of positive selection. In cases where there was no clue to the candidate gene's function, but it nevertheless seemed of interest, a request for generation of a knock-out mouse was initiated to improve our understanding the selected gene's role in the organism.

Within this framework, we chose to model variants ranging from 'safe' choices with a strong prior expectation about the phenotype to more risky ones where the phenotype was essentially unknown, including several non-synonymous variants, but also some synonymous or non-coding ones.

3.2.2.2. Mouse strain generation and phenotyping

The mouse line generation (including mutation design, mutagenesis, transgenic technologies to transfer the allele into the germ line, genotyping and quality control), colony management, primary phenotyping (Appendix C) and parts of the secondary phenotyping (if needed) were/are to be entirely performed by the Wellcome Trust Sanger Institute Mouse Pipelines (institute core facility: www.sanger.ac.uk/science/groups/mouse-pipelines) upon our request. The genome editing technology employed for the generation of the new strains (both deletion alleles and point mutations) was initially blastocyst microinjection of targeted mutant mouse embryonic stem cell (mESC), then CRISPR/Cas9-mediated mutagenesis by single-cell zygote cytoplasmic microinjection, both in the C57BL/6N background. The progeny derived from the microinjection experiment was bred to allow the transmission of the mutations into the germline of the F1 mice that were genotyped by either end-point PCR or real-time qPCR to demonstrate that the desired allelic structure had been produced. Mice were then bred to homozygosity (if viable in this stage) and sufficient numbers for phenotyping. The standardised primary phenotyping, encompassing a set of phenotypic tests covering more than 600 clinical parameters, is being applied to cohorts of 7 mutant males and 7 mutant females for each of the mutant strains and matched controls (7 males and 7 females per week). This high-throughput screen can be divided into 3

general categories: developmental, *in vivo* (reproduction, infection and immunity, musculoskeletal system, metabolism and endocrinology), and necropsy and blood analysis. A full list of tests performed as of January 2016 can be found in Appendix C.

3.2.3. Results and discussion

3.2.3.1. Knock-outs

We generated new knock-out mouse lines for highly scoring candidate variants falling in genes with no prior knowledge of their functionality. The rationale behind this strategy was that turning off the activity of a mouse ortholog might help to assess what biological systems were targeted by positive selection in humans. Our prioritization resulted in production and phenotyping of 6 mouse knock-out strains that lack genes showing signatures of local adaptation in humans in order to improve our understanding of their function and thus aid in formulating hypotheses about possible selected phenotypes (Table 9).

All 6 knock-outs were generated by CRISPR/Cas9 mediated critical exon deletion. Each knock-out mouse colony is undergoing primary phenotyping and if needed, a detailed secondary phenotyping will be performed. Four of the mutant lines (*Cpsf3l*$^{-/-}$, *Gpatch1*$^{-/-}$, *Herc1*$^{-/-}$ and *Prss53*$^{-/-}$) are at the primary phenotyping stage, but this has not yet been completed. A description of the selected candidates and the signature of their selection, as well as preliminary results of the primary

Table 9. List of mouse knock-out strains generated in this study. All human-mouse orthologue pairs shortlisted here are 1-to-1 orthologues. 'Top SNP' lists the SNP with the highest *FineMAV* score in the given gene, which is most likely driving the signal of selection in humans (although it is not being currently modelled for all knock-out lines); * in the case of *HERC1* the second-highest scoring SNP is given; 'Consequence' and '*FineMAV*' specify properties of Top SNPs; 'Pop.' – population with the signal of selection; '*SSI*' – Selection Support Index for each gene; 'Orthologue identity' – percentage of the mouse protein sequence matching the human protein sequence / percentage of the human protein sequence matching the mouse protein sequence; 'Stage' – current stage of each line: MI – micro-injection, PP – primary phenotyping.

Gene	Top SNP	Consequence	*FineMAV*	Pop.	*SSI*	Orthologue identity	Stage
CPSF3L	rs12142199	synonymous	11.07	EUR	0.08	94/95	PP
GPATCH1	rs10421769	missense	7.15	EUR	0.04	84/84	PP
HERC1	rs2255243*	missense	6.59	EAS	0.49	97/97	PP
LRRC36	rs8052655	missense, regulatory	16.28	AFR	0.22	81/80	MI
MAGEE2	rs1343879	stop gained	23.01	EAS	0.04	83/83	MI
PRSS53	rs11150606; rs201075024	missense; missense	13.66; 10.91	EAS; SAS	0.09	81/81	PP

phenotyping and discussion of the selection hypotheses are given for each gene separately in the next section.

3.2.3.1.1. Selected candidates

CPSF3L

CPSF3L (cleavage and polyadenylation specific factor 3-like) is the catalytic subunit (INTS11) of the Integrator complex, a protein complex containing at least 12 components, that is responsible for meditating the 3-prime end processing (cleavage and polyadenylation) of small nuclear RNAs U1 and U2, thereby affecting many biological processes (524). CPSF3L belongs to a superfamily of zinc-dependent β–lactamase fold proteins and functions as an RNA-specific endonuclease (524). Depletion of CPSF3L by RNA interference (RNAi) results in disrupted formation of the Integrator complex (525) and the arrest of HeLa cells in early G1, but does not prevent cell growth (524). This observation suggests that CPSF3L might be involved in the maturation of cellular pre-mRNAs encoding proteins required for cell cycle progression and entry into S phase (e.g. replication-dependent histone pre-mRNAs), but its precise cellular role remain unknown (524). Furthermore, it has been shown that *CPSF3L* is highly conserved from plants to humans and the suppression of the *Caenorhabditis elegans* orthologue by RNAi leads to an early lethal phenotype, while disruption of the mouse Integrator complex causes growth arrest in early blastocyst embryos (526). Another study revealed that INTS11 plays a critical role in the differentiation of pre-adipocytes and its expression level is increased during the process of differentiation into mature adipocytes, while being reduced to basal levels after the completion of differentiation (525). INTS11 silencing using siRNAs (small interfering RNAs) markedly inhibited adipose differentiation (525). Knock-down in zebrafish embryos led to impaired red blood cell differentiation, also implying its role in cell differentiation (527). Expression analysis found it expressed across various human tissues with the highest level in brain and uterus (200).

We found a strong signature of selection on rs12142199, ranking 5th in Europeans (Table 9 and Figure 28 upper panel). The signal has been also replicated by *CMS* and is supported by a few published studies. Although this variant seems to be synonymous according to well-supported transcript models (Transcription Support Level (TSL) = 1; GENCODE), it has been also reported to be an eQTL driving expression of *CPSF3L* (MuTHER and Geuvadis RNA sequencing project) (229, 528). We requested generation of *Cpsf3l* null allele in mouse that turned out to be recessive lethal (no homozygote embryos detected at E14.5 out of 33 collected). Primary phenotyping of heterozygotes is in progress.

GPATCH1

The *GPATCH1* (G patch domain containing 1) product is one of the proteins found in the catalytically-competent form of the spliceosome (C complex) (529). In eukaryotes, the spliceosome mediates the removal of introns from nascent transcripts (529). However, little is known about GPATCH1 function. Variants in this gene have been found to be associated with the bone mineral density, heel bone properties and risk of fracture, which classified *GPATCH1* as osteoporosis susceptibility gene (530, 531).

We found a strong *FineMAV* signal at a missense variant (rs10421769) ranking 55th in Europeans (Table 9 and Figure 28 lower panel). Selection on this variant was also supported by *CMS*. Interestingly, rs10421769 falls in a region proposed to be adaptively introgressed from an archaic source, and its derived allele was present at homozygous state in Neanderthals (Figure 17) (30, 129, 134, 245). Additionally, it has been shown that its derived form was already present in a 7,000-year-old Mesolithic European hunter-gatherer together with ancestral pigmentation alleles (532). However, and interestingly, the alleged GPATCH1 role in the immune system reported in this study (532) seems to arise from a confusion of a former *GPATCH1* name (*ECGP*, evolutionarily conserved G patch domain containing) with Ecgp (endothelial cell glycoprotein) encoded by a different gene and reported as receptor for OmpA expressed by *E. coli* (533). In the light of GWAS studies, it could be hypothesised that the candidate *GPATCH1* variant may

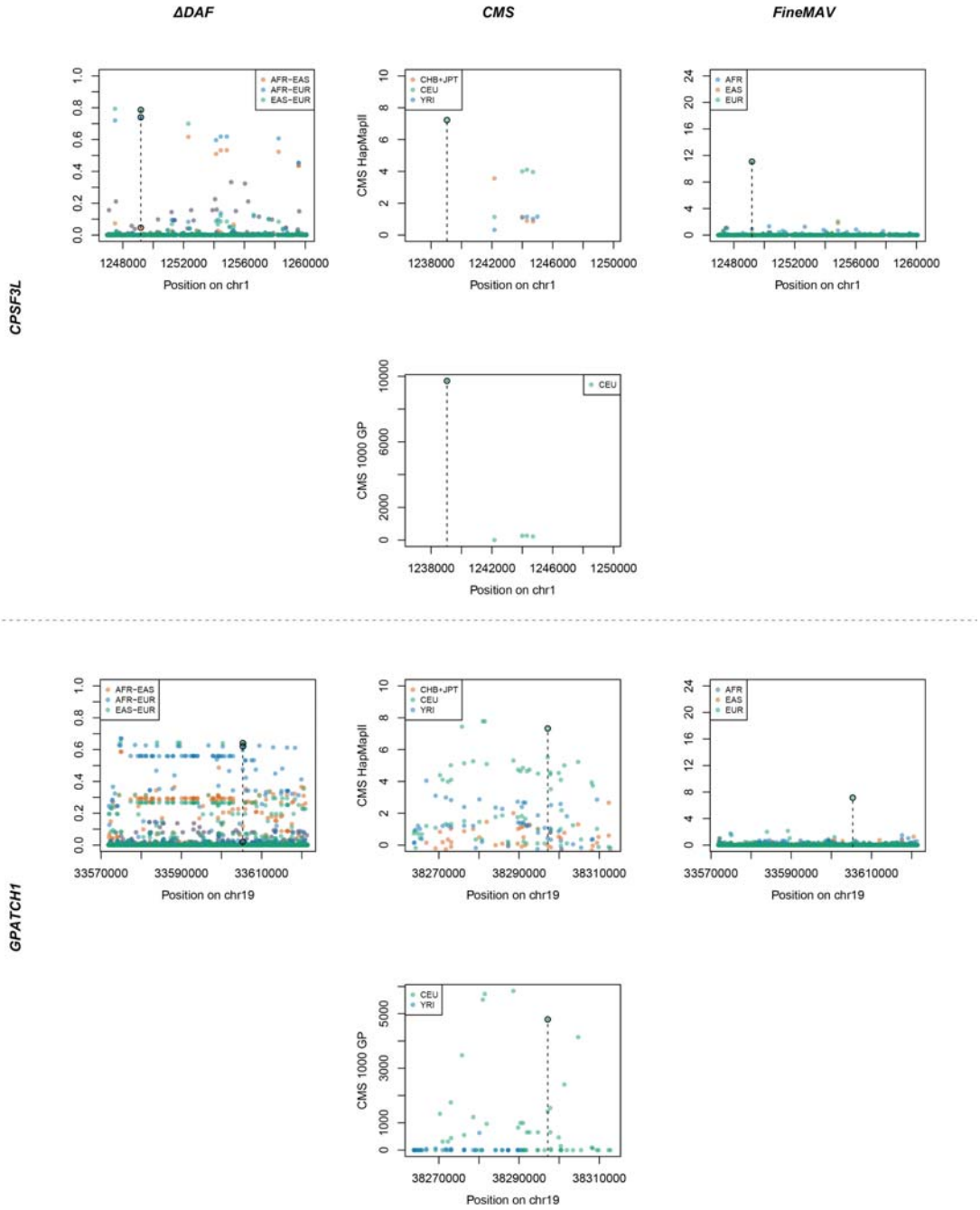

Figure 28. Comparison of three different approaches for pinpointing selected variants. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). *CMS* scores are given for booth, the pilot phase of 1000 Genomes Project (155, 185) and the phase II of the International Haplotype Map Project (HapMapII) (123, 146), all downloaded from http://www.broadinstitute.org/. 1000 Genomes Project CMS scores included windows named: region1new and region42new spanning *CPSF3L* and *GPATCH1* respectively. Variants with *CMS* values set to 'nan' or below 0 are not shown. Genomic positions are given in bp according to GRCh37 for *ΔDAF* and *FineMAV*, and build NCBI36 for *CMS*. The selected variant is marked with a dashed line. Note that the y-axis scale in the *CMS* plots is not standardised.

contribute to increased bone density and might have been selected preceding changes in skin pigmentation, as light skin pigmentation was not yet ever-present in Europe during Mesolithic times (532). Decreased vitamin D synthesis associated with dark skin pigmentation coupled with inadequate sunlight exposure at higher latitudes and a subsequent impaired bone mineralization might have accounted for the strong selective pressure during Mesolithic times that favoured variants increasing bone density. Since such hypothesis is purely speculative and expression analysis found *GPATCH1* expressed across various human tissues with the highest levels in brain, ovary and uterus (200), selective pressures could alternatively have operated on tissues other than bone.

Similarly to *Cpsf3l*, the homozygous *Gpatch1* mouse knock-out also results in embryonic lethality (no homozygote embryos detected at E14.5 out of 36 collected) and primary phenotyping of heterozygous colonies is in progress. Obtained results suggest that both *CPSF3L* and *GPATCH1* genes are crucial at the early stages of organism's development as their homozygous knockout in mouse causes embryonic lethality. However, such phenotype is not informative about putative reasons for selection. Ongoing phenotyping of heterozygotes might shed more light on *CPSF3L* and *GPATCH1* functionality by assessing biological parameters affected by their heterozygous deficiency.

LRRC36

One of the strongest *FineMAV* signals (3rd top hit in Africans) localised to a missense rs8052655 falling in *LRRC36* (leucine rich repeat containing 36) and a promoter flanking region (Table 9 and Figure 29). This gene has been repeatedly reported in selection screens and was shown to be highly expressed in human testis (200), although its function is largely unknown. Knock-out line targeting mouse *Lrrc36* is currently at the micro-injection stage.

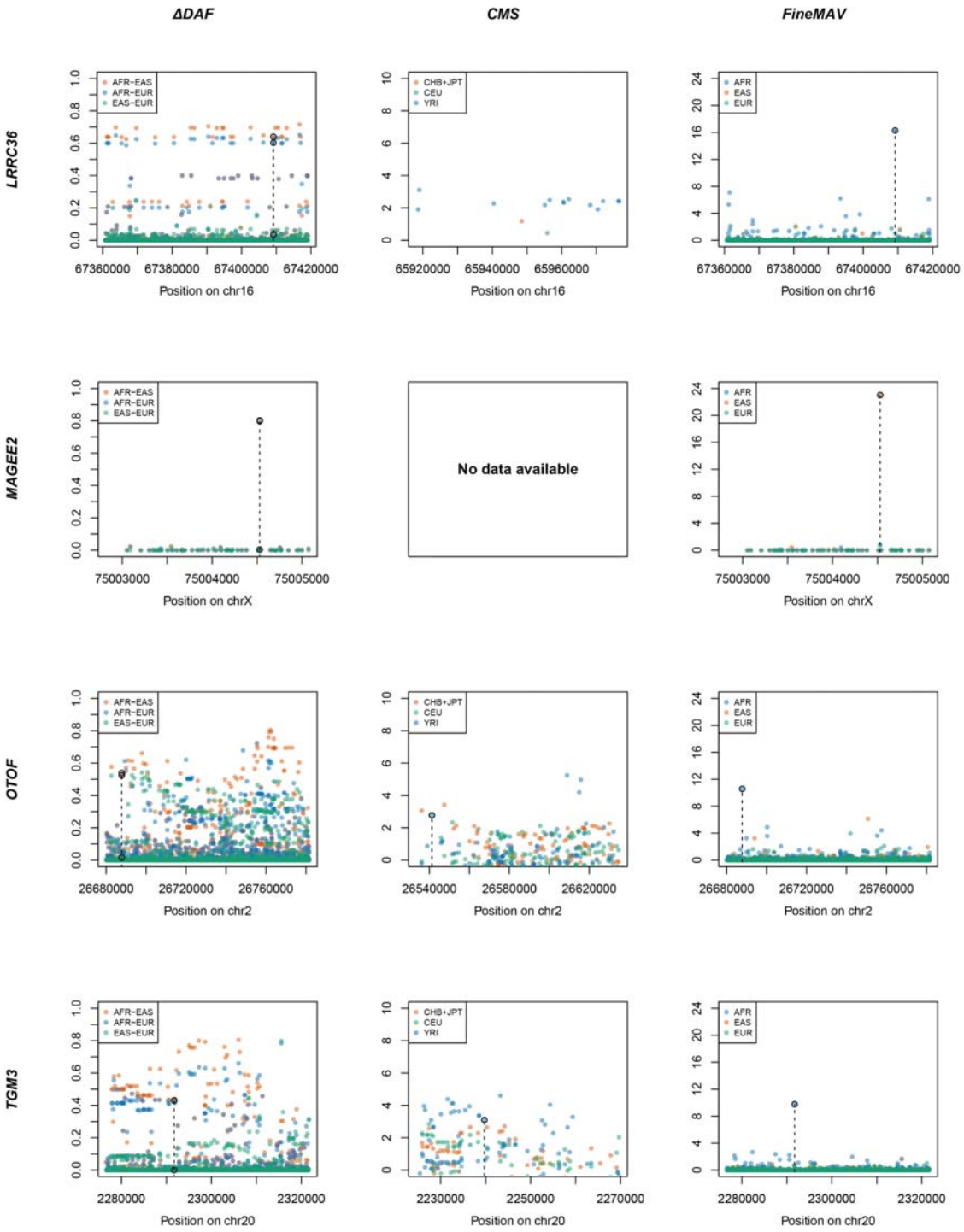

Figure 29. Comparison of three different approaches for pinpointing selected variants. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). Normalised *CMS* scores (123) were calculated using the phase II of the International Haplotype Map Project (HapMapII) (146) and downloaded from http://www.broadinstitute.org/. Variants with *CMS* values set to 'nan' or below 0 are not shown. *CMS* data for sex chromosomes is unavailable, therefore CMS scores for *MAGEE2* are missing. Genomic positions are given in bp according to GRCh37 for *ΔDAF* and *FineMAV*, and build NCBI36 for *CMS*. The selected variant is marked with a dashed line.

MAGEE2

The highest observed *FineMAV* score in East Asians mapped to a nonsense mutation (rs1343879) in enigmatic *MAGEE2* (MAGE family member E2 expressed in the brain (200)) on chromosome X (Table 9 and Figure 29). This finding is particularly interesting as sex chromosomes have usually been omitted in selection screens. Nevertheless, a selection signal in this gene was picked up by Yngvadottir *et al*., who observed lower diversity in haplotypes carrying the stop allele and concluded that the truncated MAGEE2 conferred a selective advantage in East Asia (206); the *MAGEE2* transcript containing the stop variant was predicted to avoid nonsense-mediated decay and encodes a protein truncated by about 77%. It is difficult to predict whether or not the truncated protein is non-functional. Assuming that the truncated human product is a loss-of-function variant, the *Magee2* null mouse would itself model the biological consequence of the selected human stop allele. Null *Magee2* strain is currently at the micro-injection stage.

PRSS53

PRSS53 (protease, serine 53) encodes one of the polyserine proteases called polyserase-3 (POL3S), which has the biochemical property of hydrolyzing peptide bonds but whose functional role is largely unknown (534). Proteases make up the human degradome involved in a wide variety of biological processes including embryonic development, blood coagulation, tissue remodelling, wound healing, cell-cycle progression, angiogenesis, apoptosis, autophagy and senescence (534). Moreover, it is now well established that proteases participate in these key biological events through the selective and limited cleavage of specific substrates (534). Polyserase-3 was shown to be expressed in most tissues and tumor cell lines analysed suggesting that this enzyme may contribute to tumor development and progression (535). Another study showed that POL3S may play a substantial role in the function of pancreatic islet β-cells, and it has been classified as a potential diabetes-associated gene (536). Finally, genome-wide association analysis identified polyserase-3 as a psoriasis susceptibility locus that showed the strongest

differential expression between psoriatic and normal skin (2.66-fold increase in lesional skin compared to control skin) (537). The broad range of associated traits did not allow formulation precise hypothesis about the selection in *PRSS53* and, therefore, a mouse knock-out of this gene was generated.

Our interest in this gene arose from two variants that were independently picked up in two different populations: a missense rs11150606 being the 6th top scoring variant in East Asians (also supported by *CMS* and previous reports) and the nonsynonymous rs201075024 scoring highest in South Asians. Both alleles fall in close proximity, only 10 bp apart (Table 9 and Figure 30 upper panel), which might indicate a similar functional consequence and convergent evolution. However, in parallel to our study, Adhikari *et al*. recently showed that *PRSS53* is highly expressed in the hair follicle, and associated rs11150606 with hair shape in East Asians (196). The authors confirmed functionality of rs11150606 by *in vitro* assays showing that it affects processing and secretion of the gene product, potentially contributing to a straight hair phenotype (similarly to the well-established *EDAR* variant) (196). Our novel *Prss53* null mouse strain has already entered the phenotyping pipeline and the early investigation revealed abnormal vibrissae morphology. Curly vibrissae shape was observed for 30% and 80% of homozygous null females and males respectively (Figure 31 and Figure 32). Such finding further supports *PRSS53* involvement in hair shape and appropriateness of the mouse model to study this phenotype.

HERC1

HERC1 (HECT and RLD domain containing E3 ubiquitin protein ligase family member 1) encodes a giant multidomain protein that acts as a guanine nucleotide exchange factor, GTPase regulator and E3 ubiquitin ligase (538). This protein is thought to be involved in membrane transport processes, protein stabilization and degradation, cell proliferation and growth (539). *HERC1* is widely expressed in all human and mouse tissues examined (200, 540). Furthermore, *HERC1* was found to be mutated in multiple tumors and its overexpression has been shown in all human tumor cell lines tested (540). HERC1 is known to interact with and destabilise the

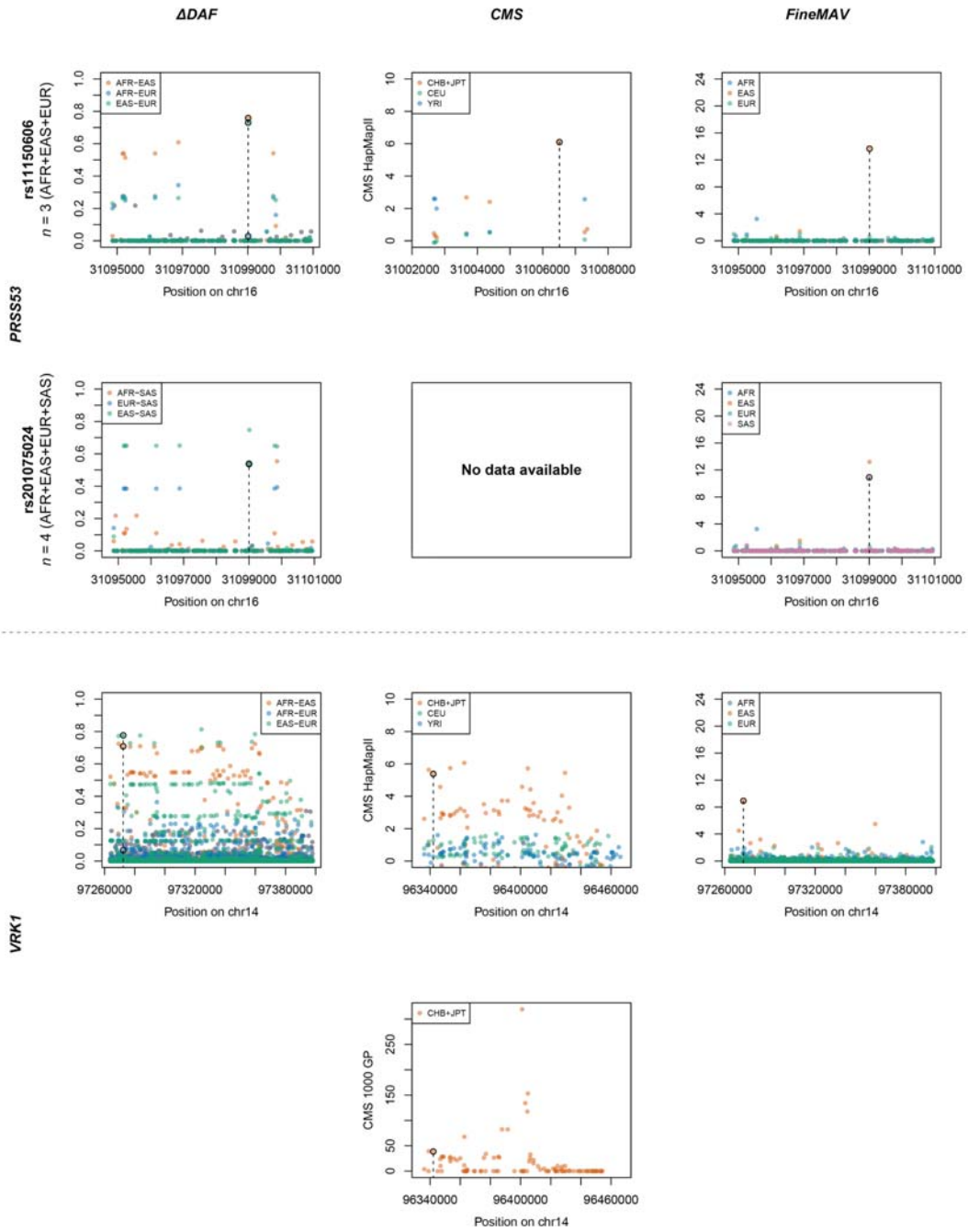

Figure 30. Comparison of three different approaches for pinpointing selected variants. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). *CMS* scores are given for the pilot phase of 1000 Genomes Project (155, 185) (if available) and the phase II of the International Haplotype Map Project (HapMapII) (123, 146), all downloaded from http://www.broadinstitute.org/. 1000 Genomes Project *CMS* scores included window named region145new, which spans *VRK1*. Variants with *CMS* values set to 'nan' or below 0 are not shown. *CMS* has not been calculated in South Asians, therefore CMS scores for rs201075024 in *PRSS53* are missing. Genomic positions are given in bp according to GRCh37 for *ΔDAF* and *FineMAV*, and build NCBI36 for *CMS*. The selected variant is marked with a dashed line. Note that the y-axis scale in the *CMS* plots is not standardised.

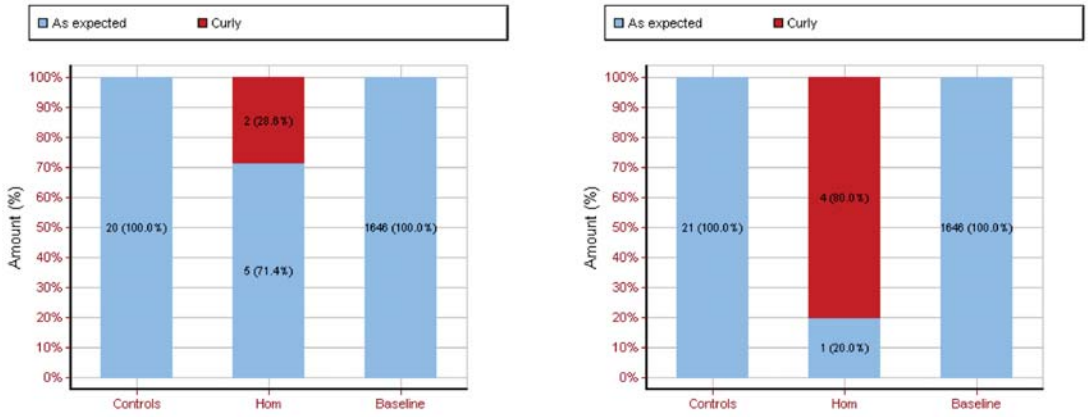

Figure 31. Vibrissae shape in *Prss53*$^{-/-}$ mice. Left panel: Females. Right panel: Males. The first number in the bar indicates the number of individuals investigated (*n*) followed by the percentage given in parentheses.

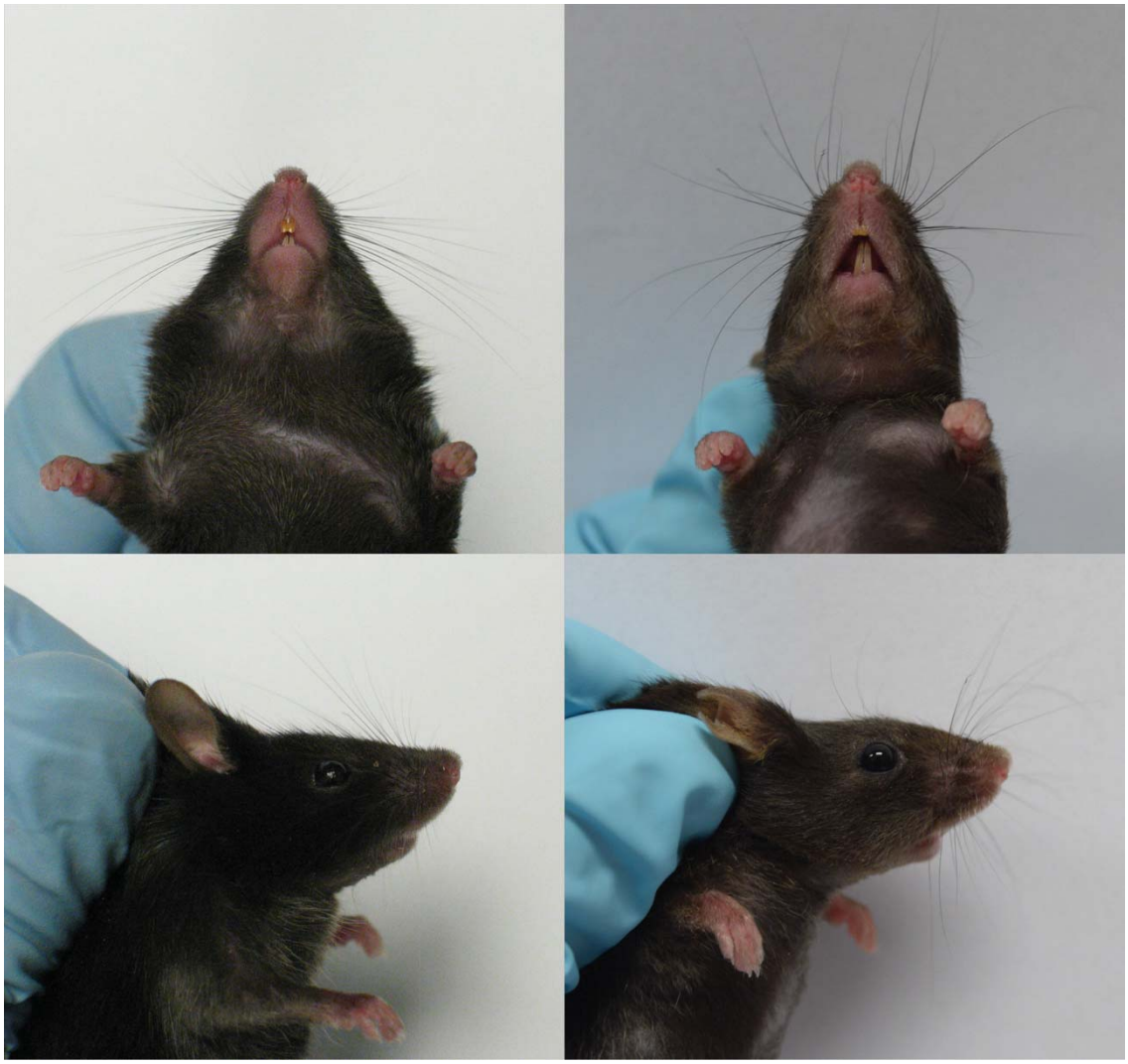

Figure 32. Vibrissae shape in *Prss53*$^{-/-}$ mice. Left panel: Control. Right panel: *Prss53* null mutant.

tumor suppressor TSC2 (541) and regulate MSH2 degradation (a DNA mismatch repair enzyme maintaining genomic integrity) (542). Its knockdown leads to a significant reduction in DNA mismatch repair capacity in human leukemia cells (542). Recent studies and selection scans have suggested that the human *HERC1* gene have been affected by local positive selection (strong *CMS* and *SSI* signal; Table 9 and Figure 33 upper panel). Marked differences in allele and haplotype frequencies between East Asian and non-East Asian populations have been reported, together with low genetic diversity in East Asia (543). The biological function of *HERC1*, however, has not been well defined.

Homozygous disruption of *Herc1* by spontaneous mutation in mouse was shown to produce a phenotype characterised by abnormal hind limb posture, decreased coordination (balance and tremor), reduced weight, decreased survival, and progressive Purkinje cell (PC) neurodegeneration leading to severe ataxia and reduced lifespan (539). Both sexes appeared to be fertile although poor breeders. All these phenotypic characteristics correlate with extensive autophagy observed in the PCs of mutant mice associated with an increase of the mutant protein level (539). Successful complete transgenic rescue was achieved with either a mouse BAC containing the normal copy of *Herc1* or with the human *HERC1* cDNA (539). It was concluded that *HERC1* has a profound impact on animal growth and the maintenance of the cerebellum structure (539), although this study did not assess the effect of this mutation on other aspects of mouse biology. Therefore, we decided to carry out a novel knock-out study with standardised primary phenotyping.

Phenotypic characterisation of the *Herc1* null mice has not been completed yet, but has already revealed a range of early observations. Homozygotes for the null allele appear to be infertile. There is also a strong trend for high-frequency hearing loss (Auditory Brain Response thresholds are elevated at 24-30kHz frequencies). Furthermore, we observed increased body weight, mostly affecting males (Figure 34). This has also been confirmand by body composition X-ray imaging showing increased total body fat amount and increased percent body fat in both sexes (Figure 35 and Figure 36). Furthermore, comprehensive plasma chemistry analysis reported abnormal levels of many parameters: sodium, chloride, high density lipoprotein (HDL), amylase, albumin (data not shown) and insulin (Figure 37). Whole blood terminal haematology analysis picked up deviation in

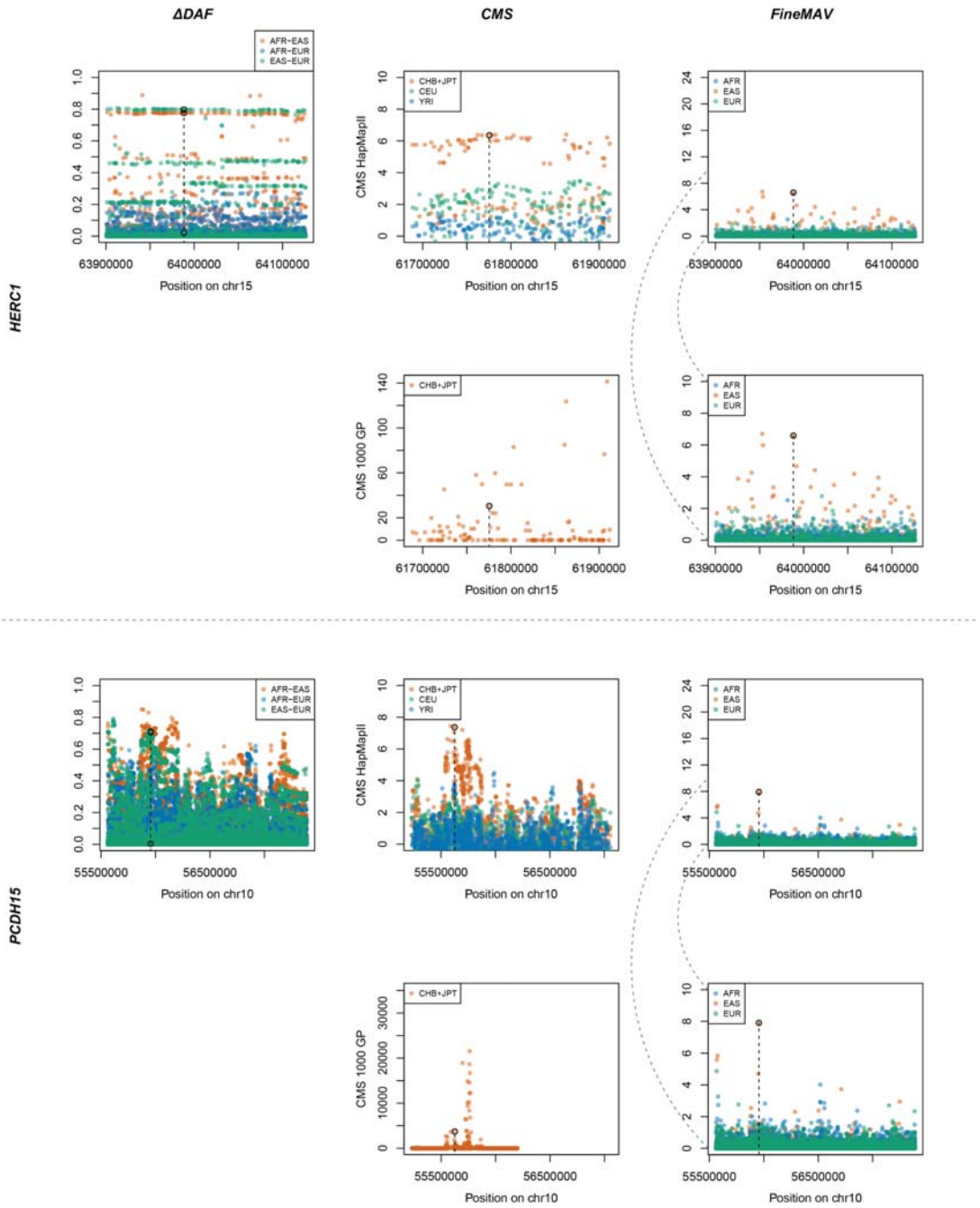

Figure 33. Comparison of three different approaches for pinpointing selected variants. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). Expanded view of the FineMAV plot is given underneath. *CMS* scores are given for booth, the pilot phase of 1000 Genomes Project (155, 185) and the phase II of the International Haplotype Map Project (HapMapII) (123, 146), all downloaded from http://www.broadinstitute.org/. 1000 Genomes Project CMS scores included windows named: region147new and region134new spanning *HERC1* and *PCDH15* respectively. Variants with *CMS* values set to 'nan' or below 0 are not shown. Genomic positions are given in bp according to GRCh37 for *ΔDAF* and *FineMAV*, and build NCBI36 for *CMS*. The selected variant is marked with a dashed line. Note that the y-axis scale in the *CMS* plots is not standardised.

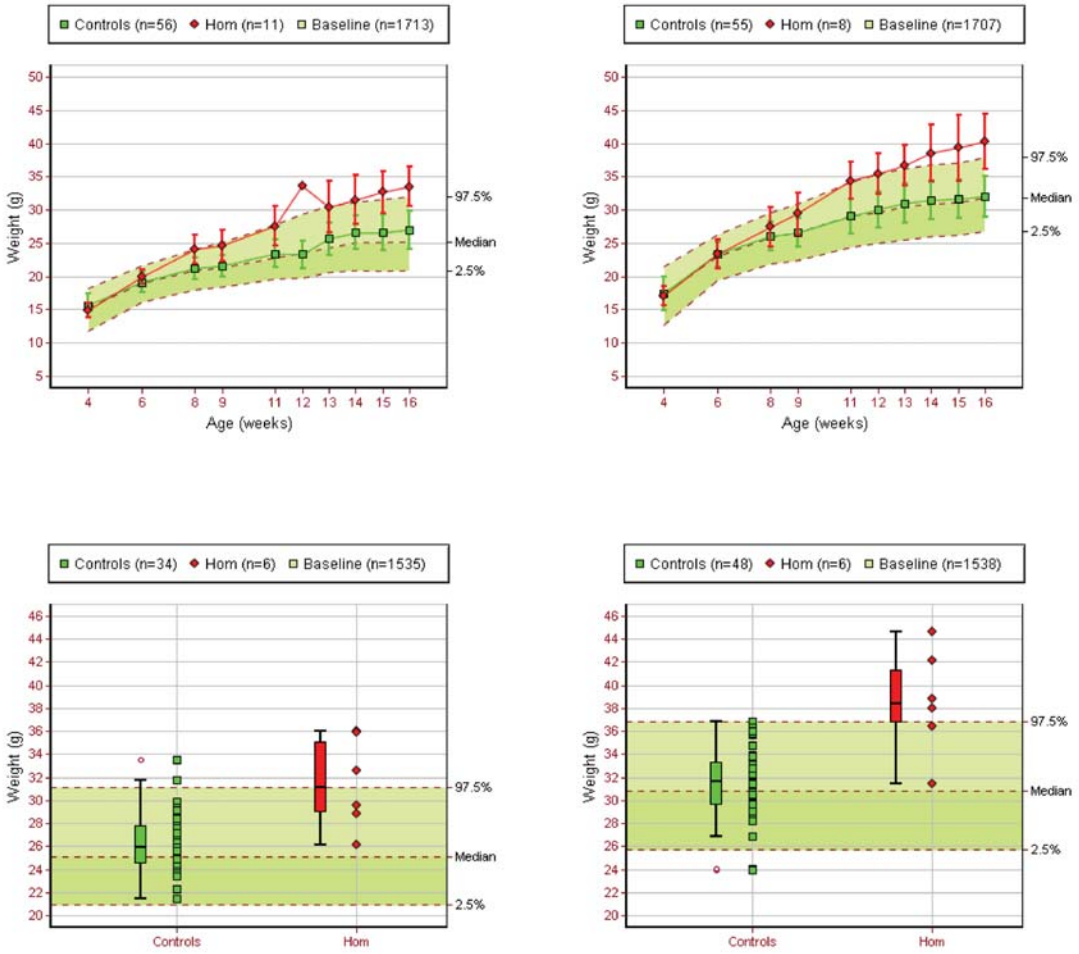

Figure 34. Increased body weight in *Herc1*$^{-/-}$ mice. Top panel: Average weight curve; Bottom panel: Body weight; Left panel: Females; Right panel: Males.

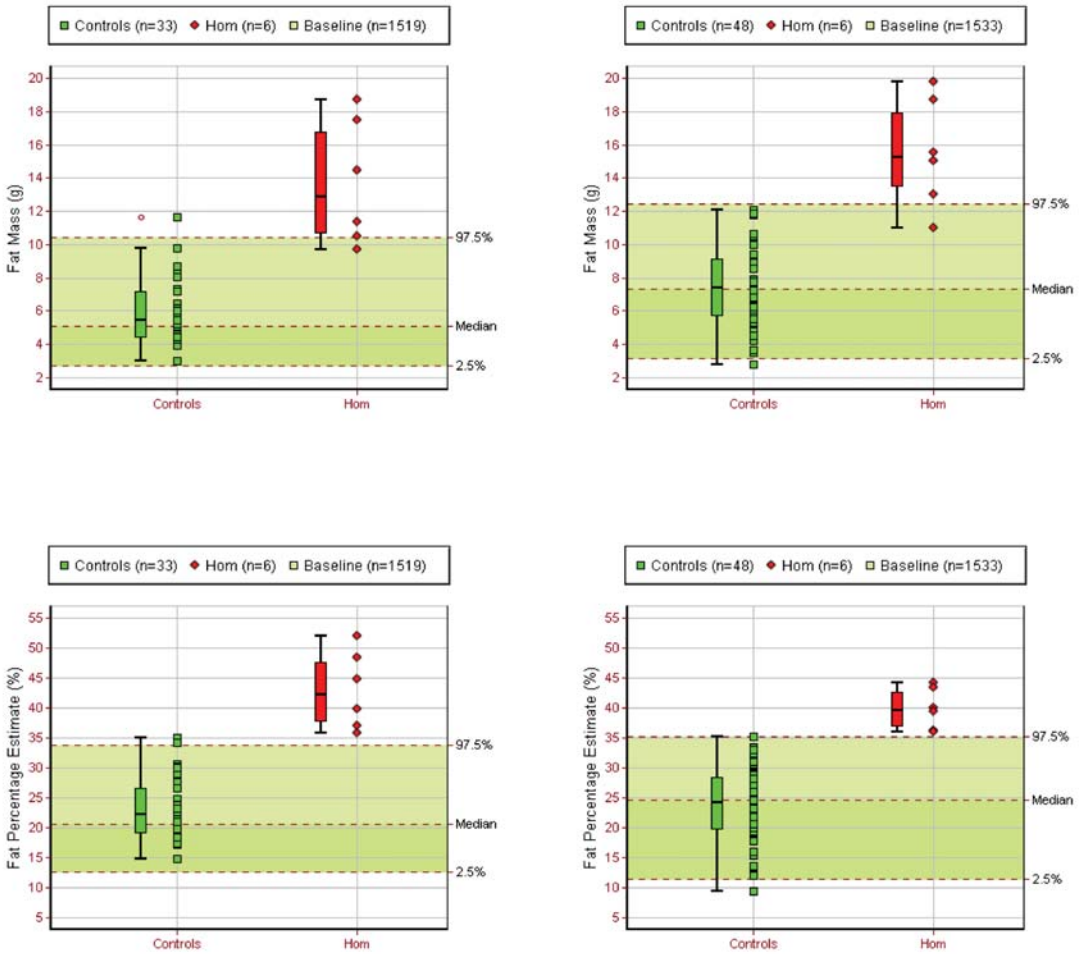

Figure 35. Increased total body fat amount in *Herc1*$^{-/-}$ mice. Top panel: Fat mass; Bottom panel: Fat percentage estimate; Left panel: Females; Right panel: Males. Body Composition was examined in anaesthetised mice using a dual energy X-ray absorptiometry machine (Lunar PIXImus II).

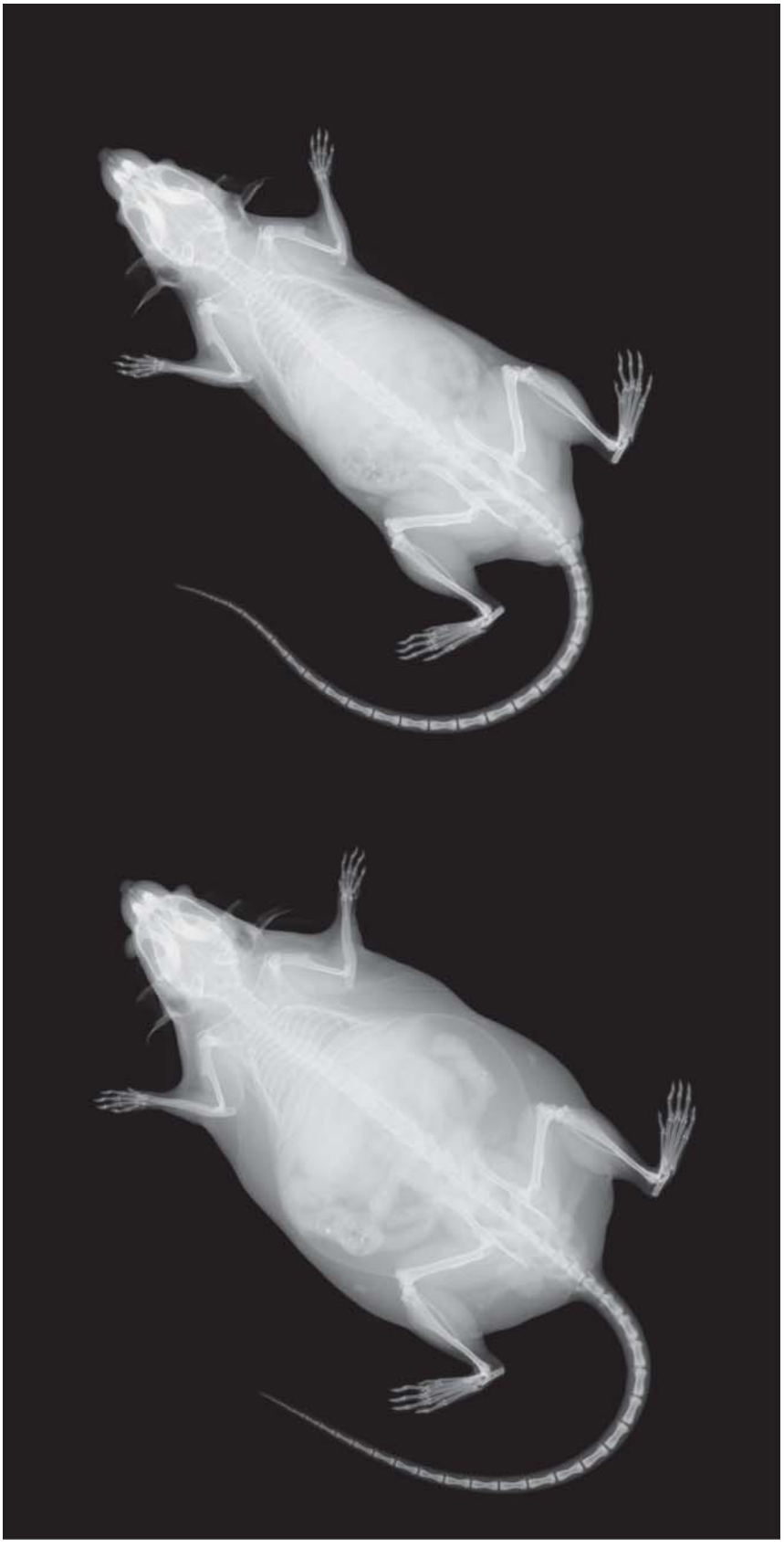

Figure 36. Body Composition X-ray imagining of a *Herc1*$^{-/-}$ mouse. Top panel: *Herc1*$^{-/-}$ male; Bottom panel: Control male. Anaesthetised mice were imaged on a dual energy X-ray absorptiometry machine (Lunar PIXImus II).

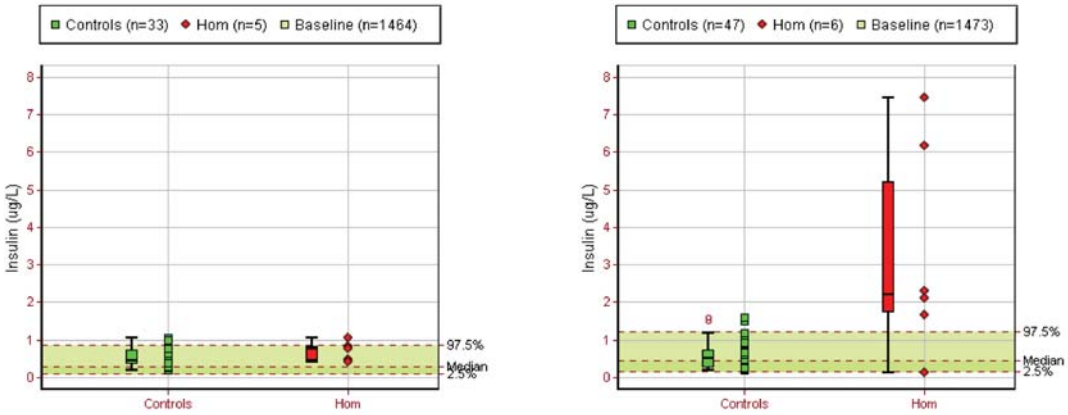

Figure 37. Increased circulating insulin level in *Herc1*$^{-/-}$ mice. Left panel: Females; Right panel: Males. Plasma insulin concentration was measured by Mesoscale Discovery (MSD) array technology in non-fasted terminally anaesthetised with ketamine/xylazine mice.

white blood cell count (increased leukocyte cell number in females) and red blood cell distribution width (increased red blood cell distribution width in females). Neurological and dysmorphology assessment did not reveal any abnormalities so far (normal paw grip, limb grasping and gait; no ataxia or tremor), although brain histopathology has not yet been performed.

Functional characterization of *Herc1*$^{-/-}$ mice model generated in this study disclosed a range of phenotypes. Interestingly, they did not recapitulate the neurodegeneration and motor impairment phenotypes of the published mouse mutant carrying a spontaneous missense mutation disrupting *Herc1* (539), although the primary phenotyping of our mutant has not been completed yet. Contrary to what we found, the spontaneous mutation affected animal growth and survival resulting in reduced weight (539). Recently reported loss of function mutations in humans manifest intellectual disability, megalocephaly, facial dysmorphism, motor development delay, hypotonia, limb and gait abnormalities, and occasionally seizures, overgrowth due to excessive tissue proliferation and brain abnormalities, with small cerebellum among others (544-546). Furthermore, *HERC1* was associated with autism spectrum disorder in humans (547). None of the published studies reported the metabolic and hearing abnormalities found in our model. It is possible that the spontaneous missense mutations in mice and humans are gain-of-function rather than loss-of-function, e.g. it has been shown that the known mutation in mouse enhances the stability of the *Herc1* protein and increases its level, which is very different to a lack of protein (539, 544). Phenotypes in human also vary between cases (544-546). It is likely that the nature of the mutations as well as possibly the effect of modifier genes leads to these phenotypic differences (544, 545). It is also possible that *HERC1* disruption affects several pathways and many systems, as its protein product is involved in membrane transport processes, potentially causing neurodevelopmental malformations and consequent abnormalities (546). The broad range of affected phenotypes complicates conclusive hypothesising about the likely selective benefits that drove adaptation, nevertheless it seems that *HERC1* is fundamental in early brain development and function (546, 548).

3.2.3.2. Knock-ins

In parallel to our knock-out studies, we requested the generation of 9 mouse models carrying the putatively selected human derived allele that met our prioritization criteria in order to test hypothesis about their causality and roles in human adaptation. We generated knock-ins of three candidate adaptive variants falling into genes targeted in our knock-out project, as well as six additional variants in cases where there was enough prior information (Table 10). We developed an interest and collaborations in four specific phenotypic categories: hair shape, reproduction energy metabolism and hearing. As the effects are expected to be rather subtle, a detailed secondary phenotyping of these humanized mice is planned on top of the standardised primary phenotyping. Two strains are at the early stages of primary phenotyping, and no gross abnormalities have been detected (Table 10). A description of the selected candidates and the signature of their selection, as well as discussion of the hypothesised phenotype potentially driving the selective force are given for each gene separately in the following sections.

Table 10. List of humanized mouse strains generated in this study. All human-mouse orthologue pairs shortlisted here are 1-to-1 orthologues. 'Top SNP' lists SNPs with the highest *FineMAV* score in the given gene, which is most likely driving the signal of selection in humans and is modelled in this study; * in the case of *HERC1* the second-highest scoring SNP was chosen. 'Consequence' and '*FineMAV*' specify properties of Top SNPs; 'Pop.' – population with the signal of selection ('NES' – Northeastern Siberian); '*SSI*' – Selection Support Index for each gene; 'Orthologue identity' – percentage of the mouse protein sequence matching the human protein sequence / percentage of the human protein sequence matching the mouse protein sequence; 'Stage' – current stage of each line: MI – micro-injection, CE – colony expansion of genotype-confirmed mice, PP – primary phenotyping.

Gene	Top SNP	Consequence	*FineMAV*	Pop.	*SSI*	Orthologue identity	Stage
CPT1A	rs80356779	missense	17.48	NES	0.05	87/87	CE
HERC1	rs2255243*	missense	6.59	EAS	0.49	97/97	PP
LRGUK	rs34890031	missense	21.32	AMR	0.04	73/73	MI
OTOF	rs17005371	missense	10.57	AFR	0.12	95/95	MI
PCHD15	rs4935502	missense, splice region	7.91	EAS	0.32	84/84	CE
PRSS53	rs11150606	missense	13.66	EAS	0.09	81/81	PP
PRSS53	rs201075024	missense	10.91	SAS	0.09	81/81	MI
TGM3	rs6048066	missense	9.77	AFR	0.12	77/77	CE
VRK1	rs2224442	regulatory	8.93	EAS	0.03	78/87	MI

3.2.3.2.1. Hair shape: *PRSS53* and *TGM3*

It has been shown that genomic regions associated with scalp hair features are enriched for signals of recent selection in humans (196). We modelled two derived alleles of the *PRSS53* gene (described in the Knock-out section) selected in East (rs11150606) and South (rs201075024) Asians, likely due to hair-related phenotypes (Table 10 and Figure 30 upper panel). These two knock-ins are being engineered alongside the *PRSS53* knock-out. The model of East Asian-specific allele has reached the initial stage of primary phenotyping, while the South Asian-specific allele model is at the micro-injection phase. It is predicted that these variants contribute to hair shape phenotype based on the genome-wide association study and functional follow-up in humans (196), but also generated here null mouse strain. It has been proposed that the straight hair phenotype has been selected outside Africa as it tends to naturally fall over the ears and neck, which could provide an adaptive advantage in cold climates relative to tightly curly hair (549). Furthermore, some have argued that straight hair enables the passage of more UV light into hair roots (and consequently into the skin) via the hair shaft, which facilitated vitamin D production at high latitudes (549, 550). Assessment of the impact of selected alleles on hair straightness will probably require detailed secondary phenotyping, as mouse hair is naturally straight and we do not expect obvious abnormalities to be picked up by the general primary screen.

Another variant that drew our attention is a missense mutation (rs6048066) in *TGM3* expressed in the cuticle of growing hair fiber and putatively selected in Africans (described in 2.3.2.5.2. Missense variants) (Table 10 and Figure 29). The mouse colony carrying the human derived allele is currently at the expansion stage preceding primary phenotyping. A selection signal in *TGM3* has also been implied in previous studies and the likely driver mutation detected by *FineMAV* scores as the 53rd top signal in Africans. We proposed that this amino acid change might cause enzyme deficiency and contribute to African hair texture. However, considering its frequency in Africa (43%) and the fixed prevalence of Afro-textured hair, this variant alone cannot explain hair curliness, but could potentially contribute to this complex and quantitative trait (the commonly-used Andre Walker hair typing system classifies hair texture into 4 types, each with 3 subcategories, resulting in

12 simplified classes used to describe different variations among individuals, although more scientific approaches have also been proposed (551)). This hypothesis is supported by the observation that *Tgm3* is a modifier of the *wal* (unlocalised) gene in mice (552). Homozygous mice carrying mutant *wal* also have a wavy coat (552). The hair curliness in double mouse mutants (*Tgm3*$^{-/-}$ *wal/wal*) is much more striking than in *wal/wal* or *Tgm3*$^{-/-}$ mutants alone, suggesting an additive effect (552).

Moreover, the hair of the *Tgm3* null mouse was also reported to be shorter than in normal mice, whilst the whiskers were twisted and thinner (553), which is consistent with observed lower hair growth rate and diameter in Africans (554, 555). It has been hypothesised that Afro-hair morphology experienced strong positive selection as the trait has been retained/preferred among many equatorial human groups (215). While sexual selection cannot be ruled out as being responsible for such pattern, a strong correlation with geography suggests rather an environmental influence. Moreover, although sub-Saharan Africans are the most genetically diverse population, curly textured hair seems to be a fixed derived feature in this region when compared to non-human primates. This points towards a strong, long-term selective pressure in the savannah environment (556). It has been suggested that Afro-textured hair may have been adaptive in Africa because the relatively sparse density of such hair, combined with its elastic helix shape, results in an airy effect that likely facilitates body-temperature regulation via improved circulation of cool air onto the scalp (215, 549). Additionally, wet tightly coiled hair does not stick to the neck and scalp which could further enhance the cooling system (549). Finally, curly hair was also argued to protect from UV light passage into the body better than straight hair (215, 549, 550).

3.2.3.2.2. Reproduction: *LRGUK* and *VRK1*

Two of the proposed knock-in alleles are predicted to have been selected due to effects on fertility, and are both at the micro-injection stage of engineering. Unquestionably, natural selection that improves reproductive fitness could act directly by modulating fertility levels. The strongest selection signal observed in

Native Americans fell on rs34890031 (missense Arg->His) in *LRGUK* (leucine rich repeats and guanylate kinase domain containing) and could be one such example (Table 10 and Figure 38). The mouse homologue is essential for multiple aspects of sperm development, assembly and function including acrosome attachment, head shaping and tail formation (248). A null mouse model caused by a nonsense mutation and nonsense-mediated mRNA decay resulted in male-specific infertility, chaotic and disorganised spermatogenesis, 81% reduction in sperm production and 13% reduction in testis weight (248). Abnormal sperm development was manifested by head and tail abnormalities and germ cell degeneration that resulted in no capacity for motility (248). *LRGUK* is predominantly expressed in human and mouse testis (200, 248).

Another selected candidate (rs2224442) falls in a promoter flanking region in the intron of *VRK1*. The region surrounding rs2224442, although non-coding, is characterised by high conservation across taxa and the presence of DNaseI hypersensitivity, and scored as the 46th top *FineMAV* variant in East Asians (Table 10 and Figure 30 lower panel). VRK1 is a protein kinase implicated in mitotic and meiotic cell cycles, cell proliferation and differentiation (233, 234, 557, 558) that plays an important role in organogenesis of sex organs and gametogenesis in multiple species (235-238). VRK1-deficient organisms show abnormality of the reproductive organs, followed by defects in germ cell development (235-238). Both sexes of *VRK1*-null mice have been reported to be infertile displaying defects in sex organs (e.g. small testis in males) and impaired oogenesis and spermatogenesis due to meiotic arrest manifested as azoospermia and lack of mature sperm in males (239-242). It might be that this regulatory variant affects the expression level of *VRK1* and modulates the maturation of gametes.

Although *VRK1* expression is highly enriched in testis compared to other human tissues (200), mutations in this gene have been linked to early-onset spinal muscular atrophy, neurogenic atrophy, ataxia, microcephaly developmental delay and intellectual disability due to disturbance of cell cycle progression (559-564). It has been also reported to act as a tumor suppressor gene that contributes to genomic stability by facilitating DNA damage responses (565-568). It is clear that a gene implicated in the coordination of diverse signaling processes and functions (especially fundamental function like cell division) might have pleiotropic effects

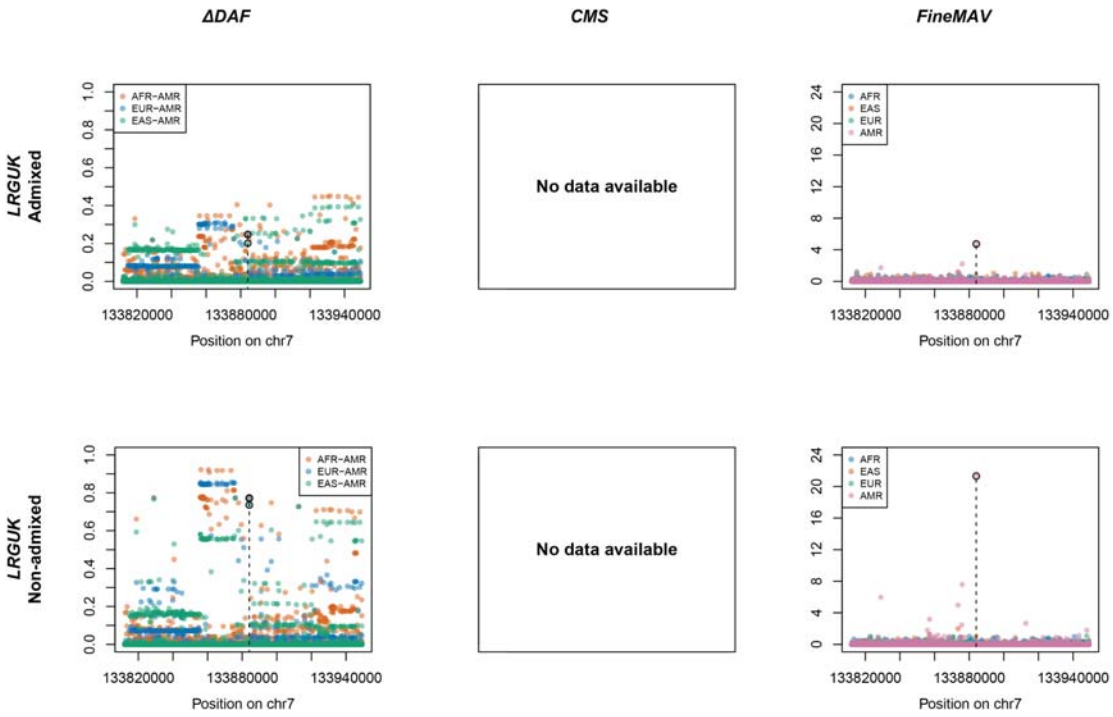

Figure 38. Signal of selection in *LRGUK* according to different approaches. *ΔDAF* and *FineMAV* scores are shown for the genomic windows spanning *LRGUK* gene. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142) including Admixed Americans (upper panel) and 24 Quechua from Peru described in 2.4. *FineMAV* application to various populations (lower panel). *CMS* has not been applied to Native American genetic data, therefore *CMS* scores are missing. Genomic positions are given in bp according to GRCh37 for *ΔDAF* and *FineMAV*. The selected variant is marked with a dashed line.

(557), therefore a comprehensive primary phenotyping assessing the function of many organs needs to be performed to identify possible reasons for adaptation at this locus.

3.2.3.2.3. Energy metabolism: *HERC1* and *CPT1A*

Energy metabolism is another field that we developed an interest in, and one that has been reported to be targeted by recent human evolution. According to the thrifty genotype hypothesis, adaptation might have favoured efficient energy expenditure to maximize energy storage, which enhanced survival in periods of food shortage but predispose to obesity and Type 2 Diabetes in a modern dietary environment (44, 45, 366, 367).

Although the *HERC1* knock-out model manifested a metabolic and auditory phenotype, these might still be secondary consequences of a neurodevelopmental disorder (described in the Knock-out section). Nevertheless, we decided to model the candidate causal variant in parallel to the knock-out line. It is a particularly complicated example as the *FineMAV* analysis revealed 3 high-scoring candidates in *HERC1* in East Asians (Figure 33): 1 intronic, 1 missense and 1 synonymous (in decreasing signal order). We proceeded with modelling the missense candidate rs2255243 (2nd scoring in *HERC1*) as its functional consequence is easier to predict and examine. However, we do not exclude the possibility of investigating the other two variants in the future. Another complication was that the mouse amino acid (Thr) does not match either the human ancestral (Ala) or derived (Gly) amino acids at this position. Therefore, we decided to model both the human ancestral and derived form of HERC1 in mouse. Both lines have entered the primary phenotyping phase.

Another variant that is a part of our modelling study comes from a published selection scan in Arctic populations (55, 56). One of the strongest signal of selection in the Siberian population was mapped to a genomic region spanning *CPT1A* and has been linked to rs80356779 variant causing Pro479Leu amino acid change, which appears to be a functional candidate for the cold adaptation in this population (55, 56). We also applied *FineMAV* to this dataset and found rs80356779 to be the

highest scoring variant in Northeast Siberians (NES) (run configuration: AFR, EAS, EUR, NES; n = 4; x = 2.98; data not shown). *CPT1A* was associated with serum metabolite levels and obesity in GWAS studies (569-572) and encodes carnitine palmitoyltransferase 1A, a liver enzyme located on the outer mitochondrial membrane, required for the import of long-chain fats into the mitochondria for use in beta-oxidation and energy production (573-575). *CPT1A* is active during fasting to maintain energy, sparing glucose for vital bodily functions by generating ketones (which serve as an alternate energy source) (366, 573-575). In the fed state when there is sufficient glucose availability, *CPT1A* is inhibited by malonyl-CoA (573, 576, 577). It has been shown that homozygosity for the Pro479Leu variant (at rs80356779) in *CPT1A* decreases enzyme thermostability and functional activity (20% of normal) and makes the enzyme relatively insensitive to malonyl-CoA inhibition (576, 578). As a result, Pro479Leu homozygosity causes *CPT1A* deficiency that impairs fatty acid oxidation, ketogenesis and fasting tolerance, also conferring risk for hypoketotic hypoglycaemia, seizures, and sudden unexpected death in infancy (SUDI) during fasting related to illness (367, 574, 575, 579, 580). It has been extensively characterised in the Inuit population, in whom it was associated with increased infant mortality (574, 575).

Paradoxically, a high prevalence of Pro479Leu variant in the *CPT1A* gene has been identified among aboriginal Arctic populations (up to 85%) suggesting selective advantage in the past (367, 573-575, 578, 581, 582). The constitutively active, malonyl-CoA resistant, Pro479Leu *CPT1A* protein maintaining increased basal rate of beta-oxidation and ketogenesis at all times may have been advantageous because of its cardioprotective role in the context of the traditional high fat diet (with little to no carbohydrates) of indigenous Arctic people, although this information alone does not fully explain a selective advantage for the variant (575, 578, 581, 583). Furthermore, such a traditional diet enriched in n-3 polyunsaturated fatty acids increases expression of *CPT1A*, which may compensate for reduced activity, and the observed deleterious effect might be caused by a recent lifestyle shift (367, 583, 584). Indeed, this variant has been linked to smaller body size and reduced body fat deposition, low serum cholesterol and triglyceride levels, reduced insulin resistance and high circulating HDL-cholesterol (575, 578, 581, 583, 585). It has been observed that indigenous Siberians, even when obese, do not

develop features of metabolic syndrome, insulin resistance and type 2 diabetes (the so-called 'healthy obese' phenotype) due to increased basal fatty acid oxidation rate (367, 581, 583). It has also been proposed this variant might protect against infection via an elevated apolipoprotein A-I level (581).

Although the causality of this mutation seems to be well established, the selected advantage is unclear and has not been explicitly tested. We want to test if the selected variant confers cold climate adaptations in order to optimise energy utilization (575). Cold adaptive processes could be expected to involve fatty acid metabolism in energy and systemic heat production, as continuous cold exposure is known to determine the mobilization and metabolism of fat (55). It has been shown that cold exposure increased fatty acid β-oxidation capacity in mice adipose tissue via increased *Cpt1a* expression (586). Mouse seems to be a suitable organism for our hypothesis testing as a mouse knock-out line phenotype recapitulates the human loss-of-function mutation. Homozygous null mice displayed embryonic lethality, while heterozygotes (~55% Cpt1a activity in the liver) were cold tolerant but exhibited decreased serum glucose and increased serum free fatty acid levels after fasting (587). Our mouse model is currently at the stage of colony expansion for primary and secondary phenotyping.

3.2.3.2.4. Hearing: *OTOF* and *PCDH15*

The high-ranking variant (84th *FineMAV* hit in East Asians; Table 10 and Figure 33 lower panel) shortlisted in our study is a nonsynonymous rs4935502 acidic-to-nonpolar (D435A) mutation in *PCDH15* (with a high *CMS* signal and strong support from the literature). This mutation alters a highly-conserved residue predicted to lie in the Ca^{2+}-binding site at the protein's cadherin-4 domain (123) and might have been selected due to an advantageous effect on some aspect of hearing. PCHD15 (protocadherin-related 15) is a member of the cadherin superfamily of integral membrane proteins that mediate calcium-dependent cell-cell adhesion (588). The *PCDH15* gene encodes three alternative isoforms differing in their cytoplasmic domains (CD1, CD2, and CD3) characterised by different expression patterns (mainly cochlea, retina, brain, lung and testis (200, 588, 589)),

suggesting that alternative splicing regulates *PCDH15* function (588). The protein product of this gene is necessary for normal retinal and cochlear functions (590). Hearing and balance use hair cells in the inner ear to transform mechanical stimuli into electrical signals (590). Mechanical force from sound waves or head movements is conveyed to hair-cell transduction channels by tip links, fine filaments formed by PCDH15 and CDH23 (591, 592). Mechanical force increases tension in tip links, which in turn conveys force to mechanosensitive ion channels to open them (592, 593). *PCDH15* was shown to play crucial role in the morphogenesis and organization of hair cell bundles and in the maintenance of retinal photoreceptor cells (590, 594). Mutations affecting these neuroepithelia in mice and rats cause profound deafness and a balance disorder due to degeneration and abnormalities of hair cells, although visual defects are not evident (589, 594, 595). Homozygotes for severe mutations exhibit hyperactivity, head-tossing, circling behaviour and impaired swimming indicative of vestibular dysfunction, along with the lack of an auditory-evoked brainstem response at the highest intensities of acoustic stimulation (589, 594, 596). Surprisingly, mice lacking PCDH15-CD1 and PCDH15-CD3 maintain hearing function (form normal hair bundles and tip links), while PCDH15-CD2-deficient mice are deaf (597). However, vestibular function remains intact in the PCDH15-CD2 mutants (597). In humans, mutations in *PCDH15* result in hearing loss, whereas more severe mutations cause Usher Syndrome Type IF (USH1F) characterised by profound deafness and vestibular dysfunction with progressive loss of vision due to retinitis pigmentosa (590). Defects in the cochlea include degeneration of hair cells and disrupted interactions between CDH23 and PCDH15 (tip-link function) (594).

On the other hand, some isoforms were detected in natural killer (NK)/T cells (598). Published studies associated *PCDH15* with extrapulmonary tuberculosis (599), late-onset Alzheimer disease (600) and response to smallpox vaccine in Hispanics (601), showing that this gene may be important in regulating humoral immunity. *PCDH15* expression was also detected in liver and pancreas, and loss-of-function mutations in the mouse orthologue caused abnormalities in the lipid profile. Similarly, *PCDH15* has been associated with anthropometric traits related to body size and adiposity (602), lipid abnormalities and increased risk of premature coronary heart disease in humans (603). All of the above might indicate

potential pleiotropic effects of *PCDH15* mutations. Our humanized mouse model is currently at the colony expansion phase for primary and secondary phenotyping.

The final variant selected for modelling is rs17005371, causing an amino acid substitution in the protein product of *OTOF* and putatively selected in Africans (30th top score; Table 10 and Figure 29). A mouse model carrying the African derived mutation is at the micro-injection phase. *OTOF* encodes otoferlin, which is expressed mainly in cochlear auditory inner hair cells (but also brain) (200, 604-606) and plays an essential role in a late step of synaptic vesicle exocytosis and neurotransmitter release at the synapse between inner hair cells and auditory nerve fibres (604, 606-608). OTOF acts as a Ca^{2+} sensor, triggering vesicle fusion at synaptic membranes (calcium dependent membrane-membrane fusion) (604, 605, 607-612). Disruptions in this gene result in synaptic disorder, impairment of auditory nerve firing and severe to profound hearing loss (605, 613). Some mutations cause temperature sensitive auditory neuropathy manifested by severe hearing loss during fever, which recovers when the body temperature returns to normal (614, 615). It might indicate that some forms of OTOF have a reduced activity as the temperature increases (616). There are several alternative splice isoform of otoferlin (604), but normal hearing is thought to require the long isoform and exon 48 (616). Insight into the molecular function of this gene was provided by mouse knock-out studies (608, 611, 612, 617). Disruption of the mouse ortholog recapitulated the hearing loss seen in human patients (608, 611, 612, 617). The null mouse had structurally normal synapses between hair cells and the auditory nerve fibre, but lacked calcium-triggered dumping of the synaptic vesicle contents (abolished exocytosis) (608, 611, 612, 617). Interestingly, both *PCDH15* and *OTOF* orthologs have undergone adaptive evolution in echolocating mammals (bats and toothed whales) implying that they might have co-evolved to optimise cochlear amplification (618, 619).

4. General discussion

4.1. Summary

Here, we return to wider questions in the field of adaptation in humans, and *FineMAV*'s contribution to it. The genetic basis of human adaptations is of great interest and has a correspondingly large literature. Most previous work has focused on investigating the mode of adaptation (classic selective sweeps vs selection on standing variation) and scanning the genome for signatures of positive selection. The current literature thus documents that classic sweeps were not common (18, 20), and are difficult to identify reliably from population-genetic data alone as attested by the limited overlap between genomic selection scans, but nevertheless have occurred and are of great interest. We have not carried out another genome-wide scan for positive selection and are not entering the debate about whether or not classic selective sweeps were common in humans. Instead, we take the view that the field now needs additional well-supported examples of variants that are driving adaptations, both to understand specific events and to inform more general questions regarding the genetic basis of human adaptation. Support comes most compellingly from model cell/organism studies, but these are low-throughput and so a way to prioritise candidates for these is needed. We provide this by combining population-genetic and functional evidence into a single quantitative measure, the *FineMAV* score, which scans millions of variants genome-wide to generate a list of individual candidate variants in order of priority. We validated our method using a meta-analysis, a handful of gold standard variants, together with available *in silico* evidence for selection. We have begun modelling a few of the candidate variants in cells or mice ourselves, and have reported progress in this area. We hope that others may benefit from this work, either directly from the human candidates we identify, or more indirectly by applying our approach to other species, as our method is applicable to any species with suitable genomic data.

We thus provide a way to move forward from the morass of genome scans for positive selection. Our study probably misses many genuine selected variants

(high false negative rate), but our prioritization aims to enrich for true positives, which is what matters for people who are going to spend years examining each individual candidate in cellular or animal models, as it has not always been possible to find a link between a seemingly strong candidate variant and reproductive fitness. For instance, the reason for selection of the *TRPV6* haplotype containing three derived non-synonymous substitutions observed in non-African populations (620) remains enigmatic, despite detailed functional characterization of selected and non-selected forms at the cellular level (621). *TRPV6* encodes a Ca^{2+} selective ion channel, which is critically involved in dietary calcium uptake and (re)absorption (621). Potential functional differences between the ancestral and derived TRPV6 proteins were investigated in cell lines by carrying out electrophysiology experiments (621). No statistically significant differences in biophysical channel function were found (621). It remains possible that the ancestral and derived forms differ in other aspects that can only be observed at the whole-organism level (621). However, none of the three candidate sites for functional differences proposed previously was supported by our *FineMAV* analysis (in both selection scenarios $n = 3$ (AFR, EAS, EUR) and $n = 2$ (AFR, EAS+EUR)) and their predicted functionality is low (*FineMAV* $\sim$1 for each variant in EAS+EUR scenario). Therefore, we see them as weak candidates for causality and would not suggest modeling them.

Modeling of human selection in cell or animal systems is challenging since relevant phenotypic consequences (often very subtle) might be overlooked. Some phenotypes might be seen only in certain conditions, such as the presence of specific pathogens or environmental stresses. Sometimes an inappropriately chosen modeling system (cell lines, tissue or organ) might miss adaptive alleles with effects that only manifest in a particular organ or at a whole organism level (22, 138). The inability to directly demonstrate phenotypic consequences in a limited set-up does not entirely rule out the possibility that a variant has been selected (17). Nonetheless, regardless of challenges like these, cell and animal models often provide the best way to test hypotheses regarding recent human evolution (138). *FineMAV* now offers a better way to identify specific variants for modelling and paves the way for identification of causative alleles driving phenotypic differences among human populations.

4.2. Next steps

As discussed earlier functional validation of candidate signals of selection is a current roadblock in the field of population genetics, limiting both our understanding of the modes and importance of positive selection, and the independent evaluation of methods to detect it. Modeling of non-pathological human genetic variation in cell or animal systems, however, has received only limited attention to date (518). The impact of each human derived mutation needs to be compared with the ancestral allele control. While cellular phenotyping in an *in vitro* set up is often restricted to a particular cell type, assays performed in model organisms provide a much broader spectrum of possibilities. Since many genes are expressed in multiple organs and could potentially affect different tissues (have pleiotropic effects), it is crucial to perform a comprehensive and multidisciplinary primary phenotypic screen measuring a variety of physiological systems (even if there is functional insight to speculate about likely phenotypic outcomes). Standardised primary phenotyping of a wild-type and mutant mouse to assay phenotypic differences between the ancestral and derived alleles might overlook subtle phenotypic differences, so detailed secondary phenotyping addressing specific organ/tissue/function will often be needed. Successful examples show that in-depth follow-up studies of putatively-selected variation using *in vitro* experiments and model organisms constitute a suitable and promising tool to test hypotheses regarding human evolution.

All mouse strains generated as a part of this study will undergo standardised primary phenotyping (in case the modelled polymorphisms were selected for different functions than predicted, or to pick up pleiotropic effects), whilst knock-in lines carrying selected human derived single point mutation are also being subjected to detailed secondary phenotyping addressing the predicted phenotype. We have already established external collaborations with experts in relevant fields to focus on energy metabolism, hearing, hair and skin phenotypes of our models.

Secondary follow-up of hair phenotypes of *TGM3* and *PRSS53* mutant lines will be carried out at the Wellcome Trust Sanger Institute Mouse Phenotyping facility led by Chris Lelliott in collaboration with Paul Schofield (Department of

Physiology, Development and Neuroscience; University of Cambridge) and John Sundberg (The Jackson Laboratory, Bar Harbor, ME, USA). The phenotyping strategy prepared by Chris Lelliott includes longitudinal dysmorphology imaging capturing hair progression over time, a hair follicle cycling test, skin integrity measured by trans-epidermal water loss and comprehensive *ex vivo* skin histopathology (haematoxylin and eosin staining, immunohistochemistry imaging and electron microscopy imaging). Hair analysis will focus on both coat hair (dorsal and ventral) and vibrissae to assess parameters like curliness, proportion of different hair types, cross-sectional ellipticity and diameter, hair density (hair follicle bulbs per skin area), hair placodes size, physical resistance or hair rigidity, presence of isopeptide bonds and defects in cross-linking. These analyses will be complemented by proteomic profiling of the shaft using mass spectrometry. TGase 3 enzymatic activity will also be assessed on protein extracted from oesophagus, the tissue with the highest *TGM3* expression in humans. The earliest experimental cohort will be available in Oct-Dec 2016.

Molecular mechanisms of energy balance of the *CPT1A* humanized mouse model will be investigated in collaboration with Sergio Rodriguez-Cuenca and Antonio Vidal-Puig (Metabolic Research Laboratories, University of Cambridge). Growth curves will be examined under different nutritional and environmental challenges including: i) classical high fat diet; ii) high polyunsaturated fatty acids (PUFAs), medium protein, low carbohydrate diet (to mimic the nutritional macronutrient composition of the artic populations); iii) food shortage/fasting; iv) thermoneutrality (28-30°C); v) cold exposure (4-8°C); vi) progressive acclimation (22/24°C to 16°C to 4°C). This study will be complemented by energy expenditure and respiratory exchange ratio measurements (evaluated using the Metatracer analyser), body composition analysis (using Time Domain Nuclear Magnetic Resonance to evaluate fat percentage and lean mass during the nutritional challenge) and basic blood biochemistry profiling during the different nutritional interventions (focusing on plasma triglycerides, free fatty acids, carnitine, glucose, insulin levels, ketone bodies, and markers of liver damage (ALT/AST ratio)). Carbohydrate and lipid metabolism are of special interest and will be followed up using glucose tolerance tests (GTT), insulin tolerance tests (ITT), lipid tolerance tests and detailed lipid profiling in plasma to evaluate the phospholipid pool and

their oxidative status. These analyses may be complemented by assessment of *Cpt1a* enzymatic activity and liver histopathology. Additionally, our Pro479Leu mouse model might also serve as a model of human sudden unexpected death in infancy (SUDI).

Finally, we plan to look at the hearing of *PCDH15*, *OTOF* and potentially *HERC1* models in great detail in collaboration with Karen Steel (King's College London). Secondary phenotyping planned by Karen Steel will focus on physiological (electrophysiological) differences between the humanized mice and controls, including auditory brainstem response screening (with extended threshold recording), tests of frequency tuning, temporal processing, adaptation, fatigue and distortion product otoacoustic emissions (DPOAEs - to examine sounds emitted in response to two simultaneous tones of different frequencies) amongst others.

In addition, phenotyping efforts will be supplemented with detailed *in silico* protein modelling that would help to understand the molecular impact of selected amino acid substitutions on the protein structure done by Tomek Stepniewski, Ramon Guixa-Gonzalez and Jana Selent (Research Programme on Biomedical Informatics, Department of Experimental and Health Sciences Universitat Pompeu Fabra, Hospital del Mar Medical Research Institute, Barcelona, Spain).

4.3. Future directions

Functional studies exploring mechanistic links between genetic diversity and phenotypic variation should also focus on addressing other modes of selection to fully understand the genetic basis of human adaptation. There are a few well-established examples of genetic variants targeted by balancing selection linked to phenotypic traits (i.e. sickle cell anemia and malaria resistance (62)), but less success has been achieved in pinpointing and validating variants driving soft sweeps and polygenic selection that leave weak signatures of selection (18, 26), although new methods addressing these questions are emerging e.g. the Singleton Density Score (*SDS*) inferring very recent changes in allele frequencies that are able to uncover polygenic shifts affecting complex traits, but also very recent hard and soft sweeps (622). Another challenge is the functional follow-up of variants with small, but non-zero size effects that have been proposed to drive phenotypic variation of many complex traits in an additive fashion (623). It is likely that such variants are insufficient to cause a detectable phenotype in isolation. One possible solution would be engineering multiple such variants in mice through genome editing in isolation and then cross-breeding the progeny in order to accumulate all candidates in one individual to examine additive effect and phenotype amplification, although it seems laborious and time consuming endeavour. Another possibility would be re-evaluating human cohorts with rich phenotyping data in testing hypotheses regarding human adaptation, thus replacing model organisms. New large-scale datasets of densely genotyped or deeply sequenced and extensively phenotyped individuals, such as the UK Biobank dataset (624, 625), might help in the assessment of co-segregation of candidate variant with phenotype directly in humans.

Further challenges include addressing the effects of other forms of genetic variation. Most functional studies focus on heritable coding variation, namely, coding variants in the nucleotide sequence of DNA. Only a handful of regulatory candidates have been functionally validated, but it seems that the bulk of human adaptation is though to be concentrated in non-coding regions driving gene expression levels (17, 140, 141, 162-164, 170-172, 626, 627). Indeed, the majority

of SNPs identified by *FineMAV* fall into regulatory, intronic or intergenic regions, but also in non-coding RNAs (ncRNA). The inability to form prior hypotheses about the function of non-coding DNA is a key factor limiting functional follow-up studies (626). Several ncRNAs have been shown to play important roles in diverse biological processes, but their functions have been largely unexplored in humans (628). It has also been shown that purifying selection has acted on conserved long intergenic ncRNAs, and a fraction of them show signals of selection similar to protein-coding genes (628, 629). A recent study proposed functional prioritisation of the hundreds of putative long ncRNAs for downstream experimental interrogation (628). Exploring signals of selection in ncRNAs is a potential further direction of this project.

Additionally, functional studies also need to tackle structural variants exhibiting signals of selection. Large allele frequency differences between populations have been reported for copy-number variants (CNVs), and this class of variants is in general believed to have contributed to hominid evolution and human adaptation (630) e.g. increase in the copy-number of the salivary amylase gene (*AMY1*) as an adaptation to a high-starch diet (631), duplication of the *HP* and *HPR* genes in Africans associated with protection against trypanosomiasis (630, 632), deletion of *UGT2B17* in East Asians (633), or selectively introgressed CNVs of archaic origin in modern Oceanic populations (630).

Finally, known genetic variation often does not fully explain the observed phenotypic variance, a phenomenon referred to as the 'missing heritability' that has been linked to gene-gene and gene-environment interactions (634). Therefore, efforts aimed at understanding human adaptation should also account for other layers of variation, like epigenetic variation, that can inform about additional mechanisms of human responses to environmental challenges. Epigenetic modifications, and in particular DNA methylations, provide information on gene activity that could contribute to phenotypic variation (635, 636). Methylation occurs on cytosine residues in the context of CpG dinucleotides which are found at gene promoters and can regulate the expression of neighbouring genes (637). A substantial portion of DNA methylation variation is controlled by inherited genetic variation (methylation quantitative trait loci; meQTLs), but it can also be affected by a broad range of environmental factors including habitat and lifestyle (638-646).

Recent studies reported extensive DNA methylation differences between major ethnic groups and a signature of selection on population-specific meQTLs (645-648). We integrated expression quantitative trait loci (eQTLs) data with our results, but have not explored meQTLs annotations, which could be another future expansion of this project. Apart from the heritable aspect of methylation-associated SNPs, it has been proposed that populations can initially respond to environmental challenges via epigenetic changes independently of underlying meQTLs, with the adaptive phenotype being achieved via genetic changes over time (648). Such short-term rapid adaptation is little-recognised and needs further investigation, as the proportion of methylated sites unexplained by underlying SNPs is substantial (643, 646, 648). On the other hand, this first line of adaptation might also influence the epitype of germline cells and potentially impact subsequent generations allowing a response to the environment through changes in gene expression (646).

A lot of work needs to be done to interpret the interplay between genotype, environment and the natural phenotypic variation occurring in the human species. Finding mechanistic links between selection candidates and Darwinian fitness seems crucial in these efforts. Our abilities to generate genetic and phenotype data on vast scales, modify genomes, and develop new analytical approaches are expanding at unprecedented rates. Predictions about the future usually turn out to miss the most important and unexpected new approaches, but there seems every reason to be optimistic that the next few years will see great advances in our understanding of human adaptation.

References

1. Darwin C. On the origin of species by means of natural selection. London,: J. Murray; 1859. ix, 1 , 502 p. p.
2. Mayr E. The growth of biological thought : diversity, evolution, and inheritance. Cambridge, Mass.: Belknap Press; 1982. ix, 974 p. p.
3. Grafen A, Ridley M. Richard Dawkins : how a scientist changed the way we think : reflections by scientists, writers, and philosophers. Oxford ; New York: Oxford University Press; 2006. xiii, 283 p. p.
4. Fisher RA. The genetical theory of natural selection. Oxford,: The Clarendon press; 1930. xiv, 272 p. p.
5. Bowler PJ. The Mendelian revolution : the emergence of hereditarian concepts in modern science and society. Baltimore: Johns Hopkins University Press; 1989. viii, 207 p. p.
6. Provine WB. The origins of theoretical population genetics. Chicago,: University of Chicago Press; 1971. xi, 201 p. p.
7. Huxley J. Evolution, the modern synthesis. London,: G. Allen & Unwin ltd; 1942. 645, 1 p. p.
8. Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome research. 2009;19(5):711-22.
9. Nielsen R. Molecular signatures of natural selection. Annual review of genetics. 2005;39:197-218.
10. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annual review of genetics. 2013;47:97-120.
11. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289-303.
12. Charlesworth B. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genetical research. 1994;63(3):213-27.
13. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS genetics. 2006;2(4):e64.
14. Kimura M. The neutral theory of molecular evolution. Cambridge Cambridgeshire ; New York: Cambridge University Press; 1983. xv, 367 p. p.
15. Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624-6.
16. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science. 2006;312(5780):1614-20.
17. Scheinfeldt LB, Tishkoff SA. Recent human adaptation: genomic approaches, interpretation and insights. Nature reviews Genetics. 2013;14(10):692-702.
18. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current biology : CB. 2010;20(4):R208-15.
19. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetical research. 1974;23(1):23-35.

20. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. Science. 2011;331(6019):920-4.
21. Novembre J, Di Rienzo A. Spatial patterns of variation due to natural selection in humans. Nature reviews Genetics. 2009;10(11):745-55.
22. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. Nature reviews Genetics. 2007;8(11):857-68.
23. Jeong C, Di Rienzo A. Adaptations to local environments in modern human populations. Current opinion in genetics & development. 2014;29:1-8.
24. Schlotterer C. Towards a molecular characterization of adaptation in local populations. Current opinion in genetics & development. 2002;12(6):683-7.
25. Przeworski M, Coop G, Wall JD. The signature of positive selection on standing genetic variation. Evolution; international journal of organic evolution. 2005;59(11):2312-23.
26. Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005;169(4):2335-52.
27. Turchin MC, Chiang CW, Palmer CD, Sankararaman S, Reich D, Genetic Investigation of ATC, et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nature genetics. 2012;44(9):1015-9.
28. Berg JJ, Coop G. A population genetic signal of polygenic adaptation. PLoS genetics. 2014;10(8):e1004412.
29. Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, et al. Population genetic differentiation of height and body mass index across Europe. Nature genetics. 2015;47(11):1357-62.
30. Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E. Evidence for archaic adaptive introgression in humans. Nature reviews Genetics. 2015;16(6):359-71.
31. Jobling MA, Hurles M, Tyler-Smith C. Human evolutionary genetics. 2nd ed. New York: Garland Science; 2013. xviii, 670 p. p.
32. Bamshad M, Wooding S, Salisbury BA, Stephens JC. Deconstructing the relationship between genetics and race. Nature reviews Genetics. 2004;5(8):598-609.
33. Jablonski NG, Chaplin G. The evolution of skin pigmentation and hair texture in people of African ancestry. Dermatologic clinics. 2014;32(2):113-21.
34. Perry GH, Dominy NJ. Evolution of the human pygmy phenotype. Trends Ecol Evol. 2009;24(4):218-25.
35. Beall CM. Tibetan and Andean contrasts in adaptation to high-altitude hypoxia. Adv Exp Med Biol. 2000;475:63-74.
36. Moore LG. Human genetic adaptation to high altitude. High Alt Med Biol. 2001;2(2):257-79.
37. Li WH, Sadler LA. Low nucleotide diversity in man. Genetics. 1991;129(2):513-23.
38. Harpending H, Rogers A. Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet. 2000;1:361-85.
39. Fischer A, Wiebe V, Paabo S, Przeworski M. Evidence for a complex demographic history of chimpanzees. Molecular biology and evolution. 2004;21(5):799-808.

40. Kittles RA, Weiss KM. Race, ancestry, and genes: implications for defining disease risk. Annu Rev Genomics Hum Genet. 2003;4:33-67.
41. Youmans WJ. Appleton's popular science monthly. New York: D. Appleton and Co.; 1895.
42. Rodriguez JA, Marigorta UM, Navarro A. Integrating genomics into evolutionary medicine. Current opinion in genetics & development. 2014;29:97-102.
43. Wilson TW, Grim CE. Biohistory of slavery and blood pressure differences in blacks today. A hypothesis. Hypertension. 1991;17(1 Suppl):I122-8.
44. Neel JV. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? American journal of human genetics. 1962;14:353-62.
45. Di Rienzo A, Hudson RR. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends in genetics : TIG. 2005;21(11):596-601.
46. Minster RL, Hawley NL, Su CT, Sun G, Kershaw EE, Cheng H, et al. A thrifty variant in CREBRF strongly influences body mass index in Samoans. Nature genetics. 2016;48(9):1049-54.
47. Ayub Q, Moutsianas L, Chen Y, Panoutsopoulou K, Colonna V, Pagani L, et al. Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. American journal of human genetics. 2014;94(2):176-85.
48. Williams GC, Nesse RM. The dawn of Darwinian medicine. Q Rev Biol. 1991;66(1):1-22.
49. Ewald PW. Evolutionary biology and the treatment of signs and symptoms of infectious disease. J Theor Biol. 1980;86(1):169-76.
50. Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, et al. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. American journal of human genetics. 2006;78(4):659-70.
51. Novembre J, Galvani AP, Slatkin M. The geographic spread of the CCR5 Delta32 HIV-resistance allele. PLoS biology. 2005;3(11):e339.
52. Lindesmith L, Moe C, Marionneau S, Ruvoen N, Jiang X, Lindblad L, et al. Human susceptibility and resistance to Norwalk virus infection. Nature medicine. 2003;9(5):548-53.
53. Raj T, Kuchroo M, Replogle JM, Raychaudhuri S, Stranger BE, De Jager PL. Common risk alleles for inflammatory diseases are targets of recent positive selection. American journal of human genetics. 2013;92(4):517-29.
54. Biswas S, Akey JM. Genomic insights into positive selection. Trends in genetics : TIG. 2006;22(8):437-46.
55. Cardona A, Pagani L, Antao T, Lawson DJ, Eichstaedt CA, Yngvadottir B, et al. Genome-wide analysis of cold adaptation in indigenous Siberian populations. PloS one. 2014;9(5):e98076.
56. Clemente FJ, Cardona A, Inchley CE, Peter BM, Jacobs G, Pagani L, et al. A Selective Sweep on a Deleterious Mutation in CPT1A in Arctic Populations. American journal of human genetics. 2014;95(5):584-9.
57. Ko WY, Rajan P, Gomez F, Scheinfeldt L, An P, Winkler CA, et al. Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. American journal of human genetics. 2013;93(1):54-66.

58. Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science. 2010;329(5993):841-5.
59. Spyropoulos B, Moens PB, Davidson J, Lowden JA. Heterozygote advantage in Tay-Sachs carriers? American journal of human genetics. 1981;33(3):375-80.
60. Schroeder SA, Gaughan DM, Swift M. Protection against bronchial asthma by CFTR delta F508 mutation: a heterozygote advantage in cystic fibrosis. Nature medicine. 1995;1(7):703-5.
61. Woolf LI, McBean MS, Woolf FM, Cahalane SF. Phenylketonuria as a balanced polymorphism: the nature of the heterozygote advantage. Ann Hum Genet. 1975;38(4):461-9.
62. Allison AC. Protection afforded by sickle-cell trait against subtertian malareal infection. British medical journal. 1954;1(4857):290-4.
63. Bamshad M, Wooding SP. Signatures of natural selection in the human genome. Nature reviews Genetics. 2003;4(2):99-111.
64. Gillespie JH. Genetic drift in an infinite population. The pseudohitchhiking model. Genetics. 2000;155(2):909-19.
65. Gillespie JH. Is the population size of a species relevant to its evolution? Evolution; international journal of organic evolution. 2001;55(11):2161-9.
66. Zlotogora J. Multiple mutations responsible for frequent genetic diseases in isolated populations. European journal of human genetics : EJHG. 2007;15(3):272-8.
67. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304(5670):581-4.
68. Cavalli-Sforza LL. Population structure and human evolution. Proc R Soc Lond B Biol Sci. 1966;164(995):362-79.
69. Burbano HA, Green RE, Maricic T, Lalueza-Fox C, de la Rasilla M, Rosas A, et al. Analysis of human accelerated DNA regions using archaic hominin genomes. PloS one. 2012;7(3):e32877.
70. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011;478(7370):476-82.
71. Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, et al. Forces shaping the fastest evolving regions in the human genome. PLoS genetics. 2006;2(10):e168.
72. Prabhakar S, Noonan JP, Paabo S, Rubin EM. Accelerated evolution of conserved noncoding sequences in humans. Science. 2006;314(5800):786.
73. Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 1998;148(3):929-36.
74. Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. Molecular biology and evolution. 1999;16(10):1315-28.
75. Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular biology and evolution. 2002;19(6):908-17.

76. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular biology and evolution. 2000;17(1):32-43.
77. Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 1998;46(4):409-18.
78. Wong WS, Nielsen R. Detecting selection in noncoding regions of nucleotide sequences. Genetics. 2004;167(2):949-58.
79. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351(6328):652-4.
80. Hudson RR, Kreitman M, Aguade M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116(1):153-9.
81. Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics. 1973;74(1):175-95.
82. Wright S. The genetical structure of populations. Ann Eugen. 1951;15(4):323-54.
83. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. Genome research. 2002;12(12):1805-14.
84. Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol. 2004;13(4):969-80.
85. Porter AH. A test for deviation from island-model population structure. Mol Ecol. 2003;12(4):903-15.
86. Kayser M, Brauer S, Stoneking M. A genome scan to detect candidate regions influenced by local natural selection in human populations. Molecular biology and evolution. 2003;20(6):893-900.
87. Vitalis R, Dawson K, Boursot P. Interpretation of variation across marker loci as evidence of selection. Genetics. 2001;158(4):1811-23.
88. Meirmans PG, Hedrick PW. Assessing population structure: F(ST) and related measures. Mol Ecol Resour. 2011;11(1):5-18.
89. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, et al. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics. 2010;186(1):241-62.
90. Excoffier L, Hofer T, Foll M. Detecting loci under selection in a hierarchically structured population. Heredity (Edinb). 2009;103(4):285-98.
91. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, et al. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics. 2004;1(4):274-86.
92. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science. 2010;329(5987):75-8.
93. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome research. 2010;20(3):393-402.
94. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics. 2013;193(3):929-41.
95. Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, et al. Human genomic regions with exceptionally high levels of population differentiation

identified from 911 whole-genome sequences. Genome biology. 2014;15(6):R88.
96. Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterlander M, et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(13):4832-7.
97. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528(7583):499-503.
98. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585-95.
99. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993;133(3):693-709.
100. Fu YX. New statistical tests of neutrality for DNA samples from a population. Genetics. 1996;143(1):557-70.
101. Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 1997;147(2):915-25.
102. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000;155(3):1405-13.
103. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 2002;160(2):765-77.
104. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419(6909):832-7.
105. Depaulis F, Veuille M. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Molecular biology and evolution. 1998;15(12):1788-90.
106. Kelly JK. A test of neutrality based on interlocus associations. Genetics. 1997;146(3):1197-206.
107. Kim Y, Nielsen R. Linkage disequilibrium as a signature of selective sweeps. Genetics. 2004;167(3):1513-24.
108. Zhang C, Bailey DK, Awad T, Liu G, Xing G, Cao M, et al. A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinformatics. 2006;22(17):2122-8.
109. Hanchard NA, Rockett KA, Spencer C, Coop G, Pinder M, Jallow M, et al. Screening for recently selected alleles by analysis of human haplotype similarity. American journal of human genetics. 2006;78(1):153-9.
110. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS biology. 2006;4(3):e72.
111. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449(7164):913-8.
112. Tang K, Thornton KR, Stoneking M. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS biology. 2007;5(7):e171.
113. Albrechtsen A, Moltke I, Nielsen R. Natural selection and the distribution of identity-by-descent in the human genome. Genetics. 2010;186(1):295-308.

114. Cai Z, Camp NJ, Cannon-Albright L, Thomas A. Identification of regions of positive selection using Shared Genomic Segment analysis. European journal of human genetics : EJHG. 2011;19(6):667-71.
115. Han L, Abney M. Using identity by descent estimation with dense genotype data to detect positive selection. European journal of human genetics : EJHG. 2013;21(2):205-11.
116. Wiener P, Pong-Wong R. A regression-based approach to selection mapping. J Hered. 2011;102(3):294-305.
117. Zeng K, Fu YX, Shi S, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 2006;174(3):1431-9.
118. Ayub Q, Yngvadottir B, Chen Y, Xue Y, Hu M, Vernes SC, et al. FOXP2 targets show evidence of positive selection in European populations. American journal of human genetics. 2013;92(5):696-706.
119. Ewens WJ. The sampling theory of selectively neutral alleles. Theoretical population biology. 1972;3(1):87-112.
120. Watterson GA. The homozygosity test of neutrality. Genetics. 1978;88(2):405-17.
121. Zeng K, Shi S, Wu CI. Compound tests for the detection of hitchhiking under positive selection. Molecular biology and evolution. 2007;24(8):1898-908.
122. Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andres AM, Albrechtsen A, et al. Darwinian and demographic forces affecting human protein coding genes. Genome research. 2009;19(5):838-49.
123. Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science. 2010;327(5967):883-6.
124. Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, et al. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics. 2015;31(24):3946-52.
125. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710-22.
126. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Molecular biology and evolution. 2011;28(8):2239-52.
127. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192(3):1065-93.
128. Plagnol V, Wall JD. Possible ancestral structure in human populations. PLoS genetics. 2006;2(7):e105.
129. Vernot B, Akey JM. Resurrecting surviving Neandertal lineages from modern human genomes. Science. 2014;343(6174):1017-21.
130. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, Moltke I, et al. Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. Science. 2014;346(6213):1113-8.
131. Sankararaman S, Patterson N, Li H, Paabo S, Reich D. The date of interbreeding between Neandertals and modern humans. PLoS genetics. 2012;8(10):e1002947.
132. Mendez FL, Watkins JC, Hammer MF. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. American journal of human genetics. 2012;91(2):265-74.

133. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505(7481):43-9.
134. Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507(7492):354-7.
135. Pavlidis P, Jensen JD, Stephan W, Stamatakis A. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. Molecular biology and evolution. 2012;29(10):3237-48.
136. Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome research. 2006;16(8):980-9.
137. Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2010;365(1537):185-205.
138. Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, et al. Modeling recent human evolution in mice by expression of a selected EDAR variant. Cell. 2013;152(4):691-702.
139. Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science. 2005;310(5755):1782-6.
140. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I. Identification of a variant associated with adult-type hypolactasia. Nature genetics. 2002;30(2):233-7.
141. Olds LC, Sibley E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Human molecular genetics. 2003;12(18):2333-40.
142. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
143. Liu X, Ong RT, Pillai EN, Elzein AM, Small KS, Clark TG, et al. Detecting and characterizing genomic signatures of positive selection in global populations. American journal of human genetics. 2013;92(6):866-81.
144. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. Natural selection has driven population differentiation in modern humans. Nature genetics. 2008;40(3):340-5.
145. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, et al. Genomic regions exhibiting positive selection identified from dense genotype data. Genome research. 2005;15(11):1553-65.
146. International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449(7164):851-61.
147. Fagny M, Patin E, Enard D, Barreiro LB, Quintana-Murci L, Laval G. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Molecular biology and evolution. 2014;31(7):1850-68.
148. Hofer T, Foll M, Excoffier L. Evolutionary forces shaping genomic islands of population differentiation in humans. BMC genomics. 2012;13:107.
149. Tennessen JA, Akey JM. Parallel adaptive divergence among geographically diverse human populations. PLoS genetics. 2011;7(6):e1002127.

150. Johansson A, Gyllensten U. Identification of local selective sweeps in human populations since the exodus from Africa. Hereditas. 2008;145(3):126-37.
151. Kimura R, Fujimoto A, Tokunaga K, Ohashi J. A practical genome scan for population-specific strong selective sweeps that have reached fixation. PloS one. 2007;2(3):e286.
152. Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, et al. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PloS one. 2009;4(11):e7888.
153. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome research. 2009;19(5):826-37.
154. Rafajlovic M, Klassmann A, Eriksson A, Wiehe T, Mehlig B. Demography-adjusted tests of neutrality based on genome-wide SNP data. Theoretical population biology. 2014;95:1-12.
155. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, et al. Identifying recent adaptations in large-scale genomic data. Cell. 2013;152(4):703-13.
156. Zhong M, Lange K, Papp JC, Fan R. A powerful score test to detect positive selection in genome-wide scans. European journal of human genetics : EJHG. 2010;18(10):1148-59.
157. Wang ET, Kodama G, Baldi P, Moyzis RK. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(1):135-40.
158. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. Localizing recent adaptive evolution in the human genome. PLoS genetics. 2007;3(6):e90.
159. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
160. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic acids research. 2015;43(Database issue):D662-9.
161. Mou C, Thomason HA, Willan PM, Clowes C, Harris WE, Drew CF, et al. Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form. Human mutation. 2008;29(12):1405-11.
162. Sturm RA, Duffy DL, Zhao ZZ, Leite FP, Stark MS, Hayward NK, et al. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. American journal of human genetics. 2008;82(2):424-31.
163. Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer KW, et al. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. Human genetics. 2008;123(2):177-87.
164. Visser M, Kayser M, Palstra RJ. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome research. 2012;22(3):446-55.
165. Engelken J, Carnero-Montoro E, Pybus M, Andrews GK, Lalueza-Fox C, Comas D, et al. Extreme population differences in the human zinc transporter ZIP4

(SLC39A4) are explained by positive selection in Sub-Saharan Africa. PLoS genetics. 2014;10(2):e1004128.
166. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319(5866):1100-4.
167. Peacock E, Whiteley P. Perlegen sciences, inc. Pharmacogenomics. 2005;6(4):439-42.
168. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics. 2014;46(3):310-5.
169. Breiman L. Classification and regression trees. Belmont, Calif.: Wadsworth International Group; 1984. x, 358 p. p.
170. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. 2008;134(1):25-36.
171. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975;188(4184):107-16.
172. Wray GA. The evolutionary significance of cis-regulatory mutations. Nature reviews Genetics. 2007;8(3):206-16.
173. Tournamille C, Colin Y, Cartron JP, Le Van Kim C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nature genetics. 1995;10(2):224-8.
174. Iwamoto S, Li J, Omi T, Ikemoto S, Kajii E. Identification of a novel exon and spliced form of Duffy mRNA that is the predominant transcript in both erythroid and postcapillary venule endothelium. Blood. 1996;87(1):378-85.
175. Iwamoto S, Li J, Sugimoto N, Okuda H, Kajii E. Characterization of the Duffy gene promoter: evidence for tissue-specific abolishment of expression in Fy(a-b-) of black individuals. Biochemical and biophysical research communications. 1996;222(3):852-9.
176. Miller LH, Mason SJ, Clyde DF, McGinniss MH. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. The New England journal of medicine. 1976;295(6):302-4.
177. Yoshiura K, Kinoshita A, Ishida T, Ninokata A, Ishikawa T, Kaname T, et al. A SNP in the ABCC11 gene is the determinant of human earwax type. Nature genetics. 2006;38(3):324-30.
178. Martin A, Saathoff M, Kuhn F, Max H, Terstegen L, Natsch A. A functional ABCC11 allele is essential in the biochemical formation of human axillary odor. The Journal of investigative dermatology. 2010;130(2):529-40.
179. Tsetskhladze ZR, Canfield VA, Ang KC, Wentzel SM, Reid KP, Berg AS, et al. Functional assessment of human coding mutations affecting skin pigmentation using zebrafish. PloS one. 2012;7(10):e47398.
180. Graf J, Hodgson R, van Daal A. Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. Human mutation. 2005;25(3):278-84.
181. Cook AL, Chen W, Thurber AE, Smit DJ, Smith AG, Bladen TG, et al. Analysis of cultured human melanocytes based on polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci. The Journal of investigative dermatology. 2009;129(2):392-405.

182. Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome research. 2008;18(11):1829-43.
183. Peng B, Kimmel M. simuPOP: a forward-time population genetics simulation environment. Bioinformatics. 2005;21(18):3686-7.
184. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS genetics. 2009;5(10):e1000695.
185. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.
186. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nature genetics. 2015;47(3):276-83.
187. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nature methods. 2014;11(3):294-6.
188. Edwards M, Bigham A, Tan J, Li S, Gozdzik A, Ross K, et al. Association of the OCA2 polymorphism His615Arg with melanin content in east Asian populations: further evidence of convergent evolution of skin pigmentation. PLoS genetics. 2010;6(3):e1000867.
189. Eaton K, Edwards M, Krithika S, Cook G, Norton H, Parra EJ. Association study confirms the role of two OCA2 polymorphisms in normal skin pigmentation variation in East Asian populations. American journal of human biology : the official journal of the Human Biology Council. 2015;27(4):520-5.
190. Donnelly MP, Paschou P, Grigorenko E, Gurwitz D, Barta C, Lu RB, et al. A global view of the OCA2-HERC2 region and pigmentation. Human genetics. 2012;131(5):683-96.
191. Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, et al. A genomewide association study of skin pigmentation in a South Asian population. American journal of human genetics. 2007;81(6):1119-32.
192. Hutton SM, Spritz RA. A comprehensive genetic study of autosomal recessive ocular albinism in Caucasian patients. Investigative ophthalmology & visual science. 2008;49(3):868-72.
193. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nature genetics. 2007;39(12):1443-52.
194. Visser M, Palstra RJ, Kayser M. Human skin color is influenced by an intergenic DNA polymorphism regulating transcription of the nearby BNC2 pigmentation gene. Human molecular genetics. 2014;23(21):5750-62.
195. Kudo T, Iwasaki H, Nishihara S, Shinya N, Ando T, Narimatsu I, et al. Molecular genetic analysis of the human Lewis histo-blood group system. II. Secretor gene inactivation by a novel single missense mutation A385T in Japanese nonsecretor individuals. The Journal of biological chemistry. 1996;271(16):9830-7.
196. Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacon-Duque JC, et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nature communications. 2016;7:10815.

197. Mackenzie FE, Romero R, Williams D, Gillingwater T, Hilton H, Dick J, et al. Upregulation of PKD1L2 provokes a complex neuromuscular disease in the mouse. Human molecular genetics. 2009;18(19):3553-66.
198. Oh HJ, Li Y, Lau YF. Sry associates with the heterochromatin protein 1 complex by interacting with a KRAB domain protein. Biology of reproduction. 2005;72(2):407-15.
199. Young P, Ehler E, Gautel M. Obscurin, a giant sarcomeric Rho guanine nucleotide exchange factor protein involved in sarcomere assembly. The Journal of cell biology. 2001;154(1):123-36.
200. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648-60.
201. Randazzo D, Giacomello E, Lorenzini S, Rossi D, Pierantozzi E, Blaauw B, et al. Obscurin is required for ankyrinB-dependent dystrophin localization and sarcolemma integrity. The Journal of cell biology. 2013;200(4):523-36.
202. Tardif S, Wilson MD, Wagner R, Hunt P, Gertsenstein M, Nagy A, et al. Zonadhesin is essential for species specificity of sperm adhesion to the egg zona pellucida. The Journal of biological chemistry. 2010;285(32):24863-70.
203. Wassarman PM, Jovine L, Litscher ES. A profile of fertilization in mammals. Nature cell biology. 2001;3(2):E59-64.
204. Gasper J, Swanson WJ. Molecular population genetics of the gene encoding the human fertilization protein zonadhesin reveals rapid adaptive evolution. American journal of human genetics. 2006;79(5):820-30.
205. Tardif S, Brady HA, Breazeale KR, Bi M, Thompson LD, Bruemmer JE, et al. Zonadhesin D3-polypeptides vary among species but are similar in Equus species capable of interbreeding. Biology of reproduction. 2010;82(2):413-21.
206. Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, et al. A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. American journal of human genetics. 2009;84(2):224-34.
207. Thibaut S, Cavusoglu N, de Becker E, Zerbib F, Bednarczyk A, Schaeffer C, et al. Transglutaminase-3 enzyme: a putative actor in human hair shaft scaffolding? The Journal of investigative dermatology. 2009;129(2):449-59.
208. John S, Thiebach L, Frie C, Mokkapati S, Bechtel M, Nischt R, et al. Epidermal transglutaminase (TGase 3) is required for proper hair development, but not the formation of the epidermal barrier. PloS one. 2012;7(4):e34252.
209. Bognar P, Nemeth I, Mayer B, Haluszka D, Wikonkal N, Ostorhazi E, et al. Reduced inflammatory threshold indicates skin barrier defect in transglutaminase 3 knockout mice. The Journal of investigative dermatology. 2014;134(1):105-11.
210. Brennan BM, Huynh MT, Rabah MA, Shaw HE, Bisaillon JJ, Radden LA, 2nd, et al. The mouse wellhaarig (we) mutations result from defects in epidermal-type transglutaminase 3 (Tgm3). Molecular genetics and metabolism. 2015;116(3):187-91.
211. Steinert PM, Parry DA, Marekov LN. Trichohyalin mechanically strengthens the hair follicle: multiple cross-bridging roles in the inner root shealth. The Journal of biological chemistry. 2003;278(42):41409-19.
212. FB UB, Cau L, Tafazzoli A, Mechin MC, Wolf S, Romano MT, et al. Mutations in Three Genes Encoding Proteins Involved in Hair Shaft Formation Cause

Uncombable Hair Syndrome. American journal of human genetics. 2016;99(6):1292-304.
213. Fujimoto A, Nishida N, Kimura R, Miyagawa T, Yuliwulandari R, Batubara L, et al. FGFR2 is associated with hair thickness in Asian populations. Journal of human genetics. 2009;54(8):461-5.
214. Laatsch CN, Durbin-Johnson BP, Rocke DM, Mukwana S, Newland AB, Flagler MJ, et al. Human hair shaft proteomic profiling: individual differences, site specificity and cuticle analysis. PeerJ. 2014;2:e506.
215. Robbins CR. Chemical and physical behavior of human hair. 4th ed. New York: Springer; 2002. xvii, 483 p. p.
216. Iolascon A, King MJ, Robertson S, Avvisati RA, Vitiello F, Asci R, et al. A genomic deletion causes truncation of alpha-spectrin and ellipto-poikilocytosis. Blood cells, molecules & diseases. 2011;46(3):195-200.
217. Salomao M, An X, Guo X, Gratzer WB, Mohandas N, Baines AJ. Mammalian alpha I-spectrin is a neofunctionalized polypeptide adapted to small highly deformable erythrocytes. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(3):643-8.
218. Burke JP, Van Zyl D, Zail SS, Coetzer TL. Reduced spectrin-ankyrin binding in a South African hereditary elliptocytosis kindred homozygous for spectrin St Claude. Blood. 1998;92(7):2591-2.
219. Soranzo N, Sanna S, Wheeler E, Gieger C, Radke D, Dupuis J, et al. Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. Diabetes. 2010;59(12):3229-39.
220. Ding K, Shameer K, Jouni H, Masys DR, Jarvik GP, Kho AN, et al. Genetic Loci implicated in erythroid differentiation and cell cycle regulation are associated with red blood cell traits. Mayo Clinic proceedings. 2012;87(5):461-74.
221. Ganesh SK, Zakai NA, van Rooij FJ, Soranzo N, Smith AV, Nalls MA, et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nature genetics. 2009;41(11):1191-8.
222. Birkenmeier CS, McFarland-Starr EC, Barker JE. Chromosomal location of three spectrin genes: relationship to the inherited hemolytic anemias of mouse and man. Proceedings of the National Academy of Sciences of the United States of America. 1988;85(21):8121-5.
223. Grossmann A, Maggio-Price L, Shiota FM, Liggitt DV. Pathologic features associated with decreased longevity of mutant sphha/sphha mice with chronic hemolytic anemia: similarities to sequelae of sickle cell anemia in humans. Laboratory animal science. 1993;43(3):217-21.
224. Lecomte MC, Dhermy D, Gautero H, Bournier O, Galand C, Boivin P. [Hereditary elliptocytosis in West Africa: frequency and repartition of spectrin variants]. Comptes rendus de l'Academie des sciences Serie III, Sciences de la vie. 1988;306(2):43-6.
225. Delaunay J, Dhermy D. Mutations involving the spectrin heterodimer contact site: clinical expression and alterations in specific function. Seminars in hematology. 1993;30(1):21-33.
226. Delaunay J. The molecular basis of hereditary red cell membrane disorders. Blood reviews. 2007;21(1):1-20.
227. Schulman S, Roth EF, Jr., Cheng B, Rybicki AC, Sussman, II, Wong M, et al. Growth of Plasmodium falciparum in human erythrocytes containing

abnormal membrane proteins. Proceedings of the National Academy of Sciences of the United States of America. 1990;87(18):7339-43.
228. Shear HL, Roth EF, Jr., Ng C, Nagel RL. Resistance to malaria in ankyrin and spectrin deficient mice. British journal of haematology. 1991;78(4):555-60.
229. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506-11.
230. Spiekerkoetter U, Sun B, Khuchua Z, Bennett MJ, Strauss AW. Molecular and phenotypic heterogeneity in mitochondrial trifunctional protein deficiency due to beta-subunit mutations. Human mutation. 2003;21(6):598-607.
231. Purevsuren J, Fukao T, Hasegawa Y, Kobayashi H, Li H, Mushimoto Y, et al. Clinical and molecular aspects of Japanese patients with mitochondrial trifunctional protein deficiency. Molecular genetics and metabolism. 2009;98(4):372-7.
232. Naiki M, Ochi N, Kato YS, Purevsuren J, Yamada K, Kimura R, et al. Mutations in HADHB, which encodes the beta-subunit of mitochondrial trifunctional protein, cause infantile onset hypoparathyroidism and peripheral polyneuropathy. American journal of medical genetics Part A. 2014;164A(5):1180-7.
233. Valbuena A, Lopez-Sanchez I, Lazo PA. Human VRK1 is an early response gene and its loss causes a block in cell cycle progression. PloS one. 2008;3(2):e1642.
234. Valbuena A, Sanz-Garcia M, Lopez-Sanchez I, Vega FM, Lazo PA. Roles of VRK1 as a new player in the control of biological processes required for cell division. Cell Signal. 2011;23(8):1267-72.
235. Lancaster OM, Breuer M, Cullen CF, Ito T, Ohkura H. The meiotic recombination checkpoint suppresses NHK-1 kinase to prevent reorganisation of the oocyte nucleus in Drosophila. PLoS genetics. 2010;6(10):e1001179.
236. Waters K, Yang AZ, Reinke V. Genome-wide analysis of germ cell proliferation in C.elegans identifies VRK-1 as a key regulator of CEP-1/p53. Dev Biol. 2010;344(2):1011-25.
237. Dobrzynska A, Askjaer P. Vaccinia-related kinase 1 is required for early uterine development in Caenorhabditis elegans. Dev Biol. 2016;411(2):246-56.
238. Ivanovska I, Khandan T, Ito T, Orr-Weaver TL. A histone code in meiosis: the histone kinase, NHK-1, is required for proper chromosomal architecture in Drosophila oocytes. Genes Dev. 2005;19(21):2571-82.
239. Kim J, Choi YH, Chang S, Kim KT, Je JH. Defective folliculogenesis in female mice lacking Vaccinia-related kinase 1. Sci Rep. 2012;2:468.
240. Schober CS, Aydiner F, Booth CJ, Seli E, Reinke V. The kinase VRK1 is required for normal meiotic progression in mammalian oogenesis. Mech Dev. 2011;128(3-4):178-90.
241. Choi YH, Park CH, Kim W, Ling H, Kang A, Chang MW, et al. Vaccinia-related kinase 1 is required for the maintenance of undifferentiated spermatogonia in mouse male germ cells. PloS one. 2010;5(12):e15254.
242. Wiebe MS, Nichols RJ, Molitor TP, Lindgren JK, Traktman P. Mice deficient in the serine/threonine protein kinase VRK1 are infertile due to a progressive loss of spermatogonia. Biology of reproduction. 2010;82(1):182-93.

243. Ionita-Laza I, Capanu M, De Rubeis S, McCallum K, Buxbaum JD. Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism. PLoS genetics. 2014;10(12):e1004729.
244. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-76.
245. Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova JL, et al. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. American journal of human genetics. 2016;98(1):5-21.
246. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338(6104):222-6.
247. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. American journal of human genetics. 2015;96(6):986-91.
248. Liu Y, DeBoer K, de Kretser DM, O'Donnell L, O'Connor AE, Merriner DJ, et al. LRGUK-1 is required for basal body and manchette function during spermatogenesis and male fertility. PLoS genetics. 2015;11(3):e1005090.
249. Wu BY, Lee SP, Hsiao HC, Chiu H, Chen CY, Yeo YH, et al. Matriptase expression and zymogen activation in human pilosebaceous unit. J Histochem Cytochem. 2014;62(1):50-9.
250. Alef T, Torres S, Hausser I, Metze D, Tursen U, Lestringant GG, et al. Ichthyosis, follicular atrophoderma, and hypotrichosis caused by mutations in ST14 is associated with impaired profilaggrin processing. The Journal of investigative dermatology. 2009;129(4):862-9.
251. List K, Currie B, Scharschmidt TC, Szabo R, Shireman J, Molinolo A, et al. Autosomal ichthyosis with hypotrichosis syndrome displays low matriptase proteolytic activity and is phenocopied in ST14 hypomorphic mice. The Journal of biological chemistry. 2007;282(50):36714-23.
252. Basel-Vanagaite L, Attia R, Ishida-Yamamoto A, Rainshtein L, Ben Amitai D, Lurie R, et al. Autosomal recessive ichthyosis with hypotrichosis caused by a mutation in ST14, encoding type II transmembrane serine protease matriptase. American journal of human genetics. 2007;80(3):467-77.
253. List K, Szabo R, Molinolo A, Nielsen BS, Bugge TH. Delineation of matriptase protein expression by enzymatic gene trapping suggests diverging roles in barrier function, hair formation, and squamous cell carcinogenesis. Am J Pathol. 2006;168(5):1513-25.
254. List K, Haudenschild CC, Szabo R, Chen W, Wahl SM, Swaim W, et al. Matriptase/MT-SP1 is required for postnatal survival, epidermal barrier function, hair follicle development, and thymic homeostasis. Oncogene. 2002;21(23):3765-79.
255. Harris H. The relation of hair-growth on the body to baldness. Br J Dermatol Syph. 1947;59(8-9):300-9.
256. Hindley SW, Damon A. Some genetic traits in Solomon Island populations. IV. Mid-phalangeal hair. Am J Phys Anthropol. 1973;39(2):191-4.

257. Chen YW, Wang JK, Chou FP, Wu BY, Hsiao HC, Chiu H, et al. Matriptase regulates proliferation and early, but not terminal, differentiation of human keratinocytes. The Journal of investigative dermatology. 2014;134(2):405-14.
258. Buzza MS, Netzel-Arnett S, Shea-Donohue T, Zhao A, Lin CY, List K, et al. Membrane-anchored serine protease matriptase regulates epithelial barrier formation and permeability in the intestine. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(9):4200-5.
259. List K, Kosa P, Szabo R, Bey AL, Wang CB, Molinolo A, et al. Epithelial integrity is maintained by a matriptase-dependent proteolytic pathway. Am J Pathol. 2009;175(4):1453-63.
260. Szabo R, Kosa P, List K, Bugge TH. Loss of matriptase suppression underlies spint1 mutation-associated ichthyosis and postnatal lethality. Am J Pathol. 2009;174(6):2015-22.
261. Scharschmidt TC, List K, Grice EA, Szabo R, Program NCS, Renaud G, et al. Matriptase-deficient mice exhibit ichthyotic skin with a selective shift in skin microbiota. The Journal of investigative dermatology. 2009;129(10):2435-42.
262. Netzel-Arnett S, Currie BM, Szabo R, Lin CY, Chen LM, Chai KX, et al. Evidence for a matriptase-prostasin proteolytic cascade regulating terminal epidermal differentiation. The Journal of biological chemistry. 2006;281(44):32941-5.
263. Yin H, Kosa P, Liu X, Swaim WD, Lai Z, Cabrera-Perez J, et al. Matriptase deletion initiates a Sjogren's syndrome-like disease in mice. PloS one. 2014;9(2):e82852.
264. Beaulieu A, Gravel E, Cloutier A, Marois I, Colombo E, Desilets A, et al. Matriptase proteolytically activates influenza virus and promotes multicycle replication in the human airway epithelium. J Virol. 2013;87(8):4237-51.
265. Baron J, Tarnow C, Mayoli-Nussle D, Schilling E, Meyer D, Hammami M, et al. Matriptase, HAT, and TMPRSS2 activate the hemagglutinin of H9N2 influenza A viruses. J Virol. 2013;87(3):1811-20.
266. Hamilton BS, Gludish DW, Whittaker GR. Cleavage activation of the human-adapted influenza virus subtypes by matriptase reveals both subtype and strain specificities. J Virol. 2012;86(19):10579-86.
267. Lebreton A, Lakisic G, Job V, Fritsch L, Tham TN, Camejo A, et al. A bacterial protein targets the BAHD1 chromatin complex to stimulate type III interferon response. Science. 2011;331(6022):1319-21.
268. Zhu C, Xiao F, Hong J, Wang K, Liu X, Cai D, et al. EFTUD2 Is a Novel Innate Immune Regulator Restricting Hepatitis C Virus Infection through the RIG-I/MDA5 Pathway. J Virol. 2015;89(13):6608-18.
269. Kim HJ, Kim CH, Kim MJ, Ryu JH, Seong SY, Kim S, et al. The Induction of Pattern-Recognition Receptor Expression against Influenza A Virus through Duox2-Derived Reactive Oxygen Species in Nasal Mucosa. Am J Respir Cell Mol Biol. 2015;53(4):525-35.
270. Ebrahim M, Mirzaei V, Bidaki R, Shabani Z, Daneshvar H, Karimi-Googheri M, et al. Are RIG-1 and MDA5 Expressions Associated with Chronic HBV Infection? Viral Immunol. 2015;28(9):504-8.

271. Cao X, Ding Q, Lu J, Tao W, Huang B, Zhao Y, et al. MDA5 plays a critical role in interferon response during hepatitis C virus infection. J Hepatol. 2015;62(4):771-8.
272. Hoffmann FS, Schmidt A, Dittmann Chevillotte M, Wisskirchen C, Hellmuth J, Willms S, et al. Polymorphisms in melanoma differentiation-associated gene 5 link protein function to clearance of hepatitis C virus. Hepatology. 2015;61(2):460-70.
273. Grandvaux N, Guan X, Yoboua F, Zucchini N, Fink K, Doyon P, et al. Sustained activation of interferon regulatory factor 3 during infection by paramyxoviruses requires MDA5. J Innate Immun. 2014;6(5):650-62.
274. Runge S, Sparrer KM, Lassig C, Hembach K, Baum A, Garcia-Sastre A, et al. In vivo ligands of MDA5 and RIG-I in measles virus-infected cells. PLoS Pathog. 2014;10(4):e1004081.
275. Pang L, Gong X, Liu N, Xie G, Gao W, Kong G, et al. A polymorphism in melanoma differentiation-associated gene 5 may be a risk factor for enterovirus 71 infection. Clin Microbiol Infect. 2014;20(10):O711-7.
276. Feng Q, Langereis MA, Lork M, Nguyen M, Hato SV, Lanke K, et al. Enterovirus 2Apro targets MDA5 and MAVS in infected cells. J Virol. 2014;88(6):3369-78.
277. Nasirudeen AM, Wong HH, Thien P, Xu S, Lam KP, Liu DX. RIG-I, MDA5 and TLR3 synergistically play an important role in restriction of dengue virus infection. PLoS Negl Trop Dis. 2011;5(1):e926.
278. Broquet AH, Hirata Y, McAllister CS, Kagnoff MF. RIG-I/MDA5/MAVS are required to signal a protective IFN response in rotavirus-infected intestinal epithelium. J Immunol. 2011;186(3):1618-26.
279. Melchjorsen J, Rintahaka J, Soby S, Horan KA, Poltajainen A, Ostergaard L, et al. Early innate recognition of herpes simplex virus in human primary macrophages is mediated via the MDA5/MAVS-dependent and MDA5/MAVS/RNA polymerase III-independent pathways. J Virol. 2010;84(21):11350-8.
280. Ikegame S, Takeda M, Ohno S, Nakatsu Y, Nakanishi Y, Yanagi Y. Both RIG-I and MDA5 RNA helicases contribute to the induction of alpha/beta interferon in measles virus-infected human cells. J Virol. 2010;84(1):372-9.
281. Lifland AW, Jung J, Alonas E, Zurla C, Crowe JE, Jr., Santangelo PJ. Human respiratory syncytial virus nucleoprotein and inclusion bodies antagonize the innate immune response mediated by MDA5 and MAVS. J Virol. 2012;86(15):8245-58.
282. Siren J, Imaizumi T, Sarkar D, Pietila T, Noah DL, Lin R, et al. Retinoic acid inducible gene-I and mda-5 are involved in influenza A virus-induced expression of antiviral cytokines. Microbes Infect. 2006;8(8):2013-20.
283. Berghall H, Siren J, Sarkar D, Julkunen I, Fisher PB, Vainionpaa R, et al. The interferon-inducible RNA helicase, mda-5, is involved in measles virus-induced expression of antiviral cytokines. Microbes Infect. 2006;8(8):2138-44.
284. Zalinger ZB, Elliott R, Rose KM, Weiss SR. MDA5 Is Critical to Host Defense during Infection with Murine Coronavirus. J Virol. 2015;89(24):12330-40.
285. Lu HL, Liao F. Melanoma differentiation-associated gene 5 senses hepatitis B virus and activates innate immune signaling to suppress virus replication. J Immunol. 2013;191(6):3264-76.

286. McCartney EM, Beard MR. Impact of alcohol on hepatitis C virus replication and interferon signaling. World J Gastroenterol. 2010;16(11):1337-43.
287. Gitlin L, Benoit L, Song C, Cella M, Gilfillan S, Holtzman MJ, et al. Melanoma differentiation-associated gene 5 (MDA5) is involved in the innate immune response to Paramyxoviridae infection in vivo. PLoS Pathog. 2010;6(1):e1000734.
288. Kato H, Takeuchi O, Sato S, Yoneyama M, Yamamoto M, Matsui K, et al. Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. Nature. 2006;441(7089):101-5.
289. Goldman FD, Ballas ZK, Schutte BC, Kemp J, Hollenback C, Noraz N, et al. Defective expression of p56lck in an infant with severe combined immunodeficiency. J Clin Invest. 1998;102(2):421-9.
290. Molina TJ, Kishihara K, Siderovski DP, van Ewijk W, Narendran A, Timms E, et al. Profound block in thymocyte development in mice lacking p56lck. Nature. 1992;357(6374):161-4.
291. Levin SD, Anderson SJ, Forbush KA, Perlmutter RM. A dominant-negative transgene defines a role for p56lck in thymopoiesis. The EMBO journal. 1993;12(4):1671-80.
292. Welte T, Leitenberg D, Dittel BN, al-Ramadi BK, Xie B, Chin YE, et al. STAT5 interaction with the T cell receptor complex and stimulation of T cell proliferation. Science. 1999;283(5399):222-5.
293. Kim PW, Sun ZY, Blacklow SC, Wagner G, Eck MJ. A zinc clasp structure tethers Lck to T cell coreceptors CD4 and CD8. Science. 2003;301(5640):1725-8.
294. Sawabe T, Horiuchi T, Nakamura M, Tsukamoto H, Nakahara K, Harashima SI, et al. Defect of lck in a patient with common variable immunodeficiency. Int J Mol Med. 2001;7(6):609-14.
295. Hauck F, Randriamampita C, Martin E, Gerart S, Lambert N, Lim A, et al. Primary T-cell immunodeficiency with immunodysregulation caused by autosomal recessive LCK deficiency. J Allergy Clin Immunol. 2012;130(5):1144-52 e11.
296. Legname G, Seddon B, Lovatt M, Tomlinson P, Sarner N, Tolaini M, et al. Inducible expression of a p56Lck transgene reveals a central role for Lck in the differentiation of CD4 SP thymocytes. Immunity. 2000;12(5):537-46.
297. Chiang YJ, Hodes RJ. Regulation of T cell development by c-Cbl: essential role of Lck. Int Immunol. 2015;27(5):245-51.
298. Karlas A, Machuy N, Shin Y, Pleissner KP, Artarini A, Heuer D, et al. Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. Nature. 2010;463(7282):818-22.
299. Sun CT, Lo WY, Wang IH, Lo YH, Shiou SR, Lai CK, et al. Transcription repression of human hepatitis B virus genes by negative regulatory element-binding protein/SON. The Journal of biological chemistry. 2001;276(26):24059-67.
300. Komori T, Doi A, Furuta H, Wakao H, Nakao N, Nakazato M, et al. Regulation of ghrelin signaling by a leptin-induced gene, negative regulatory element-binding protein, in the hypothalamic neurons. The Journal of biological chemistry. 2010;285(48):37884-94.
301. Tschop M, Smiley DL, Heiman ML. Ghrelin induces adiposity in rodents. Nature. 2000;407(6806):908-13.

302. Nakazato M, Murakami N, Date Y, Kojima M, Matsuo H, Kangawa K, et al. A role for ghrelin in the central regulation of feeding. Nature. 2001;409(6817):194-8.
303. Theander-Carrillo C, Wiedmer P, Cettour-Rose P, Nogueiras R, Perez-Tilve D, Pfluger P, et al. Ghrelin action in the brain controls adipocyte metabolism. J Clin Invest. 2006;116(7):1983-93.
304. Wortley KE, Anderson KD, Garcia K, Murray JD, Malinova L, Liu R, et al. Genetic deletion of ghrelin does not decrease food intake but influences metabolic fuel preference. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(21):8227-32.
305. Chebani Y, Marion C, Zizzari P, Chettab K, Pastor M, Korostelev M, et al. Enhanced responsiveness of Ghsr Q343X rats to ghrelin results in enhanced adiposity without increased appetite. Sci Signal. 2016;9(424):ra39.
306. Wortley KE, del Rincon JP, Murray JD, Garcia K, Iida K, Thorner MO, et al. Absence of ghrelin protects against early-onset obesity. J Clin Invest. 2005;115(12):3573-8.
307. Zigman JM, Nakano Y, Coppari R, Balthasar N, Marcus JN, Lee CE, et al. Mice lacking ghrelin receptors resist the development of diet-induced obesity. J Clin Invest. 2005;115(12):3564-72.
308. Li RL, Sherbet DP, Elsbernd BL, Goldstein JL, Brown MS, Zhao TJ. Profound hypoglycemia in starved, ghrelin-deficient mice is caused by decreased gluconeogenesis and reversed by lactate or fatty acids. The Journal of biological chemistry. 2012;287(22):17942-50.
309. Szentirmai E, Kapas L, Sun Y, Smith RG, Krueger JM. The preproghrelin gene is required for the normal integration of thermoregulation and sleep in mice. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(33):14069-74.
310. Farquhar J, Heiman M, Wong AC, Wach R, Chessex P, Chanoine JP. Elevated umbilical cord ghrelin concentrations in small for gestational age neonates. J Clin Endocrinol Metab. 2003;88(9):4324-7.
311. Zhang JV, Ren PG, Avsian-Kretchmer O, Luo CW, Rauch R, Klein C, et al. Obestatin, a peptide encoded by the ghrelin gene, opposes ghrelin's effects on food intake. Science. 2005;310(5750):996-9.
312. Nouh O, Abd Elfattah MM, Hassouna AA. Association between ghrelin levels and BMD: a cross sectional trial. Gynecol Endocrinol. 2012;28(7):570-2.
313. Shojima N, Hara K, Fujita H, Horikoshi M, Takahashi N, Takamoto I, et al. Depletion of homeodomain-interacting protein kinase 3 impairs insulin secretion and glucose tolerance in mice. Diabetologia. 2012;55(12):3318-30.
314. Hahm S, Mizuno TM, Wu TJ, Wisor JP, Priest CA, Kozak CA, et al. Targeted deletion of the Vgf gene indicates that the encoded secretory peptide precursor plays a novel role in the regulation of energy balance. Neuron. 1999;23(3):537-48.
315. Watson E, Fargali S, Okamoto H, Sadahiro M, Gordon RE, Chakraborty T, et al. Analysis of knockout mice suggests a role for VGF in the control of fat storage and energy expenditure. BMC Physiol. 2009;9:19.
316. D'Amato F, Noli B, Angioni L, Cossu E, Incani M, Messana I, et al. VGF Peptide Profiles in Type 2 Diabetic Patients' Plasma and in Obese Mice. PloS one. 2015;10(11):e0142333.

317. Rahimi M, Vinciguerra M, Daghighi M, Ozcan B, Akbarkhanzadeh V, Sheedfar F, et al. Age-related obesity and type 2 diabetes dysregulate neuronal associated genes and proteins in humans. Oncotarget. 2015;6(30):29818-32.
318. Kim JW, Rhee M, Park JH, Yamaguchi H, Sasaki K, Minamino N, et al. Chronic effects of neuroendocrine regulatory peptide (NERP-1 and -2) on insulin secretion and gene expression in pancreatic beta-cells. Biochemical and biophysical research communications. 2015;457(2):148-53.
319. Bartolomucci A, La Corte G, Possenti R, Locatelli V, Rigamonti AE, Torsello A, et al. TLQP-21, a VGF-derived peptide, increases energy expenditure and prevents the early phase of diet-induced obesity. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(39):14584-9.
320. Watson E, Hahm S, Mizuno TM, Windsor J, Montgomery C, Scherer PE, et al. VGF ablation blocks the development of hyperinsulinemia and hyperglycemia in several mouse models of obesity. Endocrinology. 2005;146(12):5151-63.
321. Henn BM, Botigue LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS genetics. 2012;8(1):e1002397.
322. Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, et al. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. American journal of human genetics. 2012;91(1):83-96.
323. Fry AE, Ghansa A, Small KS, Palma A, Auburn S, Diakite M, et al. Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. Human molecular genetics. 2009;18(14):2683-92.
324. Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, Orago AS, et al. Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. American journal of human genetics. 2007;81(2):234-42.
325. Love-Gregory L, Sherva R, Sun L, Wasson J, Schappe T, Doria A, et al. Variants in the CD36 gene associate with the metabolic syndrome and high-density lipoprotein cholesterol. Human molecular genetics. 2008;17(11):1695-704.
326. Matsuo Y, Yokoyama R, Yokoyama S. The genes for human alcohol dehydrogenases beta 1 and beta 2 differ by only one nucleotide. Eur J Biochem. 1989;183(2):317-20.
327. Jornvall H, Hempel J, Vallee BL, Bosron WF, Li TK. Human liver alcohol dehydrogenase: amino acid substitution in the beta 2 beta 2 Oriental isozyme explains functional properties, establishes an active site structure, and parallels mutational exchanges in the yeast enzyme. Proceedings of the National Academy of Sciences of the United States of America. 1984;81(10):3024-8.
328. Hurley TD, Edenberg HJ, Bosron WF. Expression and kinetic characterization of variants of human beta 1 beta 1 alcohol dehydrogenase containing substitutions at amino acid 47. The Journal of biological chemistry. 1990;265(27):16366-72.

329. Han Y, Gu S, Oota H, Osier MV, Pakstis AJ, Speed WC, et al. Evidence of positive selection on a class I ADH locus. American journal of human genetics. 2007;80(3):441-56.
330. Mendez R, Hake LE, Andresson T, Littlepage LE, Ruderman JV, Richter JD. Phosphorylation of CPE binding factor by Eg2 regulates translation of c-mos mRNA. Nature. 2000;404(6775):302-7.
331. Prasad CK, Mahadevan M, MacNicol MC, MacNicol AM. Mos 3' UTR regulatory differences underlie species-specific temporal patterns of Mos mRNA cytoplasmic polyadenylation and translational recruitment during oocyte maturation. Mol Reprod Dev. 2008;75(8):1258-68.
332. Hashimoto N, Watanabe N, Furuta Y, Tamemoto H, Sagata N, Yokoyama M, et al. Parthenogenetic activation of oocytes in c-mos-deficient mice. Nature. 1994;370(6484):68-71.
333. Singh B, Arlinghaus RB. Mos and the cell cycle. Prog Cell Cycle Res. 1997;3:251-9.
334. Haber M, Gauguier D, Youhanna S, Patterson N, Moorjani P, Botigue LR, et al. Genome-wide diversity in the levant reveals recent structuring by culture. PLoS genetics. 2013;9(2):e1003316.
335. Huang AL, Chen X, Hoon MA, Chandrashekar J, Guo W, Trankner D, et al. The cells and logic for mammalian sour taste detection. Nature. 2006;442(7105):934-8.
336. Ishimaru Y, Inada H, Kubota M, Zhuang H, Tominaga M, Matsunami H. Transient receptor potential family members PKD1L3 and PKD2L1 form a candidate sour taste receptor. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(33):12569-74.
337. Jalalvand E, Robertson B, Tostivint H, Wallen P, Grillner S. The Spinal Cord Has an Intrinsic System for the Control of pH. Current biology : CB. 2016;26(10):1346-51.
338. Horio N, Yoshida R, Yasumatsu K, Yanagawa Y, Ishimaru Y, Matsunami H, et al. Sour taste responses in mice lacking PKD channels. PloS one. 2011;6(5):e20007.
339. Ishimaru Y. Molecular mechanisms underlying the reception and transmission of sour taste information. Biosci Biotechnol Biochem. 2015;79(2):171-6.
340. Yu B, Zheng Y, Alexander D, Morrison AC, Coresh J, Boerwinkle E. Genetic determinants influencing human serum metabolome among African Americans. PLoS genetics. 2014;10(3):e1004212.
341. Wu JH, Lemaitre RN, Manichaikul A, Guan W, Tanaka T, Foy M, et al. Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. Circ Cardiovasc Genet. 2013;6(2):171-83.
342. Demirkan A, van Duijn CM, Ugocsai P, Isaacs A, Pramstaller PP, Liebisch G, et al. Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. PLoS genetics. 2012;8(2):e1002490.
343. Tremblay F, Revett T, Huard C, Zhang Y, Tobin JF, Martinez RV, et al. Bidirectional modulation of adipogenesis by the secreted protein

Ccdc80/DRO1/URB. The Journal of biological chemistry. 2009;284(12):8136-47.
344. Aoki K, Sun YJ, Aoki S, Wada K, Wada E. Cloning, expression, and mapping of a gene that is upregulated in adipose tissue of mice deficient in bombesin receptor subtype-3. Biochemical and biophysical research communications. 2002;290(4):1282-8.
345. Okada T, Nishizawa H, Kurata A, Tamba S, Sonoda M, Yasui A, et al. URB is abundantly expressed in adipose tissue and dysregulated in obesity. Biochemical and biophysical research communications. 2008;367(2):370-6.
346. Tremblay F, Huard C, Dow J, Gareski T, Will S, Richard AM, et al. Loss of coiled-coil domain containing 80 negatively modulates glucose homeostasis in diet-induced obese mice. Endocrinology. 2012;153(9):4290-303.
347. Li L, Zhang Y, Du H, He P, Li G, Liu X, et al. [Correlation between the expression level of coiled-coil domain-containing protein 80 and obesity]. Zhonghua Yu Fang Yi Xue Za Zhi. 2015;49(3):248-53.
348. Kowalski MP, Dubouix-Bourandy A, Bajmoczi M, Golan DE, Zaidi T, Coutinho-Sledge YS, et al. Host resistance to lung infection mediated by major vault protein in epithelial cells. Science. 2007;317(5834):130-2.
349. Suprenant KA, Bloom N, Fang J, Lushington G. The major vault protein is related to the toxic anion resistance protein (TelA) family. J Exp Biol. 2007;210(Pt 6):946-55.
350. Dortet L, Mostowy S, Samba-Louaka A, Gouin E, Nahori MA, Wiemer EA, et al. Recruitment of the major vault protein by InlK: a Listeria monocytogenes strategy to avoid autophagy. PLoS Pathog. 2011;7(8):e1002168.
351. Lara PC, Pruschy M, Zimmermann M, Henriquez-Hernandez LA. MVP and vaults: a role in the radiation response. Radiat Oncol. 2011;6:148.
352. Lara PC, Lloret M, Clavo B, Apolinario RM, Henriquez-Hernandez LA, Bordon E, et al. Severe hypoxia induces chemo-resistance in clinical cervical tumors through MVP over-expression. Radiat Oncol. 2009;4:29.
353. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461(7263):489-94.
354. Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, et al. Genetic evidence on the origins of Indian caste populations. Genome research. 2001;11(6):994-1004.
355. Alchon SA. A pest in the land : new world epidemics in a global perspective. 1st ed. Albuquerque: University of New Mexico Press; 2003. ix, 214 p. p.
356. Cook ND. Demographic collapse, Indian Peru, 1520-1620. Cambridge Cambridgeshire ; New York: Cambridge University Press; 1981. x, 310 p. p.
357. Aberth J. The first horseman : disease in human history. Upper Saddle River, N.J.: Pearson Prentice Hall; 2007. xii, 177 p. p.
358. Francis JM. Iberia and the Americas : culture, politics, and history : a multidisciplinary encyclopedia. Santa Barbara, Calif.: ABC-CLIO; 2006.
359. Cook ND. Born to die : disease and New World conquest, 1492-1650. Cambridge ; New York: Cambridge University Press; 1998. xiii, 248 p. p.
360. Fumagalli M, Cagliani R, Riva S, Pozzoli U, Biasin M, Piacentini L, et al. Population genetics of IFIH1: ancient population structure, local selection, and implications for susceptibility to type 1 diabetes. Molecular biology and evolution. 2010;27(11):2555-66.

361. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009;324(5925):387-9.
362. Liu S, Wang H, Jin Y, Podolsky R, Reddy MV, Pedersen J, et al. IFIH1 polymorphisms are significantly associated with type 1 diabetes and IFIH1 gene expression in peripheral blood mononuclear cells. Human molecular genetics. 2009;18(2):358-65.
363. Schell LM, Gallo MV. Overweight and obesity among North American Indian infants, children, and youth. American journal of human biology : the official journal of the Human Biology Council. 2012;24(3):302-13.
364. Story M, Evans M, Fabsitz RR, Clay TE, Holy Rock B, Broussard B. The epidemic of obesity in American Indian communities and the need for childhood obesity-prevention programs. Am J Clin Nutr. 1999;69(4 Suppl):747S-54S.
365. Ravussin E. Energy metabolism in obesity. Studies in the Pima Indians. Diabetes Care. 1993;16(1):232-8.
366. Wang SP, Yang H, Wu JW, Gauthier N, Fukao T, Mitchell GA. Metabolism as a tool for understanding human brain evolution: lipid energy metabolism as an example. J Hum Evol. 2014;77:41-9.
367. Gessner BD, Gillingham MB, Birch S, Wood T, Koeller DM. Evidence for an association between infant mortality and a carnitine palmitoyltransferase 1A genetic variant. Pediatrics. 2010;126(5):945-51.
368. Boyce VL, Swinburn BA. The traditional Pima Indian diet. Composition and adaptation for use in a dietary intervention study. Diabetes Care. 1993;16(1):369-71.
369. Smith CJ, Nelson RG, Hardy SA, Manahan EM, Bennett PH, Knowler WC. Survey of the diet of Pima Indians using quantitative food frequency assessment and 24-hour recall. Diabetic Renal Disease Study. J Am Diet Assoc. 1996;96(8):778-84.
370. Tekola-Ayele F, Adeyemo A, Chen G, Hailu E, Aseffa A, Davey G, et al. Novel genomic signals of recent selection in an Ethiopian population. European journal of human genetics : EJHG. 2015;23(8):1085-92.
371. Carrillo-Larco RM, Bernabe-Ortiz A, Pillay TD, Gilman RH, Sanchez JF, Poterico JA, et al. Obesity risk in rural, urban and rural-to-urban migrants: prospective results of the PERU MIGRANT study. Int J Obes (Lond). 2016;40(1):181-5.
372. Antiporta DA, Smeeth L, Gilman RH, Miranda JJ. Length of urban residence and obesity among within-country rural-to-urban Andean migrants. Public Health Nutr. 2016;19(7):1270-8.
373. Lindgarde F, Ercilla MB, Correa LR, Ahren B. Body adiposity, insulin, and leptin in subgroups of Peruvian Amerindians. High Alt Med Biol. 2004;5(1):27-31.
374. Revilla L, Lopez T, Sanchez S, Yasuda M, Sanjines G. [Prevalence of hypertension and diabetes in residents from Lima and Callao, Peru]. Rev Peru Med Exp Salud Publica. 2014;31(3):437-44.
375. Nunez-Robles E, Huapaya-Pizarro C, Torres-Lao R, Esquivel-Leon S, Suarez-Moreno V, Yasuda-Espinoza M, et al. [Prevalence of cardiovascular and metabolic risk factors in school students, university students, and women from community-based organizations in the districts of Lima, Callao, la

Libertad and Arequipa, Peru 2011]. Rev Peru Med Exp Salud Publica. 2014;31(4):652-9.
376. Lozano-Rojas G, Cabello-Morales E, Hernadez-Diaz H, Loza-Munarriz C. [Prevalence of overweight and obesity in adolescents from an urban district of Lima, Peru 2012]. Rev Peru Med Exp Salud Publica. 2014;31(3):494-500.
377. Bustamante A, Maia J. [Weight status and cardiorespiratory fitness in school students in the central region of Peru]. Rev Peru Med Exp Salud Publica. 2013;30(3):399-407.
378. Nam EW, Sharma B, Kim HY, Paja DJ, Yoon YM, Lee SH, et al. Obesity and Hypertension among School-going Adolescents in Peru. J Lifestyle Med. 2015;5(2):60-7.
379. Mohanna S, Baracco R, Seclen S. Lipid profile, waist circumference, and body mass index in a high altitude population. High Alt Med Biol. 2006;7(3):245-55.
380. Medina-Lezama J, Zea-Diaz H, Morey-Vargas OL, Bolanos-Salazar JF, Munoz-Atahualpa E, Postigo-MacDowall M, et al. Prevalence of the metabolic syndrome in Peruvian Andean hispanics: the PREVENCION study. Diabetes Res Clin Pract. 2007;78(2):270-81.
381. Baracco R, Mohanna S, Seclen S. A comparison of the prevalence of metabolic syndrome and its components in high and low altitude populations in peru. Metab Syndr Relat Disord. 2007;5(1):55-62.
382. Seclen SN, Rosas ME, Arias AJ, Huayta E, Medina CA. Prevalence of diabetes and impaired fasting glucose in Peru: report from PERUDIAB, a national urban population-based longitudinal study. BMJ Open Diabetes Res Care. 2015;3(1):e000110.
383. Abrha S, Shiferaw S, Ahmed KY. Overweight and obesity and its socio-demographic correlates among urban Ethiopian women: evidence from the 2011 EDHS. BMC Public Health. 2016;16:636.
384. Helelo TP, Gelaw YA, Adane AA. Prevalence and associated factors of hypertension among adults in Durame Town, Southern Ethiopia. PloS one. 2014;9(11):e112790.
385. Moges B, Amare B, Fantahun B, Kassu A. High prevalence of overweight, obesity, and hypertension with increased risk to cardiovascular disorders among adults in northwest Ethiopia: a cross sectional study. BMC Cardiovasc Disord. 2014;14:155.
386. Tebekaw Y, Teller C, Colon-Ramos U. The burden of underweight and overweight among women in Addis Ababa, Ethiopia. BMC Public Health. 2014;14:1126.
387. Tadesse T, Alemu H. Hypertension and associated factors among university students in Gondar, Ethiopia: a cross-sectional study. BMC Public Health. 2014;14:937.
388. Awoke A, Awoke T, Alemu S, Megabiaw B. Prevalence and associated factors of hypertension among adults in Gondar, Northwest Ethiopia: a community based cross-sectional study. BMC Cardiovasc Disord. 2012;12:113.
389. Regev-Tobias H, Reifen R, Endevelt R, Havkin O, Cohen E, Stern G, et al. Dietary acculturation and increasing rates of obesity in Ethiopian women living in Israel. Nutrition. 2012;28(1):30-4.

390. Tesfaye F, Byass P, Wall S. Population based prevalence of high blood pressure among adults in Addis Ababa: uncovering a silent epidemic. BMC Cardiovasc Disord. 2009;9:39.
391. Holden JE, Stone CK, Clark CM, Brown WD, Nickles RJ, Stanley C, et al. Enhanced cardiac metabolism of plasma glucose in high-altitude natives: adaptation against chronic hypoxia. J Appl Physiol (1985). 1995;79(1):222-8.
392. Hochachka PW, Clark CM, Brown WD, Stanley C, Stone CK, Nickles RJ, et al. The brain at high altitude: hypometabolism as a defense against chronic hypoxia? J Cereb Blood Flow Metab. 1994;14(4):671-9.
393. Wang P, Ha AY, Kidd KK, Koehle MS, Rupert JL. A variant of the endothelial nitric oxide synthase gene (NOS3) associated with AMS susceptibility is less common in the Quechua, a high altitude Native population. High Alt Med Biol. 2010;11(1):27-30.
394. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant pleiotropy in human complex diseases and traits. American journal of human genetics. 2011;89(5):607-18.
395. Malnic B, Godfrey PA, Buck LB. The human olfactory receptor gene family. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(8):2584-9.
396. Buettner JA, Glusman G, Ben-Arie N, Ramos P, Lancet D, Evans GA. Organization and evolution of olfactory receptor genes on human chromosome 11. Genomics. 1998;53(1):56-68.
397. Gilad Y, Segre D, Skorecki K, Nachman MW, Lancet D, Sharon D. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. Nature genetics. 2000;26(2):221-4.
398. Gaillard I, Rouquier S, Giorgi D. Olfactory receptors. Cell Mol Life Sci. 2004;61(4):456-69.
399. Zozulya S, Echeverri F, Nguyen T. The human olfactory receptor repertoire. Genome biology. 2001;2(6):RESEARCH0018.
400. Younger RM, Amadou C, Bethel G, Ehlers A, Lindahl KF, Forbes S, et al. Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. Genome research. 2001;11(4):519-30.
401. Mainland JD, Keller A, Li YR, Zhou T, Trimmer C, Snyder LL, et al. The missense of smell: functional variability in the human odorant receptor repertoire. Nat Neurosci. 2014;17(1):114-20.
402. Keller A, Zhuang H, Chi Q, Vosshall LB, Matsunami H. Genetic variation in a human odorant receptor alters odour perception. Nature. 2007;449(7161):468-72.
403. Jaeger SR, McRae JF, Bava CM, Beresford MK, Hunter D, Jia Y, et al. A Mendelian trait for olfactory sensitivity affects odor experience and food selection. Current biology : CB. 2013;23(16):1601-5.
404. McRae JF, Mainland JD, Jaeger SR, Adipietro KA, Matsunami H, Newcomb RD. Genetic variation in the odorant receptor OR2J3 is associated with the ability to detect the "grassy" smelling odor, cis-3-hexen-1-ol. Chem Senses. 2012;37(7):585-93.
405. Katada S, Hirokawa T, Oka Y, Suwa M, Touhara K. Structural basis for a broad but selective ligand spectrum of a mouse olfactory receptor: mapping the

odorant-binding site. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2005;25(7):1806-15.
406. Hasin-Brumshtein Y, Lancet D, Olender T. Human olfaction: from genomic variation to phenotypic diversity. Trends in genetics : TIG. 2009;25(4):178-84.
407. Adipietro KA, Mainland JD, Matsunami H. Functional evolution of mammalian odorant receptors. PLoS genetics. 2012;8(7):e1002821.
408. Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome research. 2010;20(1):1-9.
409. Keller A, Vosshall LB. Better smelling through genetics: mammalian odor perception. Curr Opin Neurobiol. 2008;18(4):364-9.
410. Capittini C, Martinetti M, Cuccia M. MHC variation, mate choice and natural selection: the scent of evolution. Riv Biol. 2008;101(3):463-80.
411. Apfelbach R, Blanchard CD, Blanchard RJ, Hayes RA, McGregor IS. The effects of predator odors in mammalian prey species: a review of field and laboratory studies. Neurosci Biobehav Rev. 2005;29(8):1123-44.
412. Gilad Y, Man O, Paabo S, Lancet D. Human specific loss of olfactory receptor genes. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(6):3324-7.
413. Niimura Y, Nei M. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PloS one. 2007;2(8):e708.
414. Gimelbrant AA, Skaletsky H, Chess A. Selective pressures on the olfactory receptor repertoire since the human-chimpanzee divergence. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(24):9019-22.
415. Glusman G, Yanai I, Rubin I, Lancet D. The complete human olfactory subgenome. Genome research. 2001;11(5):685-702.
416. Rouquier S, Blancher A, Giorgi D. The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. Proceedings of the National Academy of Sciences of the United States of America. 2000;97(6):2870-4.
417. Sharon D, Glusman G, Pilpel Y, Khen M, Gruetzner F, Haaf T, et al. Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. Genomics. 1999;61(1):24-36.
418. Go Y, Niimura Y. Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. Molecular biology and evolution. 2008;25(9):1897-907.
419. Gilad Y, Lancet D. Population differences in the human functional olfactory repertoire. Molecular biology and evolution. 2003;20(3):307-14.
420. Pierron D, Cortes NG, Letellier T, Grossman LI. Current relaxation of selection on the human genome: tolerance of deleterious mutations on olfactory receptors. Mol Phylogenet Evol. 2013;66(2):558-64.
421. Gilad Y, Bustamante CD, Lancet D, Paabo S. Natural selection on the olfactory receptor gene family in humans and chimpanzees. American journal of human genetics. 2003;73(3):489-501.
422. Moreno-Estrada A, Casals F, Ramirez-Soriano A, Oliva B, Calafell F, Bertranpetit J, et al. Signatures of selection in the human olfactory receptor OR5I1 gene. Molecular biology and evolution. 2008;25(1):144-54.

423. Saito H, Chi Q, Zhuang H, Matsunami H, Mainland JD. Odor coding by a Mammalian receptor repertoire. Sci Signal. 2009;2(60):ra9.
424. Trimmer C, Snyder LL, Mainland JD. High-throughput analysis of mammalian olfactory receptors: measurement of receptor activation via luciferase activity. J Vis Exp. 2014(88).
425. Zhuang H, Matsunami H. Evaluating cell-surface expression and measuring activation of mammalian odorant receptors in heterologous cells. Nat Protoc. 2008;3(9):1402-13.
426. Saito H, Kubota M, Roberts RW, Chi Q, Matsunami H. RTP family members induce functional expression of mammalian odorant receptors. Cell. 2004;119(5):679-91.
427. Zhuang H, Matsunami H. Synergism of accessory factors in functional expression of mammalian odorant receptors. The Journal of biological chemistry. 2007;282(20):15284-93.
428. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. Nucleic acids research. 2016;44(D1):D746-52.
429. Flegel C, Manteniotis S, Osthold S, Hatt H, Gisselmann G. Expression profile of ectopic olfactory receptors determined by deep sequencing. PloS one. 2013;8(2):e55368.
430. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Molecular biology and evolution. 1999;16(1):37-48.
431. Feldmesser E, Olender T, Khen M, Yanai I, Ophir R, Lancet D. Widespread ectopic expression of olfactory receptor genes. BMC genomics. 2006;7:121.
432. De la Cruz O, Blekhman R, Zhang X, Nicolae D, Firestein S, Gilad Y. A signature of evolutionary constraint on a subset of ectopically expressed olfactory receptor genes. Molecular biology and evolution. 2009;26(3):491-4.
433. Braun T, Voland P, Kunz L, Prinz C, Gratzl M. Enterochromaffin cells of the human gut: sensors for spices and odorants. Gastroenterology. 2007;132(5):1890-901.
434. Vanderhaeghen P, Schurmans S, Vassart G, Parmentier M. Molecular cloning and chromosomal mapping of olfactory receptor genes expressed in the male germ line: evidence for their wide distribution in the human genome. Biochemical and biophysical research communications. 1997;237(2):283-7.
435. Vanderhaeghen P, Schurmans S, Vassart G, Parmentier M. Specific repertoire of olfactory receptor genes in the male germ cells of several mammalian species. Genomics. 1997;39(3):239-46.
436. Spehr M, Gisselmann G, Poplawski A, Riffell JA, Wetzel CH, Zimmer RK, et al. Identification of a testicular odorant receptor mediating human sperm chemotaxis. Science. 2003;299(5615):2054-8.
437. Veitinger T, Riffell JR, Veitinger S, Nascimento JM, Triller A, Chandsawangbhuwana C, et al. Chemosensory Ca2+ dynamics correlate with diverse behavioral phenotypes in human sperm. The Journal of biological chemistry. 2011;286(19):17311-25.
438. Spehr M, Schwane K, Heilmann S, Gisselmann G, Hummel T, Hatt H. Dual capacity of a human olfactory receptor. Current biology : CB. 2004;14(19):R832-3.

439. Gakamsky A, Armon L, Eisenbach M. Behavioral response of human spermatozoa to a concentration jump of chemoattractants or intracellular cyclic nucleotides. Hum Reprod. 2009;24(5):1152-63.
440. Ziegler A, Dohr G, Uchanska-Ziegler B. Possible roles for products of polymorphic MHC and linked olfactory receptor genes during selection processes in reproduction. Am J Reprod Immunol. 2002;48(1):34-42.
441. Dreyer WJ. The area code hypothesis revisited: olfactory receptors and other related transmembrane receptors may function as the last digits in a cell surface code for assembling embryos. Proceedings of the National Academy of Sciences of the United States of America. 1998;95(16):9072-7.
442. Parmentier M, Libert F, Schurmans S, Schiffmann S, Lefort A, Eggerickx D, et al. Expression of members of the putative olfactory receptor gene family in mammalian germ cells. Nature. 1992;355(6359):453-5.
443. Solovieff N, Milton JN, Hartley SW, Sherva R, Sebastiani P, Dworkis DA, et al. Fetal hemoglobin in sickle cell anemia: genome-wide association studies suggest a regulatory region in the 5' olfactory receptor gene cluster. Blood. 2010;115(9):1815-22.
444. Dean A. Chromatin remodelling and the interaction between enhancers and promoters in the beta-globin locus. Brief Funct Genomic Proteomic. 2004;2(4):344-54.
445. Bulger M, Bender MA, van Doorninck JH, Wertman B, Farrell CM, Felsenfeld G, et al. Comparative structural and functional analysis of the olfactory receptor genes flanking the human and mouse beta-globin gene clusters. Proceedings of the National Academy of Sciences of the United States of America. 2000;97(26):14560-5.
446. Feingold EA, Penny LA, Nienhuis AW, Forget BG. An olfactory receptor gene is located in the extended human beta-globin gene cluster and is expressed in erythroid cells. Genomics. 1999;61(1):15-23.
447. Kim A, Kiefer CM, Dean A. Distinctive signatures of histone methylation in transcribed coding and noncoding human beta-globin sequences. Molecular and cellular biology. 2007;27(4):1271-9.
448. Epner E, Reik A, Cimbora D, Telling A, Bender MA, Fiering S, et al. The beta-globin LCR is not necessary for an open chromatin structure or developmentally regulated transcription of the native mouse beta-globin locus. Mol Cell. 1998;2(4):447-55.
449. Uda M, Galanello R, Sanna S, Lettre G, Sankaran VG, Chen W, et al. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(5):1620-5.
450. Sherva R, Sripichai O, Abel K, Ma Q, Whitacre J, Angkachatchai V, et al. Genetic modifiers of Hb E/beta0 thalassemia identified by a two-stage genome-wide association study. BMC Med Genet. 2010;11:51.
451. Bulger M, Schubeler D, Bender MA, Hamilton J, Farrell CM, Hardison RC, et al. A complex chromatin landscape revealed by patterns of nuclease sensitivity and histone modification within the mouse beta-globin locus. Molecular and cellular biology. 2003;23(15):5234-44.

452. Ferrer-Admetlla A, Sikora M, Laayouni H, Esteve A, Roubinet F, Blancher A, et al. A natural history of FUT2 polymorphism in humans. Molecular biology and evolution. 2009;26(9):1993-2003.
453. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nature reviews Genetics. 2014;15(6):379-93.
454. Imbert-Marcille BM, Barbe L, Dupe M, Le Moullac-Vaidye B, Besse B, Peltier C, et al. A FUT2 gene common polymorphism determines resistance to rotavirus A of the P[8] genotype. J Infect Dis. 2014;209(8):1227-30.
455. Tate JE, Burton AH, Boschi-Pinto C, Steele AD, Duque J, Parashar UD, et al. 2008 estimate of worldwide rotavirus-associated mortality in children younger than 5 years before the introduction of universal rotavirus vaccination programmes: a systematic review and meta-analysis. Lancet Infect Dis. 2012;12(2):136-41.
456. Debbink K, Lindesmith LC, Donaldson EF, Baric RS. Norovirus immunity and the great escape. PLoS Pathog. 2012;8(10):e1002921.
457. Huang P, Xia M, Tan M, Zhong W, Wei C, Wang L, et al. Spike protein VP8* of human rotavirus recognizes histo-blood group antigens in a type-specific manner. J Virol. 2012;86(9):4833-43.
458. Hu L, Crawford SE, Czako R, Cortes-Penfield NW, Smith DF, Le Pendu J, et al. Cell attachment protein VP8* of a human rotavirus specifically interacts with A-type histo-blood group antigen. Nature. 2012;485(7397):256-9.
459. Liu Y, Huang P, Tan M, Liu Y, Biesiada J, Meller J, et al. Rotavirus VP8*: phylogeny, host range, and interaction with histo-blood group antigens. J Virol. 2012;86(18):9899-910.
460. Ramani S, Cortes-Penfield NW, Hu L, Crawford SE, Czako R, Smith DF, et al. The VP8* domain of neonatal rotavirus strain G10P[11] binds to type II precursor glycans. J Virol. 2013;87(13):7255-64.
461. Kambhampati A, Payne DC, Costantini V, Lopman BA. Host Genetic Susceptibility to Enteric Viruses: A Systematic Review and Metaanalysis. Clin Infect Dis. 2016;62(1):11-8.
462. Marionneau S, Cailleau-Thomas A, Rocher J, Le Moullac-Vaidye B, Ruvoen N, Clement M, et al. ABH and Lewis histo-blood group antigens, a model for the meaning of oligosaccharide diversity in the face of a changing world. Biochimie. 2001;83(7):565-73.
463. Shirato H. Norovirus and histo-blood group antigens. Jpn J Infect Dis. 2011;64(2):95-103.
464. Koda Y, Tachida H, Pang H, Liu Y, Soejima M, Ghaderi AA, et al. Contrasting patterns of polymorphisms at the ABO-secretor gene (FUT2) and plasma alpha(1,3)fucosyltransferase gene (FUT6) in human populations. Genetics. 2001;158(2):747-56.
465. Kelly RJ, Rouquier S, Giorgi D, Lennon GG, Lowe JB. Sequence and expression of a candidate for the human Secretor blood group alpha(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. The Journal of biological chemistry. 1995;270(9):4640-9.
466. Yu LC, Yang YH, Broadberry RE, Chen YH, Chan YS, Lin M. Correlation of a missense mutation in the human Secretor alpha 1,2-fucosyltransferase gene with the Lewis(a+b+) phenotype: a potential molecular basis for the weak Secretor allele (Sew). Biochem J. 1995;312 (Pt 2):329-32.

467. Henry S, Mollicone R, Fernandez P, Samuelsson B, Oriol R, Larson G. Molecular basis for erythrocyte Le(a+ b+) and salivary ABH partial-secretor phenotypes: expression of a FUT2 secretor allele with an A-->T mutation at nucleotide 385 correlates with reduced alpha(1,2) fucosyltransferase activity. Glycoconj J. 1996;13(6):985-93.
468. Thorven M, Grahn A, Hedlund KO, Johansson H, Wahlfrid C, Larson G, et al. A homozygous nonsense mutation (428G-->A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. J Virol. 2005;79(24):15351-5.
469. Larsson MM, Rydell GE, Grahn A, Rodriguez-Diaz J, Akerlind B, Hutson AM, et al. Antibody prevalence and titer to norovirus (genogroup II) correlate with secretor (FUT2) but not with ABO phenotype or Lewis (FUT3) genotype. J Infect Dis. 2006;194(10):1422-7.
470. Marionneau S, Airaud F, Bovin NV, Le Pendu J, Ruvoen-Clouet N. Influence of the combined ABO, FUT2, and FUT3 polymorphism on susceptibility to Norwalk virus attachment. J Infect Dis. 2005;192(6):1071-7.
471. Liu P, Wang X, Lee JC, Teunis P, Hu S, Paradise HT, et al. Genetic susceptibility to norovirus GII.3 and GII.4 infections in Chinese pediatric diarrheal disease. Pediatr Infect Dis J. 2014;33(11):e305-9.
472. Carlsson B, Kindberg E, Buesa J, Rydell GE, Lidon MF, Montava R, et al. The G428A nonsense mutation in FUT2 provides strong but not absolute protection against symptomatic GII.4 Norovirus infection. PloS one. 2009;4(5):e5593.
473. Magalhaes A, Rossez Y, Robbe-Masselot C, Maes E, Gomes J, Shevtsova A, et al. Muc5ac gastric mucin glycosylation is shaped by FUT2 activity and functionally impacts Helicobacter pylori binding. Sci Rep. 2016;6:25575.
474. Ilver D, Arnqvist A, Ogren J, Frick IM, Kersulyte D, Incecik ET, et al. Helicobacter pylori adhesin binding fucosylated histo-blood group antigens revealed by retagging. Science. 1998;279(5349):373-7.
475. Boren T, Falk P, Roth KA, Larson G, Normark S. Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. Science. 1993;262(5141):1892-5.
476. Azevedo M, Eriksson S, Mendes N, Serpa J, Figueiredo C, Resende LP, et al. Infection by Helicobacter pylori expressing the BabA adhesin is influenced by the secretor phenotype. J Pathol. 2008;215(3):308-16.
477. Moore ME, Boren T, Solnick JV. Life at the margins: modulation of attachment proteins in Helicobacter pylori. Gut Microbes. 2011;2(1):42-6.
478. Polk DB, Peek RM, Jr. Helicobacter pylori: gastric cancer and beyond. Nat Rev Cancer. 2010;10(6):403-14.
479. Schreiber S, Konradt M, Groll C, Scheid P, Hanauer G, Werling HO, et al. The spatial orientation of Helicobacter pylori in the gastric mucus. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(14):5024-9.
480. Ali S, Niang MA, N'Doye I, Critchlow CW, Hawes SE, Hill AV, et al. Secretor polymorphism and human immunodeficiency virus infection in Senegalese women. J Infect Dis. 2000;181(2):737-9.
481. Blackwell CC, James VS, Davidson S, Wyld R, Brettle RP, Robertson RJ, et al. Secretor status and heterosexual transmission of HIV. BMJ. 1991;303(6806):825-6.

482. Kindberg E, Hejdeman B, Bratt G, Wahren B, Lindblom B, Hinkula J, et al. A nonsense mutation (428G-->A) in the fucosyltransferase FUT2 gene affects the progression of HIV-1 infection. AIDS. 2006;20(5):685-9.
483. Raza MW, Blackwell CC, Molyneaux P, James VS, Ogilvie MM, Inglis JM, et al. Association between secretor status and respiratory viral illness. BMJ. 1991;303(6806):815-8.
484. Wacklin P, Tuimala J, Nikkila J, Sebastian T, Makivuokko H, Alakulppi N, et al. Faecal microbiota composition in adults is associated with the FUT2 gene determining the secretor status. PloS one. 2014;9(4):e94863.
485. Tong M, McHardy I, Ruegger P, Goudarzi M, Kashyap PC, Haritunians T, et al. Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. ISME J. 2014;8(11):2193-206.
486. Rausch P, Rehman A, Kunzel S, Hasler R, Ott SJ, Schreiber S, et al. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(47):19030-5.
487. Wacklin P, Makivuokko H, Alakulppi N, Nikkila J, Tenkanen H, Rabina J, et al. Secretor genotype (FUT2 gene) is strongly associated with the composition of Bifidobacteria in the human intestine. PloS one. 2011;6(5):e20113.
488. Pham TA, Clare S, Goulding D, Arasteh JM, Stares MD, Browne HP, et al. Epithelial IL-22RA1-mediated fucosylation promotes intestinal colonization resistance to an opportunistic pathogen. Cell host & microbe. 2014;16(4):504-16.
489. Pickard JM, Maurice CF, Kinnebrew MA, Abt MC, Schenten D, Golovkina TV, et al. Rapid fucosylation of intestinal epithelium sustains host-commensal symbiosis in sickness. Nature. 2014;514(7524):638-41.
490. Nanthakumar NN, Meng D, Newburg DS. Glucocorticoids and microbiota regulate ontogeny of intestinal fucosyltransferase 2 requisite for gut homeostasis. Glycobiology. 2013;23(10):1131-41.
491. McGuckin MA, Linden SK, Sutton P, Florin TH. Mucin dynamics and enteric pathogens. Nat Rev Microbiol. 2011;9(4):265-78.
492. Xiao Y, Wang XQ, Yu Y, Guo Y, Xu X, Gong L, et al. Comprehensive mutation screening for 10 genes in Chinese patients suffering very early onset inflammatory bowel disease. World J Gastroenterol. 2016;22(24):5578-88.
493. Hu DY, Shao XX, Xu CL, Xia SL, Yu LQ, Jiang LJ, et al. Associations of FUT2 and FUT3 gene polymorphisms with Crohn's disease in Chinese patients. J Gastroenterol Hepatol. 2014;29(10):1778-85.
494. Kaur M, Panikkath D, Yan X, Liu Z, Berel D, Li D, et al. Perianal Crohn's Disease is Associated with Distal Colonic Disease, Stricturing Disease Behavior, IBD-Associated Serologies and Genetic Variation in the JAK-STAT Pathway. Inflamm Bowel Dis. 2016;22(4):862-9.
495. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature genetics. 2010;42(12):1118-25.
496. McGovern DP, Jones MR, Taylor KD, Marciante K, Yan X, Dubinsky M, et al. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Human molecular genetics. 2010;19(17):3468-76.

497. Aheman A, Luo HS, Gao F. Association of fucosyltransferase 2 gene variants with ulcerative colitis in Han and Uyghur patients in China. World J Gastroenterol. 2012;18(34):4758-64.
498. Rupp C, Friedrich K, Folseraas T, Wannhoff A, Bode KA, Weiss KH, et al. Fut2 genotype is a risk factor for dominant stenosis and biliary candida infections in primary sclerosing cholangitis. Aliment Pharmacol Ther. 2014;39(8):873-82.
499. Henriksen EK, Melum E, Karlsen TH. Update on primary sclerosing cholangitis genetics. Curr Opin Gastroenterol. 2014;30(3):310-9.
500. Folseraas T, Melum E, Rausch P, Juran BD, Ellinghaus E, Shiryaev A, et al. Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. J Hepatol. 2012;57(2):366-75.
501. Parmar AS, Alakulppi N, Paavola-Sakki P, Kurppa K, Halme L, Farkkila M, et al. Association study of FUT2 (rs601338) with celiac disease and inflammatory bowel disease in the Finnish population. Tissue Antigens. 2012;80(6):488-93.
502. Soejima M, Pang H, Koda Y. Genetic variation of FUT2 in a Ghanaian population: identification of four novel mutations and inference of balancing selection. Ann Hematol. 2007;86(3):199-204.
503. Walsh EC, Sabeti P, Hutcheson HB, Fry B, Schaffner SF, de Bakker PI, et al. Searching for signals of evolutionary selection in 168 genes related to immune function. Human genetics. 2006;119(1-2):92-102.
504. Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, et al. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome research. 2009;19(2):199-212.
505. Lis-Kuberka J, Katnik-Prastowska I, Berghausen-Mazur M, Orczyk-Pawilowicz M. Lectin-based analysis of fucosylated glycoproteins of human skim milk during 47 days of lactation. Glycoconj J. 2015;32(9):665-74.
506. Guix S, Asanaka M, Katayama K, Crawford SE, Neill FH, Atmar RL, et al. Norwalk virus RNA is infectious in mammalian cells. J Virol. 2007;81(22):12238-48.
507. Silva LM, Carvalho AS, Guillon P, Seixas S, Azevedo M, Almeida R, et al. Infection-associated FUT2 (Fucosyltransferase 2) genetic variation and impact on functionality assessed by in vivo studies. Glycoconj J. 2010;27(1):61-8.
508. Kaufman WL, Kocman I, Agrawal V, Rahn HP, Besser D, Gossen M. Homogeneity and persistence of transgene expression by omitting antibiotic selection in cell line isolation. Nucleic acids research. 2008;36(17):e111.
509. Krishnan M, Park JM, Cao F, Wang D, Paulmurugan R, Tseng JR, et al. Effects of epigenetic modulation on reporter gene expression: implications for stem cell imaging. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2006;20(1):106-8.
510. Davis RH. The age of model organisms. Nature reviews Genetics. 2004;5(1):69-76.
511. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annual review of genetics. 2005;39:309-38.
512. Studer RA, Robinson-Rechavi M. How confident can we be that orthologs are similar, but paralogs differ? Trends in genetics : TIG. 2009;25(5):210-6.

513. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol. 2011;7(6):e1002073.
514. Gharib WH, Robinson-Rechavi M. When orthologs diverge between human and mouse. Brief Bioinform. 2011;12(5):436-41.
515. Rosenthal N, Brown S. The mouse ascending: perspectives for human-disease models. Nature cell biology. 2007;9(9):993-9.
516. Hardouin SN, Nagy A. Mouse models for human disease. Clin Genet. 2000;57(4):237-44.
517. Cox RD, Brown SD. Rodent models of genetic disease. Current opinion in genetics & development. 2003;13(3):278-83.
518. Enard W. Mouse models of human evolution. Current opinion in genetics & development. 2014;29:75-80.
519. Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, et al. Myosin gene mutation correlates with anatomical changes in the human lineage. Nature. 2004;428(6981):415-8.
520. Hedlund M, Tangvoranuntakul P, Takematsu H, Long JM, Housley GD, Kozutsumi Y, et al. N-glycolylneuraminic acid deficiency in mice: implications for human biology and evolution. Molecular and cellular biology. 2007;27(12):4340-6.
521. Enard W, Gehre S, Hammerschmidt K, Holter SM, Blass T, Somel M, et al. A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell. 2009;137(5):961-71.
522. Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, et al. The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell. 2013;154(1):185-96.
523. McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, et al. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature. 2011;471(7337):216-9.
524. Dominski Z, Yang XC, Purdy M, Wagner EJ, Marzluff WF. A CPSF-73 homologue is required for cell cycle progression but not cell growth and interacts with a protein having features of CPSF-100. Molecular and cellular biology. 2005;25(4):1489-500.
525. Otani Y, Nakatsu Y, Sakoda H, Fukushima T, Fujishiro M, Kushiyama A, et al. Integrator complex plays an essential role in adipose differentiation. Biochemical and biophysical research communications. 2013;434(2):197-202.
526. Hata T, Nakayama M. Targeted disruption of the murine large nuclear KIAA1440/Ints1 protein causes growth arrest in early blastocyst stage embryos and eventual apoptotic cell death. Biochimica et biophysica acta. 2007;1773(7):1039-51.
527. Tao S, Cai Y, Sampath K. The Integrator subunits function in hematopoiesis by modulating Smad/BMP signaling. Development. 2009;136(16):2757-65.
528. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nature genetics. 2012;44(10):1084-9.
529. Jurica MS, Licklider LJ, Gygi SR, Grigorieff N, Moore MJ. Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. Rna. 2002;8(4):426-39.

530. Estrada K, Styrkarsdottir U, Evangelou E, Hsu YH, Duncan EL, Ntzani EE, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nature genetics. 2012;44(5):491-501.
531. Moayyeri A, Hsu YH, Karasik D, Estrada K, Xiao SM, Nielson C, et al. Genetic determinants of heel bone properties: genome-wide association meta-analysis and replication in the GEFOS/GENOMOS consortium. Human molecular genetics. 2014.
532. Olalde I, Allentoft ME, Sanchez-Quinto F, Santpere G, Chiang CW, DeGiorgio M, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature. 2014;507(7491):225-8.
533. Krishnan S, Chen S, Turcatel G, Arditi M, Prasadarao NV. Regulation of Toll-like receptor 2 interaction with Ecgp96 controls Escherichia coli K1 invasion of brain endothelial cells. Cellular microbiology. 2013;15(1):63-81.
534. Cal S, Moncada-Pazos A, Lopez-Otin C. Expanding the complexity of the human degradome: polyserases and their tandem serine protease domains. Frontiers in bioscience : a journal and virtual library. 2007;12:4661-9.
535. Cal S, Peinado JR, Llamazares M, Quesada V, Moncada-Pazos A, Garabaya C, et al. Identification and characterization of human polyserase-3, a novel protein with tandem serine-protease domains in the same polypeptide chain. BMC biochemistry. 2006;7:9.
536. Tonne JM, Sakuma T, Deeds MC, Munoz-Gomez M, Barry MA, Kudva YC, et al. Global gene expression profiling of pancreatic islets in mice during streptozotocin-induced beta-cell damage and pancreatic Glp-1 gene therapy. Disease models & mechanisms. 2013;6(5):1236-45.
537. Stuart PE, Nair RP, Ellinghaus E, Ding J, Tejasvi T, Gudjonsson JE, et al. Genome-wide association analysis identifies three psoriasis susceptibility loci. Nature genetics. 2010;42(11):1000-4.
538. Garcia-Gonzalo FR, Cruz C, Munoz P, Mazurek S, Eigenbrodt E, Ventura F, et al. Interaction between HERC1 and M2-type pyruvate kinase. FEBS letters. 2003;539(1-3):78-84.
539. Mashimo T, Hadjebi O, Amair-Pinedo F, Tsurumi T, Langa F, Serikawa T, et al. Progressive Purkinje cell degeneration in tambaleante mutant mice is a consequence of a missense mutation in HERC1 E3 ubiquitin ligase. PLoS genetics. 2009;5(12):e1000784.
540. Rosa JL, Casaroli-Marano RP, Buckler AJ, Vilaro S, Barbacid M. p619, a giant protein related to the chromosome condensation regulator RCC1, stimulates guanine nucleotide exchange on ARF1 and Rab proteins. The EMBO journal. 1996;15(16):4262-73.
541. Craig DW, O'Shaughnessy JA, Kiefer JA, Aldrich J, Sinari S, Moses TM, et al. Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. Molecular cancer therapeutics. 2013;12(1):104-16.
542. Diouf B, Cheng Q, Krynetskaia NF, Yang W, Cheok M, Pei D, et al. Somatic deletions of genes regulating MSH2 protein stability cause DNA mismatch repair deficiency and drug resistance in human leukemia cells. Nature medicine. 2011;17(10):1298-303.

543. Yuasa I, Umetsu K, Nishimukai H, Fukumori Y, Harihara S, Saitou N, et al. HERC1 polymorphisms: population-specific variations in haplotype composition. Cell biochemistry and function. 2009;27(6):402-5.
544. Aggarwal S, Bhowmik AD, Ramprasad VL, Murugan S, Dalal A. A splice site mutation in HERC1 leads to syndromic intellectual disability with macrocephaly and facial dysmorphism: Further delineation of the phenotypic spectrum. American journal of medical genetics Part A. 2016;170(7):1868-73.
545. Ortega-Recalde O, Beltran OI, Galvez JM, Palma-Montero A, Restrepo CM, Mateus HE, et al. Biallelic HERC1 mutations in a syndromic form of overgrowth and intellectual disability. Clin Genet. 2015;88(4):e1-3.
546. Nguyen LS, Schneider T, Rio M, Moutton S, Siquier-Pernet K, Verny F, et al. A nonsense variant in HERC1 is associated with intellectual disability, megalencephaly, thick corpus callosum and cerebellar atrophy. European journal of human genetics : EJHG. 2016;24(3):455-8.
547. Hashimoto R, Nakazawa T, Tsurusaki Y, Yasuda Y, Nagayasu K, Matsumura K, et al. Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. Journal of human genetics. 2016;61(3):199-206.
548. Bachiller S, Rybkina T, Porras-Garcia E, Perez-Villegas E, Tabares L, Armengol JA, et al. The HERC1 E3 Ubiquitin Ligase is essential for normal development and for neurotransmission at the mouse neuromuscular junction. Cell Mol Life Sci. 2015;72(15):2961-71.
549. Jablonski NG. Skin : a natural history. Berkeley: University of California Press; 2006. xiii, 266 p. p.
550. Iyengar B. The hair follicle: a specialised UV receptor in the human skin? Biol Signals Recept. 1998;7(3):188-94.
551. De la Mettrie R, Saint-Leger D, Loussouarn G, Garcel A, Porter C, Langaney A. Shape variability and classification of human hair: a worldwide approach. Hum Biol. 2007;79(3):265-81.
552. Koniukhov BV, Malinina NA, Martynov M. [The we gene is a modifier of the wal gene in mice]. Genetika. 2004;40(7):968-74.
553. Koniukhov BV, Kupriianov SD. [The mutant gene wellhaarig disturbs the differentiation of hair follicle cells in the mouse]. Ontogenez. 1990;21(1):56-62.
554. Loussouarn G. African hair growth parameters. Br J Dermatol. 2001;145(2):294-7.
555. Khumalo NP, Gumedze F. African hair length in a school population: a clue to disease pathogenesis? J Cosmet Dermatol. 2007;6(3):144-51.
556. Schwartz GG, Rosenblum LA. Allometry of primate hair density and the evolution of human hairlessness. Am J Phys Anthropol. 1981;55(1):9-12.
557. Moura DS, Fernandez IF, Marin-Royo G, Lopez-Sanchez I, Martin-Doncel E, Vega FM, et al. Oncogenic Sox2 regulates and cooperates with VRK1 in cell cycle progression and differentiation. Sci Rep. 2016;6:28532.
558. Liu J, Wang Y, He S, Xu X, Huang Y, Tang J, et al. Expression of vaccinia-related kinase 1 (VRK1) accelerates cell proliferation but overcomes cell adhesion mediated drug resistance (CAM-DR) in multiple myeloma. Hematology. 2016:1-10.

559. Renbaum P, Kellerman E, Jaron R, Geiger D, Segel R, Lee M, et al. Spinal muscular atrophy with pontocerebellar hypoplasia is caused by a mutation in the VRK1 gene. American journal of human genetics. 2009;85(2):281-9.
560. Najmabadi H, Hu H, Garshasbi M, Zemojtel T, Abedini SS, Chen W, et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature. 2011;478(7367):57-63.
561. Stoll M, Teoh H, Lee J, Reddel S, Zhu Y, Buckley M, et al. Novel motor phenotypes in patients with VRK1 mutations without pontocerebellar hypoplasia. Neurology. 2016;87(1):65-70.
562. Nguyen TP, Biliciler S, Wiszniewski W, Sheikh K. Expanding Phenotype of VRK1 Mutations in Motor Neuron Disease. J Clin Neuromuscul Dis. 2015;17(2):69-71.
563. Vinograd-Byk H, Sapir T, Cantarero L, Lazo PA, Zeligson S, Lev D, et al. The spinal muscular atrophy with pontocerebellar hypoplasia gene VRK1 regulates neuronal migration through an amyloid-beta precursor protein-dependent mechanism. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2015;35(3):936-42.
564. Gonzaga-Jauregui C, Lotze T, Jamal L, Penney S, Campbell IM, Pehlivan D, et al. Mutations in VRK1 associated with complex motor and sensory axonal neuropathy plus microcephaly. JAMA Neurol. 2013;70(12):1491-8.
565. Salzano M, Vazquez-Cedeira M, Sanz-Garcia M, Valbuena A, Blanco S, Fernandez IF, et al. Vaccinia-related kinase 1 (VRK1) confers resistance to DNA-damaging agents in human breast cancer by affecting DNA damage response. Oncotarget. 2014;5(7):1770-8.
566. Sanz-Garcia M, Monsalve DM, Sevilla A, Lazo PA. Vaccinia-related kinase 1 (VRK1) is an upstream nucleosomal kinase required for the assembly of 53BP1 foci in response to ionizing radiation-induced DNA damage. The Journal of biological chemistry. 2012;287(28):23757-68.
567. Salzano M, Sanz-Garcia M, Monsalve DM, Moura DS, Lazo PA. VRK1 chromatin kinase phosphorylates H2AX and is required for foci formation induced by DNA damage. Epigenetics. 2015;10(5):373-83.
568. Monsalve DM, Campillo-Marcos I, Salzano M, Sanz-Garcia M, Cantarero L, Lazo PA. VRK1 phosphorylates and protects NBS1 from ubiquitination and proteasomal degradation in response to DNA damage. Biochimica et biophysica acta. 2016;1863(4):760-9.
569. Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikainen LP, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nature genetics. 2012;44(3):269-76.
570. Comuzzie AG, Cole SA, Laston SL, Voruganti VS, Haack K, Gibbs RA, et al. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. PloS one. 2012;7(12):e51954.
571. Sanchez J, Priego T, Pico C, Ahrens W, De Henauw S, Fraterman A, et al. Blood cells as a source of transcriptional biomarkers of childhood obesity and its related metabolic alterations: results of the IDEFICS study. J Clin Endocrinol Metab. 2012;97(4):E648-52.
572. Saunders CL, Chiodini BD, Sham P, Lewis CM, Abkevich V, Adeyemo AA, et al. Meta-analysis of genome-wide linkage studies in BMI and obesity. Obesity. 2007;15(9):2263-75.

573. Collins SA, Sinclair G, McIntosh S, Bamforth F, Thompson R, Sobol I, et al. Carnitine palmitoyltransferase 1A (CPT1A) P479L prevalence in live newborns in Yukon, Northwest Territories, and Nunavut. Molecular genetics and metabolism. 2010;101(2-3):200-4.
574. Collins SA, Surmala P, Osborne G, Greenberg C, Bathory LW, Edmunds-Potvin S, et al. Causes and risk factors for infant mortality in Nunavut, Canada 1999-2011. BMC pediatrics. 2012;12:190.
575. Sinclair GB, Collins S, Popescu O, McFadden D, Arbour L, Vallance HD. Carnitine palmitoyltransferase I and sudden unexpected infant death in British Columbia First Nations. Pediatrics. 2012;130(5):e1162-9.
576. Brown NF, Mullur RS, Subramanian I, Esser V, Bennett MJ, Saudubray JM, et al. Molecular characterization of L-CPT I deficiency in six patients: insights into function of the native enzyme. Journal of lipid research. 2001;42(7):1134-42.
577. Akkaoui M, Cohen I, Esnous C, Lenoir V, Sournac M, Girard J, et al. Modulation of the hepatic malonyl-CoA-carnitine palmitoyltransferase 1A partnership creates a metabolic switch allowing oxidation of de novo fatty acids. Biochem J. 2009;420(3):429-38.
578. Greenberg CR, Dilling LA, Thompson GR, Seargeant LE, Haworth JC, Phillips S, et al. The paradox of the carnitine palmitoyltransferase type Ia P479L variant in Canadian Aboriginal populations. Molecular genetics and metabolism. 2009;96(4):201-7.
579. Gessner BD, Gillingham MB, Wood T, Koeller DM. Association of a genetic variant of carnitine palmitoyltransferase 1A with infections in Alaska Native children. J Pediatr. 2013;163(6):1716-21.
580. Gillingham MB, Hirschfeld M, Lowe S, Matern D, Shoemaker J, Lambert WE, et al. Impaired fasting tolerance among Alaska native children with a common carnitine palmitoyltransferase 1A sequence variant. Molecular genetics and metabolism. 2011;104(3):261-4.
581. Rajakumar C, Ban MR, Cao H, Young TK, Bjerregaard P, Hegele RA. Carnitine palmitoyltransferase IA polymorphism P479L is common in Greenland Inuit and is associated with elevated plasma apolipoprotein A-I. Journal of lipid research. 2009;50(6):1223-8.
582. Gessner BD, Gillingham MB, Johnson MA, Richards CS, Lambert WE, Sesser D, et al. Prevalence and distribution of the c.1436C-->T sequence variant of carnitine palmitoyltransferase 1A among Alaska Native infants. J Pediatr. 2011;158(1):124-9.
583. Lemas DJ, Wiener HW, O'Brien DM, Hopkins S, Stanhope KL, Havel PJ, et al. Genetic polymorphisms in carnitine palmitoyltransferase 1A gene are associated with variation in body composition and fasting lipid traits in Yup'ik Eskimos. Journal of lipid research. 2012;53(1):175-84.
584. Madsen L, Rustan AC, Vaagenes H, Berge K, Dyroy E, Berge RK. Eicosapentaenoic and docosahexaenoic acid affect mitochondrial and peroxisomal fatty acid oxidation in relation to substrate preference. Lipids. 1999;34(9):951-63.
585. Andersen MK, Jorsboe E, Sandholt CH, Grarup N, Jorgensen ME, Faergeman NJ, et al. Identification of Novel Genetic Determinants of Erythrocyte Membrane Fatty Acid Composition among Greenlanders. PLoS genetics. 2016;12(6):e1006119.

586. Reynes B, Garcia-Ruiz E, Oliver P, Palou A. Gene expression of peripheral blood mononuclear cells is affected by cold exposure. Am J Physiol Regul Integr Comp Physiol. 2015;309(8):R824-34.
587. Nyman LR, Cox KB, Hoppel CL, Kerner J, Barnoski BL, Hamm DA, et al. Homozygous carnitine palmitoyltransferase 1a (liver isoform) deficiency is lethal in the mouse. Molecular genetics and metabolism. 2005;86(1-2):179-87.
588. Ahmed ZM, Riazuddin S, Aye S, Ali RA, Venselaar H, Anwar S, et al. Gene structure and mutant alleles of PCDH15: nonsyndromic deafness DFNB23 and type 1 Usher syndrome. Human genetics. 2008;124(3):215-23.
589. Alagramam KN, Murcia CL, Kwon HY, Pawlowski KS, Wright CG, Woychik RP. The mouse Ames waltzer hearing-loss mutant is caused by mutation of Pcdh15, a novel protocadherin gene. Nature genetics. 2001;27(1):99-102.
590. Ahmed ZM, Riazuddin S, Ahmad J, Bernstein SL, Guo Y, Sabar MF, et al. PCDH15 is expressed in the neurosensory epithelium of the eye and ear and mutant alleles are responsible for both USH1F and DFNB23. Human molecular genetics. 2003;12(24):3215-23.
591. Ahmed ZM, Goodyear R, Riazuddin S, Lagziel A, Legan PK, Behra M, et al. The tip-link antigen, a protein associated with the transduction complex of sensory hair cells, is protocadherin-15. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2006;26(26):7022-34.
592. Kazmierczak P, Sakaguchi H, Tokita J, Wilson-Kubalek EM, Milligan RA, Muller U, et al. Cadherin 23 and protocadherin 15 interact to form tip-link filaments in sensory hair cells. Nature. 2007;449(7158):87-91.
593. Geng R, Sotomayor M, Kinder KJ, Gopal SR, Gerka-Stuyt J, Chen DH, et al. Noddy, a mouse harboring a missense mutation in protocadherin-15, reveals the impact of disrupting a critical interaction site between tip-link cadherins in inner ear hair cells. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2013;33(10):4395-404.
594. Zheng QY, Yan D, Ouyang XM, Du LL, Yu H, Chang B, et al. Digenic inheritance of deafness caused by mutations in genes encoding cadherin 23 and protocadherin 15 in mice and humans. Human molecular genetics. 2005;14(1):103-11.
595. Naoi K, Kuramoto T, Kuwamura Y, Gohma H, Kuwamura M, Serikawa T. Characterization of the Kyoto circling (KCI) rat carrying a spontaneous nonsense mutation in the protocadherin 15 (Pcdh15) gene. Experimental animals / Japanese Association for Laboratory Animal Science. 2009;58(1):1-10.
596. Hampton LL, Wright CG, Alagramam KN, Battey JF, Noben-Trauth K. A new spontaneous mutation in the mouse Ames waltzer gene, Pcdh15. Hearing research. 2003;180(1-2):67-75.
597. Webb SW, Grillet N, Andrade LR, Xiong W, Swarthout L, Della Santina CC, et al. Regulation of PCDH15 function in mechanosensory hair cells by alternative splicing of the cytoplasmic domain. Development. 2011;138(8):1607-17.
598. Rouget-Quermalet V, Giustiniani J, Marie-Cardine A, Beaud G, Besnard F, Loyaux D, et al. Protocadherin 15 (PCDH15): a new secreted isoform and a potential marker for NK/T cell lymphomas. Oncogene. 2006;25(19):2807-11.

599. Oki NO, Motsinger-Reif AA, Antas PR, Levy S, Holland SM, Sterling TR. Novel human genetic variants associated with extrapulmonary tuberculosis: a pilot genome wide association study. BMC research notes. 2011;4:28.
600. Grupe A, Li Y, Rowland C, Nowotny P, Hinrichs AL, Smemo S, et al. A scan of chromosome 10 identifies a novel locus showing strong association with late-onset Alzheimer disease. American journal of human genetics. 2006;78(1):78-88.
601. Ovsyannikova IG, Kennedy RB, O'Byrne M, Jacobson RM, Pankratz VS, Poland GA. Genome-wide association study of antibody response to smallpox vaccine. Vaccine. 2012;30(28):4182-9.
602. Croteau-Chonka DC, Marvelle AF, Lange EM, Lee NR, Adair LS, Lange LA, et al. Genome-wide association study of anthropometric traits and evidence of interactions with age and study year in Filipino women. Obesity. 2011;19(5):1019-27.
603. Huertas-Vazquez A, Plaisier CL, Geng R, Haas BE, Lee J, Greevenbroek MM, et al. A nonsynonymous SNP within PCDH15 is associated with lipid traits in familial combined hyperlipidemia. Human genetics. 2010;127(1):83-9.
604. Yasunaga S, Grati M, Chardenoux S, Smith TN, Friedman TB, Lalwani AK, et al. OTOF encodes multiple long and short isoforms: genetic evidence that the long ones underlie recessive deafness DFNB9. American journal of human genetics. 2000;67(3):591-600.
605. Yasunaga S, Grati M, Cohen-Salmon M, El-Amraoui A, Mustapha M, Salem N, et al. A mutation in OTOF, encoding otoferlin, a FER-1-like protein, causes DFNB9, a nonsyndromic form of deafness. Nature genetics. 1999;21(4):363-9.
606. Rouillon I, Marcolla A, Roux I, Marlin S, Feldmann D, Couderc R, et al. Results of cochlear implantation in two children with mutations in the OTOF gene. Int J Pediatr Otorhinolaryngol. 2006;70(4):689-96.
607. Pangrsic T, Reisinger E, Moser T. Otoferlin: a multi-C2 domain protein essential for hearing. Trends Neurosci. 2012;35(11):671-80.
608. Roux I, Safieddine S, Nouvian R, Grati M, Simmler MC, Bahloul A, et al. Otoferlin, defective in a human deafness form, is essential for exocytosis at the auditory ribbon synapse. Cell. 2006;127(2):277-89.
609. Lek A, Evesson FJ, Sutton RB, North KN, Cooper ST. Ferlins: regulators of vesicle fusion for auditory neurotransmission, receptor trafficking and membrane repair. Traffic. 2012;13(2):185-94.
610. Dulon D, Safieddine S, Jones SM, Petit C. Otoferlin is critical for a highly sensitive and linear calcium-dependent exocytosis at vestibular hair cell ribbon synapses. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2009;29(34):10474-87.
611. Beurg M, Safieddine S, Roux I, Bouleau Y, Petit C, Dulon D. Calcium- and otoferlin-dependent exocytosis by immature outer hair cells. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2008;28(8):1798-803.
612. Beurg M, Michalski N, Safieddine S, Bouleau Y, Schneggenburger R, Chapman ER, et al. Control of exocytosis by synaptotagmins and otoferlin in auditory hair cells. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2010;30(40):13281-90.

613. Mahdieh N, Shirkavand A, Rabbani B, Tekin M, Akbari B, Akbari MT, et al. Screening of OTOF mutations in Iran: a novel mutation and review. Int J Pediatr Otorhinolaryngol. 2012;76(11):1610-5.
614. Zhang Q, Lan L, Shi W, Yu L, Xie LY, Xiong F, et al. Temperature sensitive auditory neuropathy. Hearing research. 2016;335:53-63.
615. Starr A, Sininger Y, Winter M, Derebery MJ, Oba S, Michalewski HJ. Transient deafness due to temperature-sensitive auditory neuropathy. Ear Hear. 1998;19(3):169-79.
616. Marlin S, Feldmann D, Nguyen Y, Rouillon I, Loundon N, Jonard L, et al. Temperature-sensitive auditory neuropathy associated with an otoferlin mutation: Deafening fever! Biochemical and biophysical research communications. 2010;394(3):737-42.
617. Longo-Guess C, Gagnon LH, Bergstrom DE, Johnson KR. A missense mutation in the conserved C2B domain of otoferlin causes deafness in a new mouse model of DFNB9. Hearing research. 2007;234(1-2):21-8.
618. Hudson NJ, Baker ML, Hart NS, Wynne JW, Gu Q, Huang Z, et al. Sensory rewiring in an echolocator: genome-wide modification of retinogenic and auditory genes in the bat Myotis davidii. G3 (Bethesda). 2014;4(10):1825-35.
619. Shen YY, Liang L, Li GS, Murphy RW, Zhang YP. Parallel evolution of auditory genes for echolocation in bats and toothed whales. PLoS genetics. 2012;8(6):e1002788.
620. Akey JM, Swanson WJ, Madeoy J, Eberle M, Shriver MD. TRPV6 exhibits unusual patterns of polymorphism and divergence in worldwide populations. Human molecular genetics. 2006;15(13):2106-13.
621. Hughes DA, Tang K, Strotmann R, Schoneberg T, Prenen J, Nilius B, et al. Parallel selection on TRPV6 in human populations. PloS one. 2008;3(2):e1686.
622. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. Science. 2016;354(6313):760-4.
623. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics. 2015;47(3):291-5.
624. Matthews PM, Sudlow C. The UK Biobank. Brain. 2015;138(Pt 12):3463-5.
625. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779.
626. Zhen Y, Andolfatto P. Methods to detect selection on noncoding DNA. Methods Mol Biol. 2012;856:141-59.
627. Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. Gene expression levels are a target of recent natural selection in the human genome. Molecular biology and evolution. 2009;26(3):649-58.
628. Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M, et al. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome biology. 2016;17:19.
629. Wiberg RA, Halligan DL, Ness RW, Necsulea A, Kaessmann H, Keightley PD. Assessing Recent Selection and Functionality at Long Noncoding RNA Loci in the Mouse Genome. Genome Biol Evol. 2015;7(8):2432-44.

630. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253):aab3761.
631. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nature genetics. 2007;39(10):1256-60.
632. Hardwick RJ, Menard A, Sironi M, Milet J, Garcia A, Sese C, et al. Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis. Human genetics. 2014;133(1):69-83.
633. Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, et al. Adaptive evolution of UGT2B17 copy-number variation. American journal of human genetics. 2008;83(3):337-46.
634. Marian AJ. Elements of 'missing heritability'. Curr Opin Cardiol. 2012;27(3):197-201.
635. Pai AA, Pritchard JK, Gilad Y. The genetic and mechanistic basis for variation in gene regulation. PLoS genetics. 2015;11(1):e1004857.
636. Schubeler D. Function and information content of DNA methylation. Nature. 2015;517(7534):321-6.
637. Illingworth RS, Bird AP. CpG islands--'a rough guide'. FEBS letters. 2009;583(11):1713-20.
638. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, et al. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS genetics. 2014;10(9):e1004663.
639. Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, Sommer IE, et al. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. PloS one. 2009;4(8):e6767.
640. Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH, et al. DNA methylation profiles in monozygotic and dizygotic twins. Nature genetics. 2009;41(2):240-5.
641. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, et al. Genetic control of individual differences in gene-specific methylation in human brain. American journal of human genetics. 2010;86(3):411-9.
642. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS genetics. 2010;6(5):e1000952.
643. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome biology. 2011;12(1):R10.
644. Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477-81.
645. Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. Genome biology. 2012;13(2):R8.
646. Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, et al. DNA methylation contributes to natural human variation. Genome research. 2013;23(9):1363-72.
647. Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, et al. Genome-wide variation of cytosine modifications between European and African

populations and the implications for complex traits. Genetics. 2013;194(4):987-96.
648. Fagny M, Patin E, MacIsaac JL, Rotival M, Flutre T, Jones MJ, et al. The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nature communications. 2015;6:10047.

Appendix A

Lists of top *SIS* candidate genes

Top 7% of positively selected protein coding genes and their *SSI* values are given in the following pages.

CHR – chromosome
GENE NAME – associated gene name
SIS – Selection Support Index

CHR	GENE NAME	SSI	CHR	GENE NAME	SSI	CHR	GENE NAME	SSI	CHR	GENE NAME	SSI	CHR	GENE NAME	SSI
2	*EDAR*	0.76	14	*AP5M1*	0.20	10	*ADK*	0.16	1	*CLSPN*	0.14	9	*C9orf156*	0.12
4	*SLC30A9*	0.70	16	*CBFB*	0.20	3	*USP4*	0.16	19	*CEACAM1*	0.14	9	*HEMGN*	0.12
4	*BEND4*	0.55	16	*TPPP3*	0.20	2	*RTKN*	0.16	18	*CD226*	0.14	22	*CDC45*	0.12
15	*SLC24A5*	0.49	10	*CFAP70*	0.20	15	*CTXN2*	0.16	1	*ST6GALNAC3*	0.14	4	*SEC24B*	0.12
15	*HERC1*	0.49	16	*KCTD19*	0.20	1	*TEKT2*	0.16	2	*VSNL1*	0.14	19	*ZBTB7A*	0.12
2	*UBXN4*	0.47	20	*SLC32A1*	0.20	20	*PROCR*	0.16	20	*FAM83D*	0.14	2	*YWHAQ*	0.12
2	*LCT*	0.46	13	*MPHOSPH8*	0.20	10	*SEC24C*	0.16	15	*TRPM1*	0.14	4	*KLB*	0.12
4	*TMEM33*	0.45	2	*PCGF1*	0.20	19	*ZNF780B*	0.16	1	*HMCN1*	0.14	9	*C8G*	0.12
2	*MCM6*	0.44	5	*APC*	0.20	5	*TRIM23*	0.16	3	*CCDC36*	0.14	7	*C7orf34*	0.12
15	*MYEF2*	0.43	1	*TSTD1*	0.20	21	*WDR4*	0.16	15	*RORA*	0.14	2	*SUCLG1*	0.12
17	*BCAS3*	0.42	10	*CCDC172*	0.20	7	*IFRD1*	0.16	6	*LAMA2*	0.14	19	*CXCL17*	0.12
2	*GCC2*	0.40	3	*RPEP2*	0.19	17	*SLC25A19*	0.16	18	*RTTN*	0.14	7	*SLC26A5*	0.12
1	*ZMYM6*	0.40	3	*AMT*	0.19	6	*PREP*	0.16	2	*SLC4A5*	0.14	3	*LRIG1*	0.12
12	*PAWR*	0.39	11	*VPS37C*	0.19	16	*SLC9A5*	0.16	10	*PPP3CB*	0.14	1	*NCDN*	0.12
2	*DARS*	0.38	2	*ALMS1-IT1*	0.19	1	*AGO3*	0.16	2	*HECW2*	0.14	12	*OSBPL8*	0.12
15	*HERC2*	0.37	2	*INO80B*	0.19	13	*UFM1*	0.16	12	*TMTC3*	0.14	11	*OR5AS1*	0.12
1	*ZMYM4*	0.37	1	*BCAR3*	0.19	7	*C7orf60*	0.16	5	*C5orf34*	0.14	14	*NRXN3*	0.12
1	*ZMYM1*	0.37	10	*PIP4K2A*	0.19	5	*TMEM232*	0.16	3	*CHMP2B*	0.14	20	*CNBD2*	0.12
14	*ACTN1*	0.37	3	*NICN1*	0.19	2	*CCDC148*	0.16	10	*SCD*	0.14	16	*LONP2*	0.12
7	*LSMEM1*	0.36	16	*TMEM208*	0.19	5	*CENPK*	0.16	2	*CENPO*	0.14	2	*ADCY3*	0.12
12	*SYT1*	0.35	2	*GEN1*	0.19	16	*VKORC1*	0.16	15	*SENP8*	0.14	16	*THAP11*	0.12
19	*YIF1B*	0.34	4	*C4orf46*	0.19	5	*REEP5*	0.16	10	*DNAJC9*	0.14	4	*WDR19*	0.12
2	*R3HDM1*	0.34	16	*GABARAPL2*	0.19	6	*TAAR2*	0.16	22	*CDPF1*	0.14	7	*KCND2*	0.12
15	*SLC12A1*	0.34	15	*FBXL22*	0.19	19	*SPINT2*	0.16	9	*SETX*	0.14	8	*TEX15*	0.12
4	*DCAF4L1*	0.33	7	*DGKI*	0.19	3	*POU1F1*	0.16	16	*N4BP1*	0.14	16	*KAT8*	0.12
3	*EPHB1*	0.33	5	*SLC25A46*	0.19	22	*MRPL40*	0.16	7	*SRPK2*	0.14	16	*C16orf70*	0.12
3	*LPP*	0.32	6	*MRAP2*	0.19	2	*WDR54*	0.16	5	*PAIP1*	0.14	5	*SLC27A6*	0.12
2	*LIMS1*	0.32	11	*PGA5*	0.19	8	*PINX1*	0.16	20	*EDEM2*	0.14	6	*PHACTR1*	0.12
10	*PCDH15*	0.32	1	*OR2L13*	0.18	16	*ATP6V0D1*	0.16	6	*VNN1*	0.14	7	*MYH16*	0.12
2	*RANBP2*	0.32	3	*APEH*	0.18	14	*NAA30*	0.16	8	*MTMR9*	0.14	15	*TMEM87A*	0.12
2	*CCDC138*	0.30	2	*KCNH7*	0.18	9	*PTGS1*	0.16	12	*GOLT1B*	0.14	14	*PRKCH*	0.12
15	*APBA2*	0.30	1	*DLGAP3*	0.18	1	*ROR1*	0.16	12	*STRAP*	0.14	10	*GPR158*	0.12
12	*TMEM117*	0.30	10	*FUT11*	0.18	7	*PRKAG2*	0.16	15	*MYO5C*	0.14	12	*DDX55*	0.12
12	*PPP1R12A*	0.30	3	*CCDC71*	0.18	16	*CHST5*	0.16	14	*SNAPC1*	0.14	16	*PARN*	0.12
22	*APOL2*	0.29	2	*CXCR4*	0.18	3	*RHOA*	0.16	17	*SHMT1*	0.14	5	*SKP2*	0.12
1	*SFPQ*	0.29	19	*CATSPERG*	0.18	3	*ZBTB20*	0.16	3	*PTPLB*	0.14	22	*PPARA*	0.12
7	*TRPV6*	0.29	1	*F5*	0.18	8	*CSMD1*	0.15	1	*BLZF1*	0.14	12	*C12orf29*	0.12
2	*SULT1C2*	0.29	1	*ZBTB41*	0.18	1	*PTPN22*	0.15	2	*HADHB*	0.14	14	*KLHL28*	0.12
3	*OXSM*	0.28	6	*RPL32P1*	0.18	7	*PTCD1*	0.15	12	*ERC1*	0.13	6	*LMBRD1*	0.12
7	*TRPV5*	0.28	3	*DAG1*	0.18	1	*F13B*	0.15	20	*BMP7*	0.13	6	*UTRN*	0.12
15	*DAPK2*	0.28	16	*ELMO3*	0.18	19	*CNFN*	0.15	2	*IFIH1*	0.13	3	*SPTSSB*	0.12
2	*EXOC6B*	0.28	2	*MRPL53*	0.18	22	*CLDN5*	0.15	7	*KEL*	0.13	4	*TMEM156*	0.12
3	*SLC4A7*	0.28	6	*TAAR1*	0.18	1	*PHLDA3*	0.15	5	*C5orf28*	0.13	15	*KLF13*	0.12
10	*MRPS16*	0.28	19	*ZNF546*	0.18	6	*HLA-DPA2*	0.15	6	*NKAIN2*	0.13	4	*HMGB1P28*	0.12
15	*DUT*	0.27	16	*ABCC11*	0.18	1	*RPS6KA1*	0.15	15	*DUOX2*	0.13	20	*STK35*	0.12
10	*PNLIPRP3*	0.27	6	*CYB5R4*	0.18	1	*EVI5*	0.15	10	*P4HA1*	0.13	16	*ITGAM*	0.12
2	*SULT1C4*	0.27	7	*BUD31*	0.18	3	*MYLK*	0.15	7	*SDK1*	0.13	22	*APOL3*	0.12
22	*MYH9*	0.26	10	*SYNPO2L*	0.18	22	*PKDREJ*	0.15	22	*APOL4*	0.13	12	*GYS2*	0.12
17	*APPBP2*	0.26	3	*KLHDC8B*	0.18	10	*NUDT13*	0.15	22	*GTSE1*	0.13	9	*FKTN*	0.12
9	*BNC2*	0.26	8	*SNTG1*	0.18	8	*SLC35G5*	0.15	20	*DOK5*	0.13	17	*RNF135*	0.12
2	*GALNT5*	0.26	13	*PSPC1*	0.18	3	*FOXP1*	0.15	4	*KIAA1109*	0.13	4	*ADH1B*	0.12
3	*BAI3*	0.26	19	*PSMC4*	0.18	2	*MLPH*	0.15	17	*MED13*	0.13	17	*EVI2A*	0.12
16	*PLEKHG4*	0.26	5	*SLC45A2*	0.18	1	*F11R*	0.15	16	*TFAP4*	0.13	15	*SIN3A*	0.12
5	*MCC*	0.25	8	*XKR6*	0.18	10	*FAM149B1*	0.15	16	*BCL7C*	0.13	3	*HMGB3P12*	0.12
1	*PSMB2*	0.25	2	*CCDC142*	0.18	14	*ADAM21*	0.15	21	*TMPRSS3*	0.13	3	*TAF9BP1*	0.12
2	*WBP1*	0.25	2	*TTC31*	0.18	15	*CSNK1G1*	0.15	15	*MYO9A*	0.13	17	*DDX5*	0.12
3	*RBP2*	0.25	1	*ASPM*	0.18	18	*DTNA*	0.15	7	*ATP5J2*	0.13	1	*CDC42BPA*	0.12
3	*NGLY1*	0.25	10	*AGAP5*	0.18	16	*FBXL8*	0.15	16	*E2F4*	0.13	15	*VPS39*	0.12
5	*ARHGAP26*	0.25	2	*LBX2*	0.18	18	*CCDC102B*	0.15	16	*TRADD*	0.13	16	*ZCCHC14*	0.12
7	*TMEM168*	0.24	2	*ASAP2*	0.17	20	*EIF4E2P1*	0.15	1	*CSRP1*	0.13	15	*GNB5*	0.12
2	*ALMS1*	0.24	2	*CYP26B1*	0.17	20	*LLPHP1*	0.15	8	*PTP4A3*	0.13	5	*ADAMTS6*	0.12
3	*TBC1D5*	0.24	11	*NAV2*	0.17	1	*RPL5P4*	0.15	2	*DCTN1*	0.13	1	*ADPRHL2*	0.12
1	*OR2L2*	0.24	3	*EPHA6*	0.17	7	*PDAP1*	0.15	7	*CPSF4*	0.13	22	*C22orf39*	0.12
2	*SMC6*	0.24	1	*PVRL4*	0.17	10	*PKD2L1*	0.15	1	*TNNI1*	0.13	4	*KLHL5*	0.12
7	*CNTNAP2*	0.24	1	*USF1*	0.17	9	*DOCK8*	0.15	16	*FAM65A*	0.13	19	*LIPE*	0.12
14	*EXOC5*	0.24	1	*CRB1*	0.17	3	*C3orf84*	0.15	12	*ATXN2*	0.13	17	*PCTP*	0.12
1	*KIAA0319L*	0.24	8	*BLK*	0.17	15	*PKM*	0.15	12	*RECQL*	0.13	12	*C2CD5*	0.12
19	*C19orf33*	0.23	6	*SAMD3*	0.17	11	*VWCE*	0.15	15	*PTPN9*	0.13	16	*SDR42E1*	0.12
4	*ATP1B1P1*	0.23	4	*GABRA4*	0.17	1	*LEPR*	0.15	2	*GLI2*	0.13	6	*IL20RA*	0.12
6	*HLA-DPB1*	0.23	8	*UNC5D*	0.17	17	*SLC5A10*	0.15	14	*IGHG1*	0.13	20	*TGM3*	0.12
9	*TYRP1*	0.23	2	*NEUROD1*	0.17	3	*CD47*	0.15	1	*ADAM15*	0.13	20	*CPNE1*	0.12
16	*ADAT1*	0.23	12	*CACNA1C*	0.17	7	*GALNT11*	0.15	2	*THADA*	0.13	18	*RIT2*	0.12
17	*GRB2*	0.23	16	*FHOD1*	0.17	16	*EXOC3L1*	0.15	8	*EIF3H*	0.13	10	*COMMD3*	0.12
7	*EPHB6*	0.23	6	*HLA-DPA1*	0.17	16	*NOL3*	0.15	4	*TLR1*	0.13	12	*TBC1D30*	0.12
15	*HEXA*	0.23	6	*HLA-DPB2*	0.17	2	*PDE11A*	0.15	22	*UFD1L*	0.13	3	*FSTL1*	0.12
22	*LARGE*	0.22	10	*ZSWIM8*	0.17	1	*SLC44A5*	0.15	2	*SPR*	0.13	22	*MKL1*	0.12
10	*USP54*	0.22	14	*PPP2R5E*	0.17	14	*SYNJ2BP*	0.15	2	*DNAH6*	0.13	19	*GGN*	0.12
3	*TCTA*	0.22	10	*NRG3*	0.17	10	*SPAG6*	0.15	9	*CNTLN*	0.13	1	*COL8A2*	0.12
12	*TEAD4*	0.22	3	*COPB2*	0.17	15	*GRAMD2*	0.15	2	*NCOA1*	0.13	17	*PRPSAP2*	0.12
5	*CTNND2*	0.22	16	*LRRC29*	0.17	20	*RALGAPA2*	0.15	3	*FHIT*	0.13	11	*PRDM10*	0.12
15	*OCA2*	0.22	10	*NDST2*	0.17	10	*ANXA7*	0.14	15	*PPIB*	0.13	22	*SYNGR1*	0.11
16	*LRRC36*	0.22	5	*FGF1*	0.17	10	*ECD*	0.14	2	*PRPF40A*	0.13	1	*SCP2*	0.11
19	*FAM98C*	0.22	6	*HTR1B*	0.17	11	*POU2F3*	0.14	15	*CEMIP*	0.13	22	*HIRA*	0.11
3	*BBX*	0.22	6	*TAAR3*	0.17	17	*USP32*	0.14	2	*FMNL2*	0.13	7	*RELN*	0.11
18	*PTPRM*	0.21	1	*ARHGAP30*	0.17	15	*SNX1*	0.14	16	*WWOX*	0.13	21	*BACE2*	0.11
15	*C15orf43*	0.21	2	*PTRHD1*	0.17	16	*HSD11B2*	0.14	15	*SORD*	0.13	17	*DPRXP4*	0.11
10	*CELF2*	0.21	5	*SRP19*	0.17	16	*HSF4*	0.14	15	*ZNF609*	0.13	6	*POPDC3*	0.11
4	*LIMCH1*	0.21	20	*ACTR5*	0.17	15	*USP3*	0.14	11	*PGA3*	0.13	1	*TFAP2E*	0.11
1	*AGO4*	0.21	5	*PPWD1*	0.17	3	*SNRK*	0.14	11	*PGA4*	0.13	11	*NTM*	0.11
19	*SPRED3*	0.21	3	*NEK10*	0.17	10	*CAMK2G*	0.14	1	*AGO1*	0.13	7	*SEMA3E*	0.11
10	*CHCHD1*	0.21	16	*ABCC12*	0.17	4	*PHOX2B*	0.14	17	*C17orf64*	0.13	9	*FBXW5*	0.11
1	*COL11A1*	0.21	16	*ZDHHC1*	0.17	11	*CD5*	0.14	11	*DDB1*	0.13	14	*NPAS3*	0.11
12	*CUX2*	0.21	7	*ITGB8*	0.17	1	*ABCA4*	0.14	2	*ARL6IP6*	0.13	3	*CLSTN2*	0.11
19	*KCNK6*	0.21	7	*GRM8*	0.17	16	*ENKD1*	0.14	4	*ARHGEF38*	0.13	14	*IGHG3*	0.11
19	*PSMD8*	0.21	15	*SNX22*	0.17	15	*CELF6*	0.14	1	*NME7*	0.13	7	*CPED1*	0.11
22	*CELSR1*	0.21	19	*ZNF780A*	0.17	10	*BMS1P4*	0.14	22	*MEI1*	0.13	17	*PRKCA*	0.11
7	*ARPC1B*	0.21	2	*PARD3B*	0.17	12	*SPX*	0.14	22	*APOL1*	0.13	2	*CERKL*	0.11
9	*SMC2*	0.21	2	*MOGS*	0.16	2	*FUCA1P1*	0.14	20	*DHX35*	0.13	3	*CNTN6*	0.11
4	*ETFDH*	0.20	3	*MRPS22*	0.16	17	*PPM1D*	0.14	2	*OTOF*	0.12	8	*TCF24*	0.11
7	*ARPC1A*	0.20	10	*CTNNA3*	0.16	14	*HIF1A*	0.14	1	*SLC35F3*	0.12	2	*TET3*	0.11

CHR	GENE NAME	SSI	CHR	GENE NAME	SSI	CHR	GENE NAME	SSI	CHR	GENE NAME	SSI	CHR	GENE NAME	SSI
6	*GRIK2*	0.11	2	*ERBB4*	0.10	1	*SCAMP3*	0.09	12	*CD63*	0.08	15	*DUOXA1*	0.08
7	*VWC2*	0.11	2	*PKP4*	0.10	15	*TRIP4*	0.09	8	*FABP4*	0.08	16	*RLTPR*	0.08
18	*CCDC178*	0.11	1	*FDPS*	0.10	10	*MYOZ1*	0.09	12	*IGFBP6*	0.08	16	*FTO*	0.08
3	*LMLN*	0.11	3	*GPX1*	0.10	18	*ZNF407*	0.09	1	*SLC25A34*	0.08	17	*HDAC5*	0.08
3	*MKRN2*	0.11	20	*ROMO1*	0.10	16	*SETD1A*	0.09	13	*TBC1D4*	0.08	1	*SMYD3*	0.08
4	*STK32B*	0.11	1	*PRKAR1AP*	0.10	12	*C12orf50*	0.09	7	*BBS9*	0.08	1	*FAM69A*	0.08
20	*SPAG4*	0.11	2	*ADAM17*	0.10	12	*CDK2AP1*	0.09	20	*SRMS*	0.08	14	*FANCM*	0.08
7	*EGFR*	0.11	16	*HSD17B2*	0.10	14	*RAD51B*	0.09	2	*KCNS3*	0.08	22	*PRR34*	0.08
12	*GTF2H3*	0.11	12	*KITLG*	0.10	1	*OR2L5*	0.09	7	*ZNRF2*	0.08	2	*DUXAP1*	0.08
16	*XYLT1*	0.11	1	*PDE4B*	0.10	14	*EFS*	0.09	1	*ACAP3*	0.08	7	*OR9A2*	0.08
7	*CTTNBP2*	0.11	4	*ADH7*	0.10	11	*DLG2*	0.09	2	*CNNM3*	0.08	8	*PTTG3P*	0.08
2	*ZRANB3*	0.11	20	*RALGAPB*	0.10	10	*TCF7L2*	0.09	17	*SHPK*	0.08	13	*ST6GALNAC4P1*	0.08
12	*GUCY2C*	0.11	1	*HCN3*	0.10	12	*SH2B3*	0.09	3	*SERPINI1*	0.08	16	*UBA52P8*	0.08
7	*TPK1*	0.11	12	*MPHOSPH9*	0.10	11	*CKAP5*	0.09	12	*ST8SIA1*	0.08	6	*PBX2*	0.08
15	*SV2B*	0.11	4	*TTC39CP1*	0.10	14	*RPL36AL*	0.09	9	*SH3GL2*	0.08			
22	*ZBED4*	0.11	22	*SGSM3*	0.10	2	*ANTXR1*	0.09	2	*IQCA1*	0.08			
13	*RB1*	0.11	2	*EMX1*	0.10	8	*VPS13B*	0.09	1	*OR2AK2*	0.08			
1	*KCNN3*	0.11	4	*BMP3*	0.10	2	*ACMSD*	0.09	6	*RNF5*	0.08			
20	*ID1*	0.11	3	*CLEC3B*	0.10	16	*EEF2K*	0.09	1	*TRAPPC3*	0.08			
16	*KCTD5*	0.11	20	*ACSS1*	0.10	18	*KLHL14*	0.09	10	*HPSE2*	0.08			
1	*OR2L8*	0.11	2	*FAP*	0.10	16	*CEMP1*	0.09	17	*NF1*	0.08			
17	*POLG2*	0.11	16	*ZNF668*	0.10	16	*RBFOX1*	0.09	12	*KCNH3*	0.08			
3	*FAM208A*	0.11	14	*DCAF5*	0.10	11	*TMEM138*	0.09	11	*OAF*	0.08			
20	*CEP250*	0.11	1	*SLC9C2*	0.10	11	*TMEM216*	0.09	3	*CAMKV*	0.08			
8	*RNF170*	0.11	1	*EFNA4*	0.10	7	*ZNF394*	0.09	2	*NEB*	0.08			
9	*INVS*	0.11	20	*SCAND1*	0.10	11	*MYRF*	0.09	15	*MTMR10*	0.08			
16	*CES4A*	0.11	12	*CEP290*	0.10	3	*GOLIM4*	0.09	16	*SHCBP1*	0.08			
2	*MAP3K19*	0.11	1	*RPAP2*	0.10	4	*PPID*	0.09	1	*KIAA1107*	0.08			
3	*SEMA5B*	0.11	1	*GFI1*	0.09	8	*NRG1*	0.09	17	*MPP2*	0.08			
2	*ABCA12*	0.11	9	*KANK1*	0.09	17	*SMURF2*	0.09	7	*CYP3A5*	0.08			
8	*SGK3*	0.11	7	*PRSS1*	0.09	4	*C4orf22*	0.09	12	*EPS8*	0.08			
10	*BMI1*	0.11	20	*COX4I2*	0.09	9	*TTLL11*	0.09	12	*NAA25*	0.08			
15	*ARNT2*	0.11	22	*TNRC6B*	0.09	12	*HMGA2*	0.09	19	*PSG3*	0.08			
5	*KCNIP1*	0.11	8	*PPP1R42*	0.09	7	*ZSCAN25*	0.09	4	*RAB28*	0.08			
20	*BCL2L1*	0.11	1	*PUSL1*	0.09	6	*KCNQ5*	0.09	1	*EFNA1*	0.08			
16	*KIAA0895L*	0.11	1	*PRDM16*	0.09	6	*GPR111*	0.09	17	*NT5C*	0.08			
12	*TSPAN9*	0.11	6	*COL11A2*	0.09	7	*GTF2I*	0.09	11	*SLC35C1*	0.08			
2	*FAM49A*	0.11	1	*ZBTB7B*	0.09	12	*HECTD4*	0.09	5	*SGTB*	0.08			
7	*SUGCT*	0.11	16	*BCKDK*	0.09	8	*PCMTD1*	0.09	2	*SSFA2*	0.08			
8	*PSD3*	0.11	16	*CTF1*	0.09	16	*ACD*	0.09	16	*FAM96B*	0.08			
16	*AGRP*	0.11	16	*PRSS53*	0.09	1	*ACKR1*	0.09	14	*MGAT2*	0.08			
9	*EHMT1*	0.11	8	*SGCZ*	0.09	1	*SLC50A1*	0.09	17	*EVI2B*	0.08			
11	*AHNAK*	0.11	15	*GANC*	0.09	2	*TLX2*	0.09	7	*OR6V1*	0.08			
14	*C14orf28*	0.11	16	*TERF2IP*	0.09	2	*NECAP1P2*	0.09	2	*DNAJC5G*	0.08			
2	*ERMN*	0.11	10	*HIF1AN*	0.09	13	*NPM1P22*	0.09	19	*HMG20B*	0.08			
10	*MSS51*	0.11	11	*TENM4*	0.09	1	*OR10J3*	0.09	4	*FNIP2*	0.08			
6	*CEP162*	0.10	8	*CHRNB3*	0.09	2	*OR7E28P*	0.09	3	*FAIM*	0.08			
8	*RP1L1*	0.10	15	*GABRB3*	0.09	18	*RPS3AP49*	0.09	2	*DFNB59*	0.08			
6	*UHRF1BP1*	0.10	20	*NFS1*	0.09	2	*TAF13P2*	0.09	20	*PTK6*	0.08			
15	*FBN1*	0.10	14	*RHOJ*	0.09	1	*DCST1*	0.09	6	*PARK2*	0.08			
8	*POTEA*	0.10	17	*RAB11FIP4*	0.09	12	*SLCO1B3*	0.09	14	*PLEK2*	0.08			
3	*ANO10*	0.10	4	*RXFP1*	0.09	1	*KCNT2*	0.08	22	*CCDC134*	0.08			
2	*ACKR3*	0.10	3	*KCNAB1*	0.09	11	*OR4P4*	0.08	3	*MST1R*	0.08			
1	*SCMH1*	0.10	8	*THAP1*	0.09	11	*OR8H3*	0.08	6	*AGER*	0.08			
22	*ADSL*	0.10	1	*LHX8*	0.09	17	*DHX58*	0.08	7	*ARF5*	0.08			
5	*KIF3A*	0.10	10	*SEC31B*	0.09	17	*GNA13*	0.08	7	*ING3*	0.08			
22	*MGAT3*	0.10	16	*ORAI3*	0.09	6	*NCOA7*	0.08	8	*CHRNA6*	0.08			
8	*PTK2*	0.10	2	*RAB17*	0.09	4	*PRKG2*	0.08	6	*ANKS1A*	0.08			
3	*ARHGEF3*	0.10	5	*SEPT8*	0.09	2	*PRKRA*	0.08	3	*DUSP7*	0.08			
8	*MCMDC2*	0.10	22	*MCHR1*	0.09	2	*ARL5A*	0.08	12	*HOXC13*	0.08			
9	*NXNL2*	0.10	22	*RPL3*	0.09	8	*ST18*	0.08	16	*GFOD2*	0.08			
6	*PDE7B*	0.10	4	*SORCS2*	0.09	5	*PPP2R2B*	0.08	4	*BOD1L1*	0.08			
17	*ADAP2*	0.10	12	*TMTC2*	0.09	1	*SHC1*	0.08	4	*SMARCA5*	0.08			
2	*DIRC1*	0.10	4	*TMPRSS11B*	0.09	16	*ITGAL*	0.08	9	*OLFML2A*	0.08			
2	*RAB3GAP1*	0.10	20	*FER1L4*	0.09	2	*KDM3A*	0.08	15	*CTDSPL2*	0.08			
6	*AHI1*	0.10	12	*TCTN2*	0.09	22	*TTC28*	0.08	7	*ZNF655*	0.08			
2	*LRPPRC*	0.10	20	*RBM39*	0.09	12	*SBNO1*	0.08	22	*CACNA1I*	0.08			
2	*KIF3C*	0.10	3	*PPARG*	0.09	1	*MAGI3*	0.08	16	*CDH16*	0.08			
16	*CENPT*	0.10	8	*HOOK3*	0.09	15	*CAPN3*	0.08	14	*DNAAF2*	0.08			
15	*FAM96A*	0.10	5	*FBN2*	0.09	1	*RPL5*	0.08	17	*ATP5H*	0.08			
8	*TRPS1*	0.10	7	*ZNF789*	0.09	14	*AHSA1*	0.08	16	*RRAD*	0.08			
20	*ERGIC3*	0.10	17	*TRPV1*	0.09	11	*DAK*	0.08	21	*N6AMT1*	0.08			
1	*NOTCH2*	0.10	6	*TBC1D7*	0.09	22	*PDGFB*	0.08	19	*SIPA1L3*	0.08			
2	*ITGB1BP1*	0.10	3	*EIF4E3*	0.09	6	*EZR*	0.08	22	*TEF*	0.08			
16	*KARS*	0.10	13	*ZMYM5*	0.09	8	*FABP9*	0.08	22	*PMM1*	0.08			
12	*ANO2*	0.10	11	*F2*	0.09	11	*SCGB1D1*	0.08	12	*SLC6A15*	0.08			
1	*ANKRD45*	0.10	2	*NAT8*	0.09	12	*EIF2B1*	0.08	19	*MAST1*	0.08			
2	*CACNB4*	0.10	16	*PARD6A*	0.09	1	*CPSF3L*	0.08	9	*EDF1*	0.08			
1	*LRRC7*	0.10	1	*EFNA3*	0.09	22	*TOB2*	0.08	6	*GPSM3*	0.08			
5	*NIM1K*	0.10	12	*FAM109A*	0.09	15	*DUOXA2*	0.08	22	*ST13*	0.08			
1	*RABGGTB*	0.10	2	*SULT1C2P1*	0.09	5	*IL13*	0.08	15	*CYP19A1*	0.08			
20	*RBM12*	0.10	6	*TRDN*	0.09	4	*GAB1*	0.08	2	*APOB*	0.08			
19	*ARHGEF1*	0.10	2	*CCNT2*	0.09	4	*GPM6A*	0.08	20	*GDF5*	0.08			
3	*DGKG*	0.10	22	*ALG12*	0.09	10	*SMNDC1*	0.08	3	*FAM212A*	0.08			
16	*NUDT16L1*	0.10	2	*SULT1C3*	0.09	3	*ADAMTS9*	0.08	1	*KLHDC9*	0.08			
1	*SSU72*	0.10	5	*CDH12*	0.09	2	*GCG*	0.08	11	*OR56B4*	0.08			
10	*SORCS3*	0.10	9	*TRAF2*	0.09	1	*GLMN*	0.08	2	*ANKRD23*	0.08			
16	*MPHOSPH6*	0.10	2	*CCDC150*	0.09	1	*NFASC*	0.08	10	*CAMK1D*	0.08			
19	*RASGRP4*	0.10	13	*PRR20A*	0.09	10	*KIAA1217*	0.08	1	*MEX3A*	0.08			
15	*KIAA0101*	0.10	20	*CDH4*	0.09	7	*TTC26*	0.08	1	*RAB25*	0.08			
2	*SEMA4F*	0.10	8	*DNAJC5B*	0.09	11	*LRP4*	0.08	20	*CSTF1*	0.08			
16	*ZNF646*	0.10	11	*TMEM258*	0.09	11	*CYB561A3*	0.08	12	*OLR1*	0.08			
17	*NUP85*	0.10	7	*KMT2C*	0.09	12	*C12orf65*	0.08	8	*FAM167A*	0.08			
12	*MAPKAPK5*	0.10	17	*CA4*	0.09	16	*TSNAXIP1*	0.08	19	*ZNF14*	0.08			
16	*NUTF2*	0.10	1	*CFH*	0.09	16	*BFAR*	0.08	11	*KCNQ1*	0.08			
1	*DPM3*	0.10	12	*ATP6V0A2*	0.09	12	*KCNJ8*	0.08	2	*ACVR1*	0.08			
1	*OR2L3*	0.10	1	*HS2ST1*	0.09	1	*OR2T8*	0.08	2	*TMEM163*	0.08			
20	*PHF20*	0.10	1	*C1orf216*	0.09	16	*C16orf87*	0.08	17	*CALCOCO2*	0.08			
2	*COPS8*	0.10	15	*PIF1*	0.09	1	*SLFNL1*	0.08	20	*XKR7*	0.08			
22	*TRMU*	0.10	16	*PYCARD*	0.09	7	*GRM3*	0.08	1	*CFHR1*	0.08			

Appendix B

Lists of top *FineMAV* candidates

Top 100 *FineMAV* hits in Africans (AFR), East Asians (EAS), Europeans (EUR), Eurasians (EAS+EUR) and non-admixed Native Americans (AMR) are given in the following pages.

SNP	– Single Nucleotide Polymorphism ID
CHR	– chromosome
POS	– genomic position
DER_ALLELE	– derived allele
DAF	– population specific Derived Allele Frequency
DAF_GLOB	– Global Derived Allele Frequency (average across populations)
DAP	– Derived Allele Purity
CADD	– Combined Annotation-Dependent Depletion of derived allele
FineMAV	– population specific Fine-Mapping of Adaptive Variation
CONSEQUENCE	– most severe consequence according to ENSEMBL (NC stands for non-coding; RNA stands for different types of non-coding RNA including lincRNA, snRNA and miRNA)
GENE	– associated gene name

SNP	CHR	POS	DER_ALLELE	DAF_AFR	DAF_EAS	DAF_EUR	DAF_GLOB	DAP	CADD	FineMAV_AFR	CONSEQUENCE	GENE
rs2814778	1	159174683	C	0.96	0.00	0.01	0.32	0.98	17.74	16.73	5 prime UTR	*ACKR1*
rs7645635	3	188327555	C	0.71	0.00	0.01	0.24	0.98	23.50	16.34	Missense	*LPP*
rs8052655	16	67409180	A	0.64	0.00	0.04	0.23	0.79	32.00	16.28	Missense; Regulatory	*LRRC36*
rs6601495	8	10480377	C	0.77	0.00	0.01	0.26	0.97	21.00	15.73	Missense	*RP1L1*
rs35999740	19	22116015	T	0.48	0.00	0.01	0.16	0.92	34.00	14.94	Stop	*ZNF208*
rs13043	3	138347988	C	0.63	0.00	0.01	0.21	0.93	23.10	13.38	Missense	*FAIM*
rs1871534	8	145639681	C	0.91	0.00	0.01	0.31	0.97	14.92	13.29	Missense	*SLC39A4*
rs10743455	12	23269297	C	0.57	0.00	0.00	0.19	1.00	21.20	12.16	RNA	
rs10955754	8	116673229	A	0.62	0.00	0.01	0.21	0.97	19.83	12.05	Intron	*TRPS1*
rs12334173	7	107698612	T	0.64	0.00	0.01	0.22	0.90	20.20	11.77	Intron	*LAMB4*
rs7073075	10	108751261	A	0.68	0.00	0.02	0.24	0.89	19.33	11.77	Intron	*SORCS1*
rs7749306	6	10687746	T	0.78	0.04	0.02	0.28	0.76	19.73	11.75	Missense	*C6orf52*
rs2871874	2	121696149	G	0.59	0.00	0.01	0.20	0.92	21.70	11.75	Regulatory	*GLI2*
rs12386710	7	79665088	G	0.57	0.00	0.00	0.19	0.98	20.40	11.45	Intergenic	
rs73129583	5	91226210	A	0.60	0.00	0.00	0.20	0.98	19.35	11.43	Intergenic	
rs4246244	12	23255870	G	0.57	0.00	0.00	0.19	1.00	19.68	11.22	RNA	
rs7860464	9	710966	A	0.55	0.00	0.02	0.19	0.89	22.60	11.13	Missense	*KANK1*
rs7091076	10	108751262	G	0.68	0.00	0.02	0.24	0.89	18.23	11.10	Intron	*SORCS1*
rs11993782	8	145004042	A	0.83	0.00	0.01	0.28	0.97	13.85	11.08	Intron	*PLEC*
rs11989050	8	116781936	C	0.70	0.00	0.00	0.24	0.97	16.29	11.08	Intron	*TRPS1*
rs7840291	8	116762150	A	0.71	0.00	0.00	0.24	0.97	16.00	10.99	Intron	*TRPS1*
rs10016143	4	149615561	G	0.72	0.00	0.00	0.24	0.98	15.46	10.92	RNA	
rs1185511	17	3627619	T	0.59	0.00	0.00	0.20	1.00	18.44	10.82	Regulatory	*GSG2; ITGAE*
rs16887761	8	116781627	C	0.70	0.00	0.00	0.24	0.97	15.90	10.81	Intron	*TRPS1*
rs11860295	16	67316234	T	0.69	0.00	0.07	0.26	0.70	22.20	10.79	Missense	*PLEKHG4*
rs1346265	15	45695612	T	0.53	0.01	0.00	0.18	0.94	21.80	10.78	Regulatory	*SPATA5L1; GATM*
rs5962502	X	104535693	C	0.89	0.05	0.06	0.33	0.68	17.77	10.73	Intron	*IL1RAPL2*
rs2467852	15	45695341	A	0.53	0.01	0.00	0.18	0.94	21.50	10.65	Regulatory	*SPATA5L1; GATM*
rs11121679	1	11072691	A	0.49	0.00	0.00	0.16	0.99	21.90	10.58	Regulatory; 5 prime UTR	*TARDBP*
rs17005371	2	26687761	A	0.54	0.00	0.01	0.18	0.91	21.60	10.57	Missense	*OTOF*
rs10093431	8	54899861	T	0.61	0.00	0.02	0.21	0.87	19.97	10.55	Regulatory	*TCEA1*
rs73641908	9	14339646	C	0.57	0.00	0.00	0.19	0.99	18.39	10.48	Intron	*NFIB*
rs6830772	4	110475069	C	0.56	0.00	0.00	0.19	0.99	18.68	10.46	NC transcript exon	*SETP20*
rs67954183	X	17610532	T	0.59	0.01	0.01	0.20	0.89	19.66	10.33	Regulatory	*NHS*
rs11989722	8	116662069	T	0.63	0.00	0.01	0.21	0.97	16.68	10.28	Regulatory	*TRPS1*
rs6468210	8	33858833	G	0.52	0.00	0.01	0.18	0.95	20.70	10.24	RNA	
rs57609710	1	173560754	G	0.65	0.00	0.01	0.22	0.97	16.00	10.16	Intron	*SLC9C2*
rs181107182	X	39964323	G	0.59	0.00	0.04	0.21	0.80	21.50	10.12	Regulatory	*BCOR*
rs73636231	X	104706319	A	0.69	0.00	0.00	0.23	0.98	14.86	10.10	Intron	*IL1RAPL2*
rs73695330	8	91295266	T	0.50	0.01	0.00	0.17	0.95	21.20	10.08	RNA	*LINC00534*
rs11537667	16	70303659	T	0.30	0.00	0.00	0.10	0.98	34.00	10.05	Missense	*AARS*
rs10044990	5	107141997	A	0.50	0.00	0.00	0.17	0.97	20.40	9.98	Regulatory	*RN7SKP122*
rs55844314	3	46036264	G	0.64	0.04	0.03	0.24	0.70	22.20	9.96	Regulatory	*FYCO1*
rs6759356	2	158140806	C	0.50	0.00	0.02	0.17	0.87	22.80	9.95	Missense	*GALNT5*
rs1427463	17	62492582	T	0.80	0.03	0.12	0.32	0.56	22.00	9.93	Missense; Regulatory	*POLG2*
rs692703	15	41408736	G	0.46	0.00	0.01	0.16	0.96	22.30	9.93	Regulatory	*INO80*
rs9460973	6	24175021	A	0.49	0.00	0.02	0.17	0.84	23.80	9.92	Missense	*DCDC2*
rs698258	2	176988412	T	0.54	0.00	0.02	0.19	0.86	21.40	9.90	Intron	*HOXD9; HOXD10*
rs857418	6	14656031	C	0.54	0.00	0.01	0.19	0.91	19.96	9.86	Regulatory	
rs7406656	17	58832257	C	0.81	0.02	0.11	0.31	0.61	20.10	9.85	Intron	*BCAS3*
rs7868958	9	110905201	G	0.46	0.00	0.00	0.15	0.98	22.10	9.84	Regulatory	
rs10912654	1	173601630	T	0.70	0.00	0.01	0.23	0.98	14.36	9.80	Intron	*ANKRD45*
rs6048066	20	2291722	C	0.43	0.00	0.00	0.14	1.00	22.70	9.77	Missense	*TGM3*
rs2122635	7	131830328	A	0.87	0.06	0.07	0.33	0.62	17.96	9.73	Regulatory	*PLXNA4*
rs8034559	15	95104562	T	0.49	0.00	0.01	0.17	0.97	20.40	9.70	Intergenic	
rs10093311	8	54899788	G	0.60	0.00	0.02	0.21	0.87	18.54	9.69	Regulatory	*TCEA1*
rs7072384	10	108750921	C	0.68	0.00	0.02	0.24	0.89	15.86	9.68	Intron	*SORCS1*
rs7723522	5	100962433	T	0.71	0.03	0.03	0.26	0.75	18.02	9.67	Intergenic	
rs10111177	8	33843454	T	0.52	0.00	0.01	0.17	0.95	19.58	9.67	RNA	
rs2408536	21	24144784	A	0.47	0.00	0.00	0.16	0.99	20.70	9.56	Intergenic	
rs10503956	8	33941001	A	0.47	0.00	0.00	0.16	0.97	20.80	9.55	Intergenic	
rs521983	13	58636124	T	0.45	0.00	0.00	0.15	0.98	21.40	9.51	Intergenic	
rs28711384	2	164867367	C	0.61	0.00	0.01	0.20	0.96	16.27	9.50	RNA	
rs34155925	4	76788564	T	0.45	0.01	0.00	0.15	0.93	22.60	9.49	Missense	*PPEF2*
rs7705335	5	72743925	C	0.64	0.00	0.00	0.22	0.98	14.98	9.47	Missense; Regulatory	*FOXD1*
rs331537	11	4471276	A	0.72	0.03	0.03	0.26	0.74	17.78	9.46	Missense	*OR52K2*
rs17105909	14	37473678	A	0.56	0.00	0.00	0.19	0.99	16.97	9.42	Intron	*SLC25A21*
rs5965014	X	64988114	T	0.71	0.00	0.00	0.24	0.99	13.36	9.40	Downstream gene	*NANOGP9*
rs1657220	10	34816233	T	0.62	0.00	0.05	0.22	0.76	19.96	9.40	Regulatory	*PARD3*
rs7103594	11	78320673	A	0.63	0.00	0.01	0.21	0.97	15.34	9.38	Intergenic	
rs983763	4	185981720	G	0.72	0.00	0.00	0.24	0.99	13.26	9.35	RNA	
rs733495	8	72472156	C	0.47	0.00	0.00	0.16	1.00	19.99	9.35	RNA	
rs16945912	15	63133002	C	0.57	0.01	0.02	0.20	0.85	19.16	9.34	3 prime UTR	*TLN2*
rs6982495	8	34790660	T	0.55	0.00	0.02	0.19	0.90	18.87	9.33	Intergenic	
rs6883147	5	68176952	C	0.53	0.00	0.00	0.18	0.97	18.12	9.32	RNA	
rs7090581	10	26504928	A	0.48	0.00	0.00	0.16	0.99	19.86	9.31	Regulatory	*GAD2; MYO3A*
rs7095872	10	112435986	C	0.57	0.00	0.00	0.19	0.98	16.52	9.30	Regulatory	*RBM20*
rs6596137	5	133019116	C	0.50	0.00	0.02	0.17	0.88	21.10	9.28	Intergenic	
rs2610724	6	89383001	A	0.52	0.00	0.00	0.18	0.95	18.62	9.25	Intron	*RNGTT*
rs1128308	14	105954682	G	0.60	0.00	0.02	0.21	0.86	18.10	9.25	Regulatory	*CRIP1; C14orf80*
rs6659237	1	147026125	G	0.67	0.00	0.04	0.24	0.81	17.10	9.24	Intron	*BCL9*
rs1464836	7	108478693	A	0.52	0.00	0.01	0.17	0.95	18.69	9.23	Intergenic	
rs60081219	X	53710797	A	0.67	0.02	0.01	0.23	0.85	16.23	9.22	Regulatory	*HUWE1*
rs1253835	13	106222388	G	0.56	0.01	0.02	0.20	0.85	19.26	9.22	Intergenic	
rs7106654	11	47203983	A	0.36	0.00	0.01	0.12	0.95	27.10	9.20	Missense	*PACSIN3*
rs1871148	3	71102544	C	0.48	0.00	0.00	0.16	0.99	19.14	9.19	Intron	*FOXP1*
rs201113	10	11048264	C	0.79	0.00	0.10	0.30	0.65	17.79	9.13	Regulatory	*CELF2*
rs7268805	20	11198959	G	0.59	0.00	0.02	0.21	0.85	18.27	9.11	Regulatory	
rs16984889	X	105168689	A	0.44	0.00	0.00	0.15	0.99	20.80	9.11	Missense; Splice	*NRK*
rs6888459	5	72741402	A	0.62	0.00	0.00	0.21	0.98	14.87	9.09	Intron	*FOXD1*
rs59236623	11	114013734	C	0.50	0.00	0.00	0.17	0.99	18.38	9.06	Intron	*ZBTB16*
rs57085058	X	48910933	T	0.63	0.00	0.00	0.21	0.99	14.54	9.05	Regulatory	*CCDC120*
rs16984014	X	102979184	A	0.26	0.00	0.00	0.09	1.00	34.00	9.00	Missense	*GLRA4*
rs5469	16	72088418	A	0.43	0.00	0.00	0.14	0.98	21.10	8.98	Regulatory	*TXNL4B; HP*
rs7771665	6	144539961	C	0.62	0.00	0.01	0.21	0.92	15.76	8.98	Intergenic	
rs4876614	8	116821225	T	0.84	0.07	0.02	0.31	0.70	15.31	8.98	Intron	*TRPS1*
rs7935192	11	47004492	A	0.59	0.00	0.00	0.20	0.98	15.52	8.97	Intron	*C11orf49*
rs5971168	X	28334054	G	0.74	0.01	0.03	0.26	0.85	14.26	8.97	Intergenic	
rs7547313	1	158632531	C	0.37	0.00	0.00	0.12	0.97	24.80	8.94	Missense	*SPTA1*
rs17118461	12	56687805	T	0.59	0.00	0.00	0.20	0.99	15.32	8.93	Intron	*CS*

SNP	CHR	POS	DER_ALLELE	DAF_AFR	DAF_EAS	DAF_EUR	DAF_GLOB	DAP	CADD	FineMAV_EAS	CONSEQUENCE	GENE
rs1343879	X	75004529	A	0.03	0.83	0.03	0.30	0.79	35.00	23.01	Stop	*MAGEE2*
rs2293766	7	100371358	A	0.02	0.51	0.01	0.18	0.82	47.00	19.48	Stop	*ZAN*
rs3827760	2	109513601	G	0.00	0.87	0.01	0.30	0.95	22.30	18.42	Missense	*EDAR*
rs1800414	15	28197037	C	0.00	0.60	0.00	0.20	1.00	24.10	14.31	Missense; Splice	*OCA2*
rs72617608	Y	2863171	A	0.00	0.82	0.00	0.27	1.00	17.31	14.12	NC transcript exon	*EEF1A1P41*
rs11150606	16	31099011	C	0.01	0.76	0.03	0.27	0.84	21.30	13.66	Missense	*PRSS53*
rs3732240	2	71366957	A	0.00	0.43	0.02	0.15	0.87	32.00	12.12	Missense	*MPHOSPH10*
rs17027638	3	3197918	G	0.00	0.70	0.02	0.24	0.88	19.75	12.07	Synonymous	*CRBN*
rs34537429	2	26296089	G	0.00	0.86	0.02	0.30	0.90	15.53	12.00	Regulatory	*RAB10*
rs72627476	2	109451118	G	0.00	0.86	0.01	0.29	0.95	14.61	11.91	Intron	*CCDC138*
rs12678022	8	73837442	A	0.01	0.78	0.04	0.28	0.82	18.45	11.76	Intron	*KCNB2*
rs17822931	16	48258198	T	0.01	0.78	0.14	0.31	0.54	26.50	11.26	Missense	*ABCC11*
rs72630048	X	71715133	A	0.04	0.73	0.01	0.26	0.79	19.04	10.97	Regulatory	*HDAC8*
rs11217799	11	120164954	C	0.01	0.65	0.00	0.22	0.96	17.49	10.96	Intron	*POU2F3*
rs4411467	15	93588043	A	0.03	0.68	0.04	0.25	0.73	22.00	10.96	3 prime UTR	*RGMA*
rs3736508	11	45975130	T	0.00	0.56	0.03	0.20	0.83	23.60	10.92	Missense	*PHF21A*
rs9360004	6	93673482	A	0.02	0.71	0.01	0.25	0.88	17.46	10.88	Intergenic	
rs4688744	3	50249500	A	0.00	0.62	0.01	0.21	0.93	18.13	10.43	Regulatory	*SLC38A3*
rs12921053	16	82867623	A	0.06	0.76	0.02	0.28	0.70	19.41	10.42	Regulatory	*CDH13*
rs35685348	15	64759279	T	0.04	0.79	0.09	0.31	0.59	21.80	10.25	Intron	*ZNF609*
rs1047781	19	49206631	T	0.00	0.44	0.00	0.15	1.00	23.20	10.20	Missense	*FUT2*
rs74595980	3	50198840	C	0.00	0.63	0.01	0.21	0.92	17.44	10.15	Intron	*SEMA3F*
rs58137261	3	50187637	C	0.00	0.63	0.01	0.21	0.92	17.49	10.11	Regulatory	*SEMA3F*
rs2072053	3	50197092	T	0.00	0.63	0.01	0.21	0.92	17.47	10.10	Synonymous	*SEMA3F*
rs6609051	X	39938225	C	0.00	0.64	0.03	0.23	0.83	18.81	10.01	Intron	*BCOR*
rs16986619	2	19740980	C	0.02	0.70	0.07	0.26	0.65	22.00	9.95	Regulatory	
rs72612108	8	32306158	A	0.01	0.61	0.02	0.21	0.84	19.25	9.91	Regulatory	*NRG1*
rs2240227	19	15852242	A	0.01	0.61	0.05	0.22	0.72	22.30	9.85	Missense	*OR10H3*
rs2282440	1	31347320	A	0.01	0.52	0.01	0.18	0.83	22.50	9.79	Missense	*SDC3*
rs10496971	2	145769943	G	0.07	0.79	0.07	0.31	0.57	21.10	9.58	Regulatory; RNA	*TEX41*
rs2242406	16	75574030	T	0.01	0.70	0.04	0.25	0.78	17.33	9.56	Synonymous	*TMEM231*
rs936212	15	40581543	C	0.00	0.56	0.01	0.19	0.91	18.69	9.53	Missense	*PLCB2*
rs289817	15	63795675	A	0.08	0.91	0.13	0.37	0.48	22.00	9.49	Regulatory	*USP3*
rs245821	5	142403197	A	0.00	0.43	0.00	0.14	1.00	21.70	9.43	Regulatory	*ARHGAP26*
rs61687236	X	74860977	A	0.03	0.83	0.03	0.29	0.79	14.34	9.40	Intergenic	
rs12607385	18	47777937	A	0.00	0.43	0.01	0.14	0.93	23.70	9.36	Missense	*CFAP53*
rs17143187	7	20666235	C	0.01	0.45	0.01	0.16	0.85	24.10	9.26	Splice	*ABCB5*
rs79525262	17	48639367	A	0.01	0.82	0.12	0.31	0.61	18.46	9.22	Intron	*CACNA1G*
rs10846	19	14582468	A	0.04	0.82	0.09	0.32	0.60	18.51	9.14	Missense	*PKN1; PTGER1*
rs12231744	12	112477055	C	0.05	0.57	0.00	0.21	0.72	22.10	9.10	Missense	*NAA25*
rs72612111	8	32345444	A	0.00	0.57	0.02	0.20	0.84	18.89	9.02	Regulatory	*NRG1*
rs3843699	15	64637091	T	0.01	0.76	0.08	0.28	0.69	17.18	9.01	Regulatory	*CSNK1G1*
rs10497377	2	172926226	T	0.01	0.44	0.01	0.15	0.88	23.40	9.00	Missense; Splice	*METAP1D*
rs12036697	1	152878909	G	0.00	0.52	0.00	0.18	0.98	17.67	8.99	Regulatory	*IVL*
rs2074000	7	20685484	A	0.00	0.45	0.01	0.15	0.90	22.20	8.97	Missense	*ABCB5*
rs2224442	14	97272382	G	0.12	0.83	0.05	0.34	0.51	21.10	8.93	Regulatory	*VRK1*
rs77017835	7	100377364	A	0.03	0.51	0.01	0.18	0.76	22.90	8.89	Missense	*ZAN*
rs2236943	3	50278613	A	0.00	0.62	0.01	0.21	0.93	15.31	8.81	Intron	*GNAI2*
rs61186727	3	149801650	A	0.00	0.43	0.00	0.14	0.97	21.10	8.81	Regulatory	
rs77794375	7	100285476	T	0.02	0.50	0.01	0.18	0.79	22.40	8.77	Missense	*GIGYF1*
rs58361699	9	14241650	C	0.00	0.51	0.01	0.17	0.92	18.57	8.72	Regulatory	*NFIB*
rs2897762	X	12212919	G	0.00	0.62	0.04	0.22	0.77	18.19	8.70	Intron	*FRMPD4*
rs16938528	8	73866368	A	0.01	0.78	0.04	0.28	0.80	13.90	8.66	RNA	
rs11217775	11	120121323	C	0.02	0.62	0.03	0.22	0.78	17.99	8.62	Intron	*POU2F3*
rs1017570	12	54245951	T	0.04	0.64	0.05	0.24	0.64	21.00	8.61	Intergenic	
rs727250	X	71645222	G	0.05	0.73	0.01	0.26	0.74	15.93	8.60	Intron	*HDAC8*
rs79117849	3	50370841	G	0.00	0.62	0.01	0.21	0.93	14.85	8.57	Intron	*RASSF1*
rs6123230	20	51566360	T	0.01	0.55	0.04	0.20	0.75	20.90	8.57	Intergenic	
rs2235371	1	209964080	T	0.00	0.43	0.01	0.15	0.88	22.50	8.54	Missense; Regulatory	*IRF6*
rs2292114	15	65876969	A	0.00	0.43	0.01	0.15	0.93	21.00	8.51	Intron	*VWA9*
rs12638212	3	108159977	A	0.04	0.56	0.04	0.21	0.62	24.40	8.48	Missense; Regulatory	*MYH15*
rs12643176	4	38253288	C	0.00	0.72	0.03	0.25	0.86	13.77	8.48	Intergenic	
rs7711270	5	141870940	C	0.00	0.57	0.03	0.20	0.81	18.25	8.47	Regulatory	
rs80166022	7	83487913	G	0.01	0.61	0.01	0.21	0.90	15.24	8.44	Intergenic	
rs58402910	11	131279607	T	0.05	0.62	0.01	0.23	0.74	18.06	8.35	Intron	*NTM*
rs12622852	2	17409594	G	0.00	0.77	0.08	0.28	0.70	15.46	8.31	Regulatory	
rs10433910	4	80852826	C	0.04	0.66	0.05	0.25	0.62	20.30	8.30	Intron	*ANTXR2*
rs6612369	X	55829607	A	0.01	0.62	0.00	0.21	0.95	14.00	8.27	Intergenic	
rs3795786	1	228469903	T	0.02	0.29	0.01	0.11	0.76	37.00	8.26	Stop	*OBSCN*
rs4912861	5	141888970	C	0.00	0.53	0.05	0.19	0.74	20.90	8.23	Regulatory	
rs9384327	6	155752362	C	0.01	0.69	0.07	0.26	0.68	17.59	8.19	Intron	*NOX3*
rs12585165	13	74150425	T	0.03	0.60	0.01	0.21	0.80	17.08	8.18	RNA	*LINC00392*
rs9634714	13	34908238	A	0.01	0.78	0.05	0.28	0.78	13.39	8.16	Intergenic	
rs11014573	10	25758874	G	0.01	0.56	0.00	0.19	0.94	15.45	8.13	Intron	*GPR158*
rs12688902	X	71695388	G	0.04	0.71	0.01	0.25	0.78	14.53	8.10	Regulatory	*HDAC8*
rs2922849	8	6456740	T	0.00	0.46	0.02	0.16	0.85	20.80	8.09	Regulatory	*MCPH1*
rs12492683	3	50351598	T	0.00	0.62	0.01	0.21	0.93	14.00	8.08	Upstream gene	*HYAL1; HYAL2*
rs72630058	X	71880808	T	0.04	0.68	0.01	0.24	0.79	15.01	8.07	Intron	*PHKA1*
rs965305	15	98271950	T	0.01	0.56	0.05	0.21	0.69	20.70	8.02	RNA	*LINC00923*
rs2140929	10	80304118	A	0.00	0.45	0.00	0.15	0.97	18.21	7.99	Regulatory; RNA	*LINC00856*
rs12622360	2	60610224	A	0.00	0.61	0.06	0.23	0.70	18.82	7.99	RNA	*RNU1-32P; MIR4432*
rs3824883	11	120172688	T	0.00	0.60	0.00	0.20	0.97	13.61	7.99	NC transcript exon	*POU2F3*
rs4688683	3	50239464	T	0.00	0.61	0.01	0.21	0.93	13.88	7.95	Regulatory	*SLC38A3*
rs4935502	10	55955444	G	0.13	0.84	0.13	0.37	0.38	24.50	7.91	Missense; Splice	*PCDH15*
rs60041373	7	109419462	C	0.00	0.40	0.00	0.14	0.96	20.50	7.91	Intergenic	
rs9363019	6	93628360	G	0.02	0.73	0.13	0.29	0.53	20.50	7.89	Intergenic	
rs17027678	2	41647254	G	0.00	0.56	0.03	0.19	0.84	16.93	7.87	Intergenic	
rs17853471	18	70209177	T	0.00	0.46	0.01	0.16	0.91	18.62	7.86	Regulatory	*CBLN2*
rs12684795	9	9817047	C	0.07	0.71	0.06	0.28	0.57	19.23	7.84	Intron	*PTPRD*
rs75241782	12	112280300	T	0.05	0.58	0.00	0.21	0.72	18.71	7.76	Regulatory	*MAPKAPK5*
rs3905317	2	135744007	C	0.01	0.38	0.00	0.13	0.87	23.30	7.75	Missense	*MAP3K19*
rs17130188	10	114868162	C	0.02	0.50	0.01	0.18	0.81	19.45	7.75	Regulatory	*TCF7L2*
rs59111569	2	25956419	G	0.00	0.84	0.04	0.29	0.85	10.86	7.75	Downstream gene	*ASXL2*
rs7309681	12	112337874	T	0.05	0.57	0.00	0.21	0.72	18.68	7.73	3 prime UTR	*MAPKAPK5; ADAM1A*
rs2303893	2	26507076	T	0.02	0.74	0.07	0.28	0.67	15.69	7.72	Splice; Regulatory	*HADHB*
rs1969589	15	93588030	T	0.03	0.65	0.03	0.24	0.74	15.85	7.69	3 prime UTR	*RGMA*
rs12575998	11	131279672	T	0.06	0.62	0.01	0.23	0.72	17.04	7.60	Intron	*NTM*
rs76145088	1	209989232	G	0.00	0.43	0.01	0.15	0.88	20.20	7.60	Regulatory	
rs4145606	2	19754838	G	0.02	0.69	0.08	0.26	0.64	17.14	7.58	Intergenic	
rs10460537	2	47283824	C	0.00	0.35	0.00	0.12	0.99	21.60	7.57	Regulatory; RNA	*TTC7A; C2orf61*

SNP	CHR	POS	DER_ALLELE	DAF_AFR	DAF_EAS	DAF_EUR	DAF_GLOB	DAP	CADD	FineMAV_EUR	CONSEQUENCE	GENE
rs16891982	5	33951693	G	0.04	0.01	0.94	0.33	0.86	26.30	21.21	Missense	*SLC45A2*
rs1426654	15	48426484	A	0.07	0.01	1.00	0.36	0.75	18.75	13.99	Missense	*SLC24A5*
rs10962600	9	16795783	T	0.02	0.00	0.73	0.25	0.89	18.33	11.92	Regulatory	*BNC2*
rs12939056	17	7754993	A	0.01	0.00	0.58	0.20	0.92	22.00	11.65	Regulatory; Synonymous	*KDM6B; TMEM88*
rs12142199	1	1249187	A	0.07	0.02	0.81	0.30	0.71	19.45	11.07	Missense	*CPSF3L*
rs12881545	14	101176212	C	0.03	0.01	0.65	0.23	0.83	19.95	10.78	Regulatory	
rs11073964	15	91543761	C	0.04	0.00	0.60	0.21	0.78	22.40	10.56	Missense	*VPS33B*
rs6705778	2	104490395	G	0.04	0.00	0.71	0.25	0.81	18.05	10.36	RNA	
rs1472932	17	19220666	A	0.06	0.00	0.81	0.29	0.76	16.51	10.17	Regulatory	*EPN2*
rs4851673	2	104817402	G	0.08	0.00	0.74	0.28	0.69	19.88	10.14	Intergenic	
rs7580	X	71493691	T	0.10	0.01	0.75	0.29	0.63	21.30	10.09	Regulatory; Synonymous	*RPS4X; PIN4*
rs12350739	9	16885017	A	0.02	0.01	0.56	0.20	0.87	20.40	10.01	Regulatory	
rs1446585	2	136407479	A	0.03	0.00	0.54	0.19	0.84	21.70	9.84	Missense	*R3HDM1*
rs34938541	2	104830710	T	0.09	0.00	0.76	0.29	0.66	18.73	9.46	Intergenic	
rs17261772	2	135911422	T	0.03	0.00	0.59	0.21	0.84	19.07	9.40	Missense	*RAB3GAP1*
rs10756789	9	16720122	G	0.02	0.00	0.63	0.22	0.89	16.53	9.30	Regulatory	*BNC2*
rs12128213	1	200197538	G	0.02	0.01	0.51	0.18	0.82	21.60	9.11	Regulatory	
rs1355128	17	19299144	A	0.04	0.00	0.78	0.27	0.84	13.93	9.10	Intergenic	
rs13037496	20	31466534	T	0.05	0.01	0.48	0.18	0.70	26.50	8.99	Missense	*EFCAB8*
rs35517488	7	113362256	T	0.03	0.01	0.51	0.18	0.80	21.90	8.97	Intergenic	
rs28591622	17	19175317	A	0.06	0.00	0.81	0.29	0.76	14.24	8.77	Intron	*EPN2*
rs7637449	3	56628031	A	0.07	0.05	0.54	0.22	0.49	33.00	8.68	Missense	*CCDC66*
rs4508618	2	104763415	G	0.06	0.00	0.74	0.27	0.74	15.80	8.65	RNA	
rs2026805	9	16800789	A	0.02	0.00	0.72	0.25	0.89	13.39	8.59	Intron	*BNC2*
rs13027794	2	216577817	T	0.09	0.00	0.70	0.26	0.66	18.55	8.54	Regulatory; RNA	*LINC00607*
rs2842895	6	7106316	C	0.03	0.01	0.54	0.19	0.81	19.37	8.54	Upstream gene	*RREB1*
rs35135256	2	104788566	G	0.08	0.00	0.74	0.28	0.68	16.79	8.52	Intergenic	
rs10756812	9	16786303	A	0.02	0.00	0.69	0.24	0.89	13.69	8.50	Intron	*BNC2*
rs12913832	15	28365618	G	0.03	0.00	0.64	0.22	0.85	15.66	8.48	Intron	*HERC2*
rs1042602	11	88911696	A	0.01	0.00	0.37	0.13	0.89	25.40	8.37	Missense	*TYR*
rs12345910	9	16802118	C	0.02	0.00	0.72	0.25	0.89	12.71	8.23	Regulatory	*BNC2*
rs17675094	16	82949088	C	0.02	0.03	0.60	0.22	0.73	18.41	8.08	Intron	*CDH13*
rs11788101	9	16828961	T	0.02	0.00	0.55	0.19	0.89	15.94	7.85	Regulatory	*BNC2*
rs776912	1	10847784	A	0.06	0.00	0.65	0.24	0.71	16.93	7.81	Intron	*CASZ1*
rs4676964	3	71034748	T	0.04	0.01	0.52	0.19	0.72	20.80	7.79	Intron	*FOXP1*
rs7568863	2	104642666	A	0.05	0.00	0.59	0.21	0.76	17.29	7.74	RNA	
rs679539	18	7743989	A	0.04	0.00	0.47	0.17	0.75	21.60	7.70	Intron	*PTPRM*
rs11751128	6	155561796	T	0.01	0.00	0.28	0.10	0.81	34.00	7.62	Missense	*TIAM2*
rs10960774	9	12739313	G	0.05	0.01	0.59	0.22	0.72	17.65	7.58	Regulatory	
rs9874207	3	71019750	T	0.03	0.00	0.51	0.18	0.80	18.54	7.57	Intron	*FOXP1*
rs6057598	20	31168281	C	0.05	0.07	0.71	0.28	0.57	18.25	7.43	RNA; Intron	*NOL4L*
rs10178579	2	139231958	A	0.03	0.01	0.50	0.18	0.73	20.30	7.42	Regulatory; RNA	
rs12727510	1	214356500	T	0.03	0.01	0.62	0.22	0.80	14.91	7.41	Intergenic	
rs1249090	1	48153140	G	0.04	0.04	0.53	0.20	0.63	22.20	7.41	Regulatory	
rs11170700	12	54210635	T	0.03	0.01	0.47	0.17	0.77	20.40	7.41	RNA	*RN7SKP289*
rs10810606	9	16754982	A	0.03	0.00	0.69	0.24	0.88	12.15	7.39	Intron	*BNC2*
rs12344347	9	16769662	T	0.02	0.00	0.69	0.24	0.89	11.92	7.38	Intron	*BNC2*
rs13281090	8	66590479	A	0.05	0.00	0.53	0.19	0.74	18.60	7.36	Intron	*MTFR1*
rs12915092	15	56035767	A	0.02	0.02	0.50	0.18	0.81	18.13	7.35	Regulatory	*PRTG*
rs4357572	1	50576710	A	0.03	0.03	0.54	0.20	0.68	19.76	7.32	Intron	*ELAVL4*
rs1052023	19	42799049	T	0.01	0.00	0.39	0.13	0.91	20.50	7.29	Synonymous	*CIC*
rs1126809	11	89017961	A	0.01	0.00	0.25	0.09	0.87	33.00	7.27	Missense	*TYR*
rs561828	11	126035297	C	0.04	0.01	0.49	0.18	0.72	20.20	7.19	RNA	
rs9887642	X	71348837	A	0.06	0.03	0.79	0.29	0.68	13.31	7.17	Regulatory; 3 prime UTR	*RGAG4; NHSL2*
rs10421769	19	33605312	T	0.03	0.01	0.65	0.23	0.81	13.54	7.15	Missense	*GPATCH1*
rs2217415	2	137731295	T	0.03	0.01	0.49	0.18	0.78	18.75	7.12	Intron	*THSD7B*
rs7593295	2	104781454	C	0.06	0.00	0.74	0.27	0.74	12.92	7.06	Regulatory	
rs17856697	17	7348625	G	0.01	0.00	0.35	0.12	0.87	22.80	7.05	Missense; Regulatory	*CHRNB1; FGF11*
rs13222241	7	113393990	A	0.03	0.01	0.53	0.19	0.80	16.72	7.04	Intergenic	
rs89962	12	52915172	T	0.02	0.01	0.41	0.15	0.81	21.10	7.01	Upstream gene	*KRT5*
rs63716	X	72489427	A	0.04	0.02	0.67	0.24	0.75	13.67	6.92	Intergenic	
rs55930529	8	144905741	T	0.02	0.01	0.53	0.18	0.86	15.10	6.92	Regulatory	*PUF60*
rs11781090	8	144871633	T	0.02	0.01	0.53	0.19	0.79	16.62	6.91	Downstream gene	*SCRIB*
rs6707475	2	74710491	C	0.09	0.18	0.84	0.37	0.38	21.70	6.91	Missense; Regulatory	*TTC31; CCDC142*
rs597894	X	71551322	T	0.08	0.02	0.77	0.29	0.66	13.43	6.86	Intron	*HDAC8*
rs6862358	5	78742980	G	0.03	0.01	0.52	0.18	0.79	16.65	6.86	Regulatory	*HOMER1*
rs11789463	9	16836011	A	0.02	0.00	0.51	0.18	0.89	14.97	6.84	Regulatory	*BNC2*
rs4299484	3	108189627	T	0.01	0.00	0.29	0.10	0.89	25.80	6.74	Missense; Regulatory	*MYH15*
rs61992671	14	101531854	G	0.02	0.04	0.49	0.18	0.66	20.80	6.73	NC transcript exon; RNA	*MIR412*
rs750607	15	74734819	T	0.03	0.01	0.54	0.19	0.82	14.85	6.67	Regulatory	*UBL7*
rs977745	17	59603321	T	0.04	0.11	0.64	0.26	0.49	21.30	6.67	Intergenic	
rs28760541	17	19172505	A	0.06	0.00	0.81	0.29	0.76	10.78	6.64	Regulatory	*EPN2*
rs569791	1	38507495	G	0.03	0.05	0.53	0.20	0.60	21.00	6.63	Downstream gene	*POU3F1*
rs3904600	6	7109665	G	0.12	0.01	0.63	0.25	0.53	19.91	6.61	Regulatory	*RREB1*
rs12131971	1	38809580	C	0.04	0.01	0.45	0.17	0.70	21.00	6.59	Intergenic	
rs12864339	13	59824906	C	0.03	0.01	0.45	0.16	0.77	19.27	6.59	Intergenic	
rs1267936	19	33643243	G	0.03	0.01	0.62	0.22	0.80	13.29	6.59	Regulatory	*WDR88*
rs7499011	16	81242198	A	0.09	0.00	0.38	0.16	0.46	37.00	6.58	Stop	*PKD1L2*
rs587118	9	35074917	C	0.02	0.00	0.48	0.17	0.84	16.37	6.57	Splice	*FANCG*
rs1411428	9	16661820	G	0.02	0.00	0.59	0.20	0.89	12.48	6.56	Intron	*BNC2*
rs12488457	3	130116696	C	0.05	0.17	0.72	0.32	0.40	22.80	6.53	Missense	*COL6A5*
rs17651549	17	44061278	T	0.01	0.00	0.24	0.08	0.82	33.00	6.53	Missense	*MAPT*
rs820373	3	123404986	G	0.07	0.07	0.78	0.31	0.55	15.29	6.52	Intron	*MYLK*
rs820371	3	123404711	C	0.07	0.07	0.77	0.30	0.55	15.53	6.52	Intron	*MYLK*
rs3170660	1	27210721	T	0.02	0.00	0.37	0.13	0.80	22.00	6.50	Missense	*GPN2*
rs35244551	12	122416064	T	0.02	0.00	0.46	0.16	0.86	16.26	6.49	Intron	*WDR66*
rs11211032	1	45246243	A	0.03	0.03	0.47	0.18	0.67	20.40	6.46	NC transcript exon	*RPS15AP11*
rs9403480	6	143605472	G	0.02	0.00	0.41	0.14	0.83	18.92	6.43	Intron	*AIG1*
rs959071	17	19142226	C	0.07	0.12	0.88	0.36	0.51	14.32	6.40	Intron	*EPN2*
rs952031	2	104430586	T	0.03	0.00	0.51	0.18	0.80	15.73	6.36	Intergenic	
rs12146246	10	74574063	T	0.04	0.03	0.63	0.23	0.68	14.81	6.35	Regulatory	*MCU*
rs2972190	5	72739879	T	0.02	0.00	0.35	0.13	0.82	21.90	6.35	Regulatory	*FOXD1*
rs40305	3	123411589	A	0.06	0.07	0.76	0.30	0.57	14.66	6.34	Regulatory	*MYLK*
rs61734605	11	34916657	T	0.02	0.00	0.33	0.12	0.81	23.80	6.32	Missense; Splice	*APIP*
rs13285378	9	16671729	T	0.02	0.00	0.54	0.19	0.90	12.98	6.31	Regulatory	*BNC2*
rs9650151	8	66628852	A	0.04	0.00	0.43	0.16	0.71	20.30	6.30	Regulatory	*MTFR1; PDE7A*
rs5913330	X	80087152	C	0.04	0.05	0.68	0.26	0.64	14.30	6.30	Intergenic	
rs4988235	2	136608646	A	0.03	0.00	0.51	0.18	0.83	14.86	6.29	Intron	*MCM6*
rs522681	11	126021204	G	0.04	0.01	0.51	0.18	0.73	17.04	6.27	Intergenic	
rs12440103	15	56036068	G	0.02	0.02	0.50	0.18	0.81	15.46	6.27	Regulatory	*PRTG*

SNP	CHR	POS	DER_ALLELE	DAF_AFR	DAF_EAS+EUR	DAF_GLOB	DAP	CADD	FineMAV_EAS+EUR	CONSEQUENCE	GENE
rs12471312	2	72789764	G	0.04	0.88	0.46	0.82	20.90	15.02	Regulatory	*EXOC6B*
rs1801187	X	32380996	T	0.03	0.59	0.31	0.78	32.00	14.62	Missense	*DMD*
rs1385699	X	65824986	T	0.05	0.90	0.48	0.77	18.50	12.78	Missense	*EDA2R*
rs3753841	1	103379918	A	0.05	0.66	0.35	0.70	26.50	12.19	Missense	*COL11A1*
rs5936453	X	147943667	G	0.05	0.94	0.49	0.76	17.04	12.14	Intron	*AFF2*
rs16957091	15	43017426	T	0.04	0.76	0.40	0.75	21.20	12.05	Synonymous	*CDAN1*
rs344478	7	146410468	A	0.08	0.90	0.49	0.67	19.56	11.77	Intron	*CNTNAP2*
rs35143646	X	2856155	T	0.08	0.79	0.43	0.62	23.70	11.64	Missense	*ARSE*
rs4844074	X	147897430	A	0.06	0.91	0.48	0.74	17.19	11.57	Intron	*AFF2*
rs10736817	11	78602525	A	0.04	0.70	0.37	0.76	21.50	11.51	Intron	*TENM4*
rs5943149	X	110655636	G	0.09	0.99	0.54	0.64	17.74	11.36	Upstream gene	*DCX*
rs6637417	X	127370947	A	0.03	0.85	0.44	0.83	16.02	11.31	Intergenic	
rs2037044	2	177682929	A	0.02	0.62	0.32	0.87	20.80	11.30	RNA	
rs5985320	X	110466936	A	0.09	0.97	0.53	0.64	18.18	11.30	3 prime UTR	*PAK3*
rs2034721	17	53894516	T	0.07	0.83	0.45	0.67	20.20	11.22	Regulatory	*PCTP*
rs5921046	X	97565093	G	0.06	0.81	0.44	0.70	19.63	11.19	Intergenic	
rs1596930	2	72826665	G	0.04	0.88	0.46	0.82	15.41	11.15	Intron	*EXOC6B*
rs2303223	16	31075175	A	0.02	0.64	0.33	0.87	19.99	11.14	Synonymous	*ZNF668*
rs1343879	X	75004529	A	0.03	0.43	0.23	0.74	35.00	11.12	Stop	*MAGEE2*
rs1078540	X	110951042	C	0.06	1.00	0.53	0.74	14.98	11.08	Intron	*ALG13*
rs4722666	7	27197601	G	0.10	0.85	0.47	0.57	22.50	10.89	Regulatory	*HOXA7; HOXA9*
rs2249797	5	98107341	T	0.02	0.61	0.32	0.82	21.70	10.89	Regulatory	*RGMB*
rs17671597	16	7381106	G	0.02	0.57	0.30	0.86	22.10	10.85	Regulatory	*RBFOX1*
rs2301721	7	27196113	T	0.10	0.84	0.47	0.57	22.50	10.82	Missense; Regulatory	*HOXA7*
rs143383	20	34025983	A	0.03	0.67	0.35	0.80	19.99	10.72	Regulatory; 5 prime UTR	*GDF5*
rs17822931	16	48258198	T	0.01	0.46	0.24	0.88	26.50	10.67	Missense	*ABCC11*
rs1973791	3	187416634	A	0.05	0.67	0.36	0.72	21.90	10.64	Synonymous	*RTP2*
rs9845788	3	135914715	A	0.05	0.70	0.37	0.70	21.70	10.64	Regulatory	*MSL2*
rs6872244	5	117489412	T	0.05	0.71	0.38	0.74	20.20	10.59	RNA	
rs4788890	17	73389446	G	0.08	0.89	0.48	0.66	17.89	10.53	Regulatory	*GRB2*
rs1126565	X	19373839	A	0.05	0.98	0.51	0.79	13.68	10.51	Synonymous	*PDHA1*
rs6979	16	67691668	A	0.03	0.66	0.34	0.79	20.20	10.48	Missense	*ACD*
rs2276118	11	67288594	T	0.03	0.61	0.32	0.79	21.80	10.42	Missense	*CABP2*
rs1049205	16	4942099	T	0.03	0.59	0.31	0.76	23.30	10.39	Missense	*PPL*
rs9428174	1	120585262	G	0.06	0.79	0.42	0.71	18.57	10.36	Intron	*NOTCH2*
rs184502	3	168196092	A	0.04	0.76	0.40	0.78	17.43	10.35	Intron	*EGFEM1P*
rs1352943	3	27368875	G	0.03	0.60	0.31	0.79	22.00	10.34	Regulatory	*NEK10*
rs1576050	10	31608510	A	0.06	0.87	0.47	0.71	16.72	10.32	Regulatory	*ZEB1*
rs1593304	7	131619847	G	0.07	0.83	0.45	0.65	19.10	10.28	RNA	
-	X	153629155	G	0.13	0.92	0.52	0.53	21.10	10.28	Missense	*RPL10*
rs5921045	X	97564948	G	0.06	0.81	0.44	0.70	18.15	10.26	Intergenic	
rs2394517	6	29069299	T	0.05	0.47	0.26	0.63	35.00	10.22	Stop	*OR2J1*
rs734312	4	6303354	A	0.03	0.69	0.36	0.80	18.60	10.21	Missense	*WFS1*
rs12507582	4	169369920	T	0.02	0.48	0.25	0.81	26.20	10.18	Missense	*DDX60L*
rs1132528	5	112929013	T	0.03	0.59	0.31	0.76	22.40	10.14	Missense	*YTHDC2*
rs269868	15	45392075	A	0.16	0.94	0.55	0.46	23.40	10.08	Missense	*DUOX2*
rs1036739	2	158665196	G	0.06	0.86	0.46	0.73	16.10	10.05	Intron	*ACVR1*
rs3809482	15	43661802	T	0.03	0.61	0.32	0.81	20.40	10.04	Missense	*ZSCAN29*
rs7193955	16	48122582	A	0.11	0.84	0.47	0.54	22.20	10.03	Missense	*ABCC12*
rs6546839	2	73680508	G	0.12	0.88	0.50	0.53	21.60	10.02	Missense	*ALMS1*
rs642652	2	72622480	C	0.05	0.90	0.48	0.75	14.77	9.99	Intron	*EXOC6B*
rs16903296	5	36721292	C	0.06	0.76	0.41	0.71	18.48	9.97	Regulatory	
rs12442525	15	42149506	C	0.06	0.71	0.38	0.67	21.00	9.96	Missense	*SPTBN5*
rs11655511	17	59287269	T	0.06	0.78	0.42	0.68	18.67	9.95	Regulatory	*BCAS3*
rs1061629	14	77926011	T	0.07	0.71	0.39	0.63	22.30	9.94	Synonymous; 5 prime UTR	*AHSA1*
rs1332720	X	64815536	C	0.07	1.00	0.53	0.72	13.84	9.94	Regulatory	*MSN*
rs2294504	X	109552667	T	0.02	0.60	0.31	0.88	18.82	9.93	NC transcript exon	*AMMECR1*
rs7421376	2	217902364	G	0.02	0.54	0.28	0.86	21.40	9.91	Intergenic	
rs4442810	16	65676733	T	0.02	0.57	0.29	0.85	20.60	9.89	RNA	
rs2037912	16	4933939	C	0.03	0.59	0.31	0.78	21.30	9.84	Missense	*PPL*
rs7944370	11	8105557	G	0.03	0.61	0.32	0.80	20.20	9.83	Regulatory	*TUB*
rs3760454	17	30222002	C	0.03	0.55	0.29	0.80	22.40	9.80	Missense	*UTP6*
rs6467654	7	136093047	A	0.03	0.62	0.32	0.82	19.19	9.77	RNA	
rs10497469	2	177807577	A	0.03	0.65	0.34	0.82	18.39	9.75	Intergenic	
rs2725405	17	79220224	C	0.03	0.57	0.30	0.81	21.20	9.75	Missense	*SLC38A10*
rs2747701	6	71238105	G	0.02	0.51	0.26	0.83	22.90	9.73	Missense	*FAM135A*
rs10790451	11	121570520	G	0.05	0.76	0.41	0.74	17.21	9.70	Intergenic	
rs8082149	17	54572895	C	0.06	0.77	0.42	0.70	17.78	9.70	Intron	*ANKFN1*
rs56103503	1	154980351	T	0.02	0.64	0.33	0.84	18.24	9.69	Intron	*ZBTB7B*
rs12855681	X	98858596	G	0.09	0.87	0.48	0.62	17.81	9.68	Intergenic	
rs238928	11	113973184	C	0.03	0.78	0.41	0.81	15.36	9.65	Intron	*ZBTB16*
rs24657	16	22972853	A	0.03	0.58	0.30	0.76	22.00	9.64	Intergenic	
rs6494466	15	64508763	G	0.04	0.81	0.42	0.79	14.98	9.63	Missense; Splice	*CSNK1G1*
rs1057079	1	11205058	T	0.06	0.78	0.42	0.68	18.27	9.60	Synonymous	*MTOR*
rs2602141	15	43724646	T	0.04	0.65	0.35	0.74	19.88	9.59	Missense	*TP53BP1*
rs7210574	17	73824121	T	0.04	0.61	0.33	0.74	21.20	9.57	Regulatory	*UNC13D; UNK*
rs6894514	5	117901084	T	0.09	0.77	0.43	0.56	22.10	9.56	RNA	
rs8068946	17	70712513	G	0.04	0.68	0.36	0.75	18.72	9.55	Regulatory	*SLC39A11*
rs11150220	16	79848178	G	0.04	0.74	0.39	0.77	16.58	9.55	RNA	*LINC01228*
rs2330442	7	42380071	A	0.03	0.66	0.35	0.78	18.58	9.54	Intergenic	
rs3827760	2	109513601	G	0.00	0.44	0.22	0.97	22.30	9.54	Missense	*EDAR*
rs12347057	9	126694094	T	0.11	0.79	0.45	0.54	22.30	9.53	Upstream gene	*DENND1A*
rs5919324	X	66430640	A	0.05	0.93	0.49	0.76	13.39	9.48	Intergenic	
rs224333	20	34023962	G	0.03	0.67	0.35	0.79	17.92	9.45	Regulatory	*GDF5*
rs62173642	2	177885414	G	0.08	0.73	0.40	0.60	21.60	9.45	RNA	
rs11665349	18	67601981	C	0.15	0.98	0.56	0.49	19.56	9.44	Intron	*CD226*
rs9930441	16	79862805	T	0.03	0.64	0.34	0.77	19.26	9.42	RNA	*LINC01228*
rs10499928	7	96384778	G	0.04	0.60	0.32	0.74	21.20	9.41	Intergenic	
rs1448335	X	20875921	G	0.07	0.80	0.43	0.66	17.83	9.38	Regulatory	
rs1735679	10	31786137	A	0.06	0.88	0.47	0.71	15.15	9.38	Intron	*ZEB1*
rs3747579	16	4445327	T	0.12	0.75	0.43	0.48	25.90	9.36	Missense	*CORO7*
rs12917189	15	43023482	T	0.05	0.77	0.41	0.74	16.37	9.35	Missense	*CDAN1*
rs4888047	16	79826489	G	0.04	0.70	0.37	0.77	17.18	9.31	RNA	*LINC01229*
rs4891384	18	67624554	C	0.11	0.98	0.54	0.59	16.09	9.29	Regulatory	*CD226*
rs4850759	2	197952851	G	0.03	0.74	0.39	0.81	15.55	9.28	Intron	*ANKRD44*
rs11559290	4	159601676	T	0.15	0.92	0.53	0.48	20.90	9.27	Missense; Splice	*ETFDH*
rs1005502	8	117167843	T	0.03	0.59	0.31	0.79	19.88	9.26	RNA	*LINC00536*
rs2292249	12	83252600	G	0.03	0.59	0.31	0.78	20.10	9.26	Intron	*TMTC2*
rs10786871	10	107319562	C	0.05	0.67	0.36	0.71	19.39	9.25	Intergenic	
rs12849919	X	139115671	A	0.07	0.79	0.43	0.65	17.91	9.25	Downstream gene	*HNRNPA3P3*

SNP	CHR	POS	DER_ALLELE	DAF_AFR	DAF_AMR	DAF_EAS	DAF_EUR	DAF_GLOB	DAP	CADD	FineMAV_AMR	CONSEQUENCE	GENE
rs34890031	7	133884081	A	0.00	0.77	0.04	0.00	0.20	0.86	32.00	21.32	Missense	*LRGUK*
rs61750329	1	11585308	T	0.00	0.70	0.01	0.02	0.18	0.89	24.90	15.65	Missense	*PTCHD2*
rs142326775	10	31134425	T	0.00	0.45	0.00	0.00	0.11	0.97	35.00	15.38	Missense	*ZNF438*
rs146573098	12	113875765	A	0.00	0.57	0.00	0.00	0.14	0.99	26.50	14.89	Missense	*SDSL*
rs12421620	11	66276576	A	0.01	0.89	0.01	0.03	0.23	0.88	18.19	14.15	Missense	*DPP3*
rs1871044	11	67276158	G	0.02	0.75	0.04	0.00	0.20	0.80	22.30	13.42	Regulatory	*PITPNM1; CDK2AP2*
rs149808694	16	9811265	C	0.00	0.73	0.01	0.00	0.18	0.97	18.87	13.28	Intergenic	
rs138840395	6	157264533	A	0.00	0.73	0.00	0.00	0.18	0.98	18.38	13.13	Regulatory	*ARID1B*
rs142949209	7	147359632	T	0.00	0.66	0.00	0.00	0.17	0.99	19.93	12.97	Intron	*CNTNAP2*
rs143128534	6	56332359	T	0.00	0.64	0.00	0.00	0.16	1.00	19.83	12.62	Intron	*DST*
rs4924468	15	40729596	C	0.00	0.57	0.00	0.00	0.14	0.98	22.30	12.37	Up- Downstream gene	*BAHD1; IVD*
rs149320184	7	147375822	C	0.00	0.64	0.00	0.00	0.16	1.00	19.17	12.20	Intron	*CNTNAP2*
rs145495982	7	147379361	G	0.00	0.64	0.00	0.00	0.16	1.00	18.95	12.06	Intron	*CNTNAP2*
rs1574103	11	67057599	T	0.00	0.77	0.03	0.00	0.20	0.86	18.03	11.98	Regulatory; Synonymous	*ANKRD13D*
rs35545453	3	69031963	C	0.00	0.57	0.01	0.01	0.15	0.92	22.90	11.91	Missense; Splice	*EOGT*
rs146181399	7	147371437	A	0.00	0.64	0.00	0.00	0.16	1.00	18.55	11.81	Intron	*CNTNAP2*
rs17739774	14	47770258	T	0.00	0.82	0.00	0.06	0.22	0.81	17.84	11.80	Intron	*MDGA2*
rs150664883	18	35146206	A	0.00	0.70	0.02	0.05	0.19	0.74	22.40	11.69	Regulatory	*CELF4*
rs17819664	15	57221167	T	0.00	0.66	0.00	0.02	0.17	0.89	19.68	11.60	Regulatory	*TCF12*
rs79317492	15	57537491	C	0.00	0.66	0.00	0.02	0.17	0.89	19.40	11.43	Regulatory	*TCF12*
rs9548505	13	39422624	T	0.13	0.70	0.00	0.08	0.23	0.46	35.00	11.43	Missense	*FREM2*
rs12441769	15	57116557	T	0.00	0.66	0.00	0.02	0.17	0.90	19.13	11.38	Intron	*ZNF280D*
rs145088108	1	32740366	A	0.00	0.61	0.01	0.00	0.16	0.92	19.99	11.33	Missense	*LCK*
rs74431877	1	32193049	A	0.00	0.66	0.01	0.00	0.17	0.95	17.74	11.16	Missense	*ADGRB2*
rs74719094	14	96768380	G	0.00	0.52	0.01	0.00	0.13	0.92	23.10	11.14	Missense	*ATG2B*
rs77499580	12	98578714	A	0.00	0.61	0.00	0.00	0.15	0.98	18.57	11.11	RNA	
rs138670348	7	147375810	C	0.00	0.64	0.00	0.00	0.16	1.00	17.34	11.04	Intron	*CNTNAP2*
rs117867971	3	60269689	C	0.00	0.61	0.01	0.00	0.16	0.93	19.24	11.02	Intron	*FHIT*
rs62621285	11	130058437	A	0.00	0.57	0.00	0.03	0.15	0.82	23.60	11.02	Missense	*ST14*
rs61740840	9	124794020	T	0.00	0.57	0.01	0.02	0.15	0.84	22.90	10.98	Missense	*TTLL11*
rs79256594	3	89177432	T	0.01	0.80	0.05	0.02	0.22	0.77	17.95	10.95	Intron	*EPHA3*
rs140151892	12	49214015	G	0.00	0.59	0.02	0.00	0.15	0.88	20.90	10.91	NC transcript exon	*CACNB3*
rs12474958	2	163150282	G	0.00	0.61	0.00	0.01	0.16	0.93	19.09	10.90	Regulatory	*IFIH1*
rs143040189	12	110937404	C	0.00	0.52	0.00	0.00	0.13	0.97	21.30	10.85	Intron	*VPS29*
rs141137978	7	147379526	C	0.00	0.64	0.00	0.00	0.16	1.00	16.79	10.69	Intron	*CNTNAP2*
rs2293766	7	100371358	A	0.02	0.80	0.48	0.00	0.32	0.29	47.00	10.66	Stop	*ZAN*
rs117487309	22	41195082	A	0.01	0.77	0.01	0.01	0.20	0.90	15.24	10.55	Synonymous	*SLC25A17*
rs12422382	12	116131292	T	0.00	0.52	0.00	0.00	0.13	0.97	20.60	10.49	Intergenic	
rs78262741	9	4290185	A	0.00	0.52	0.00	0.00	0.13	0.98	20.50	10.45	Regulatory	*GLIS3*
rs189645263	11	33278862	G	0.00	0.48	0.00	0.00	0.12	0.98	22.10	10.36	Regulatory	*HIPK3*
rs12478730	2	163150279	C	0.00	0.61	0.00	0.01	0.16	0.93	17.95	10.25	Regulatory	*IFIH1*
rs6130686	20	43371253	T	0.00	0.52	0.02	0.00	0.14	0.88	22.10	10.20	Intron	*KCNK15*
rs144851788	18	50510228	C	0.01	0.52	0.00	0.00	0.13	0.94	20.60	10.10	Intron	*DCC*
rs41307728	1	229594003	T	0.02	0.43	0.00	0.01	0.11	0.85	27.60	10.09	Missense	*NUP133*
rs12540617	7	100815816	T	0.00	0.84	0.02	0.02	0.22	0.87	13.69	10.06	Synonymous	*NAT16*
rs145924266	3	138665815	T	0.04	0.75	0.03	0.00	0.21	0.76	17.66	10.05	Regulatory; 5 prime UTR	*FOXL2*
rs75681209	15	57451085	G	0.00	0.64	0.00	0.02	0.17	0.88	17.95	10.04	Regulatory	*TCF12*
rs79140822	1	98298087	G	0.00	0.73	0.03	0.00	0.19	0.87	15.78	10.03	Regulatory	*DPYD*
rs142443364	3	138660894	T	0.01	0.75	0.02	0.00	0.19	0.89	14.98	10.02	RNA	*LINC01391*
rs148564531	7	147382260	G	0.00	0.64	0.00	0.00	0.16	1.00	15.75	10.02	Intron	*CNTNAP2*
rs114571119	3	138657902	A	0.04	0.73	0.02	0.00	0.20	0.78	17.49	9.98	RNA	*LINC01391*
rs138576967	6	157270044	A	0.00	0.75	0.00	0.00	0.19	0.98	13.48	9.94	Regulatory	*ARID1B*
rs2066494	11	67165015	C	0.06	0.77	0.03	0.01	0.22	0.69	18.55	9.88	Synonymous	*RAD9A*
rs147339364	12	98446787	C	0.00	0.68	0.03	0.00	0.18	0.87	16.59	9.83	intergenic	
rs75828095	6	56512199	G	0.00	0.61	0.01	0.00	0.16	0.95	16.80	9.78	Regulatory	*DST*
rs145602591	1	32126672	T	0.00	0.57	0.02	0.00	0.15	0.87	19.80	9.77	Missense	*COL16A1*
rs116131136	7	100807230	T	0.00	0.48	0.01	0.00	0.12	0.92	22.10	9.71	Missense; Regulatory	*VGF*
rs1055280	7	135661990	T	0.00	0.50	0.01	0.00	0.13	0.89	21.80	9.68	Regulatory; 5 prime UTR	*MTPN*
rs117415954	12	98446717	C	0.00	0.68	0.03	0.00	0.18	0.87	16.31	9.66	Intergenic	
rs147152654	4	71724337	T	0.00	0.52	0.00	0.00	0.13	0.98	18.87	9.63	Intergenic	
rs188337800	7	132117258	C	0.00	0.45	0.00	0.00	0.11	1.00	21.10	9.59	Regulatory	*PLXNA4*
rs16940118	12	109754567	T	0.00	0.84	0.00	0.00	0.21	0.99	11.47	9.57	Intergenic	
rs147302393	21	34925304	G	0.00	0.41	0.00	0.00	0.10	1.00	23.30	9.53	Missense; Regulatory	*SON*
rs73009507	19	10961024	G	0.00	0.50	0.00	0.04	0.14	0.78	24.30	9.46	Missense	*C19orf38*
rs35047625	22	32000930	T	0.00	0.41	0.05	0.00	0.12	0.70	33.00	9.39	Missense; Splice	*SFI1*
rs76326377	3	28652325	G	0.00	0.73	0.08	0.05	0.22	0.61	21.20	9.38	Intron	*RBMS3*
rs79250477	15	57436183	A	0.00	0.66	0.00	0.02	0.17	0.89	15.81	9.32	Intron	*TCF12*
rs2019884	12	98481159	C	0.00	0.68	0.04	0.00	0.18	0.83	16.48	9.31	Intergenic	
rs16934463	10	33713572	G	0.00	0.48	0.01	0.00	0.12	0.89	21.90	9.31	Intergenic	
rs12440729	15	57497961	A	0.00	0.66	0.00	0.02	0.17	0.89	15.73	9.27	Intron	*TCF12*
rs145682935	1	53125880	G	0.00	0.66	0.00	0.00	0.17	0.99	14.21	9.25	Intron	*FAM159A*
rs145384767	7	147417499	C	0.00	0.61	0.00	0.00	0.15	1.00	15.05	9.23	Intron	*CNTNAP2*
rs17800083	3	70784519	T	0.00	0.57	0.00	0.01	0.15	0.92	17.61	9.23	Intergenic	
rs150384670	3	89975292	G	0.00	0.43	0.00	0.00	0.11	0.98	21.60	9.18	Intergenic	
rs140696197	5	81153904	C	0.00	0.64	0.01	0.00	0.16	0.93	15.31	9.08	Intergenic	
rs148257703	10	31441807	G	0.00	0.43	0.00	0.00	0.11	1.00	21.00	9.07	Intergenic	
rs11083711	19	44001379	C	0.02	0.68	0.09	0.05	0.21	0.54	24.80	9.06	Missense	*PHLDB3*
rs4788114	16	28995145	T	0.01	0.57	0.00	0.06	0.16	0.71	22.50	9.04	Synonymous	*SPNS1*
rs139787913	11	111601686	A	0.00	0.50	0.01	0.00	0.13	0.93	19.54	9.04	Regulatory	*PPP2R1B*
rs149326963	1	54863067	T	0.00	0.50	0.00	0.02	0.13	0.91	19.87	9.01	Intron	*SSBP3*
rs961269	12	109915143	A	0.00	0.48	0.00	0.00	0.12	1.00	18.86	9.00	Regulatory; 5 prime UTR	*KCTD10; UBE3B*
rs78662103	10	61142860	T	0.03	0.75	0.04	0.01	0.21	0.74	16.07	8.98	Intergenic	
rs185677681	1	176354064	C	0.00	0.50	0.00	0.00	0.13	1.00	17.91	8.96	Intergenic	
rs10509182	10	64760466	C	0.00	0.59	0.05	0.00	0.16	0.78	19.41	8.95	Intergenic	
rs138004015	12	31653216	T	0.00	0.61	0.01	0.00	0.16	0.93	15.71	8.94	Intron	*DENND5B*
rs3735695	7	100452239	T	0.00	0.70	0.05	0.00	0.19	0.80	15.81	8.93	Splice	*SLC12A9*
rs2287541	12	14664250	G	0.01	0.52	0.01	0.05	0.15	0.68	25.00	8.87	Missense	*PLBD1*
rs4537942	14	24901824	T	0.01	0.73	0.01	0.04	0.20	0.80	15.11	8.84	Splice	*KHNYN*
rs1816322	15	57330511	G	0.00	0.66	0.00	0.02	0.17	0.89	15.00	8.84	Intron	*TCF12*
rs2268883	2	163041387	G	0.01	0.82	0.18	0.01	0.25	0.53	20.50	8.83	Intron	*FAP*
rs140164015	14	61059047	T	0.00	0.55	0.00	0.00	0.14	0.99	16.39	8.83	Intergenic	
rs16880472	6	50562461	A	0.00	0.61	0.01	0.00	0.15	0.97	14.70	8.77	RNA	
rs145570841	10	28409298	T	0.00	0.48	0.00	0.00	0.12	1.00	18.37	8.77	Missense	*MPP7*
rs78180656	16	73587930	A	0.00	0.70	0.04	0.01	0.19	0.81	15.27	8.77	RNA	
rs77618848	20	52277148	C	0.00	0.45	0.01	0.00	0.12	0.91	21.10	8.75	Regulatory	
rs12032735	1	27757291	G	0.01	0.89	0.23	0.00	0.28	0.49	20.20	8.74	Regulatory	*WASF2*
rs150087094	7	147375742	G	0.00	0.64	0.00	0.00	0.16	1.00	13.71	8.73	Intron	*CNTNAP2*
rs148387576	1	170570525	T	0.01	0.73	0.05	0.00	0.20	0.79	15.15	8.71	RNA	
rs147396436	10	33680982	C	0.00	0.52	0.01	0.00	0.13	0.95	17.46	8.69	Intergenic	
rs4684677	3	10328453	A	0.01	0.57	0.00	0.07	0.16	0.68	22.30	8.66	Missense	*GHRL*

Appendix C

List of primary phenotyping tests

The list below specifies a standard set of phenotyping tests that were applied to all mouse strains generated in this study. Provided here information was derived from Wellcome Trust Sanger Institute Mouse Pipelines internal website (mouse.internal.sanger.ac.uk).

Homozygous viability at P14
Recessive Lethal Study
Homozygous Fertility
General Observations
Weight Curves
Neurological Assessment
Grip Strength
Dysmorphology
Indirect Calorimetry
Glucose Tolerance (ip)
Auditory Brainstem Response
Body Composition (DEXA)
X-ray Imaging
Eye Morphology
Plasma Chemistry
Insulin
Haematology Terminal
Micronuclei
PBL Terminal
Heart Weight
Brain Histopathology
Eye Histopathology

Salmonella Challenge

Citrobacter Challenge

Cytotoxic T Cell Function

Spleen Immunophenotyping

Mesenteric Lymph Node

Bone Marrow

Anti-nuclear Antibody Assay

Epidermal Immune Composition

DSS Challenge

Influenza Challenge

Trichuris Challenge

OBCD Bone

Appendix D

List of publications

A Selective Sweep on a Deleterious Mutation in *CPT1A* in Arctic Populations

American Journal of Human Genetics 2014

Clemente, F. J., Cardona, A., Inchley, C. E., Peter, B. M., Jacobs, G., Pagani, L., Lawson, D. J., Antao, T., Vicente, M., Mitt, M., DeGiorgio, M., Faltyskova, Z., Xue, Y., Ayub, Q., **Szpak, M.**, Magi, R., Eriksson, A., Manica, A., Raghavan, M., Rasmussen, M., Rasmussen, S., Willerslev, E., Vidal-Puig, A., Tyler-Smith, C., Villems, R., Nielsen, R., Metspalu, M., Malyarchuk, B., Derenko, M., Kivisild, T.

Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding

Science 2015

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., **Szpak, M.**, Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D. N., de Manuel, M., Hernandez-Rodriguez, J., Lobon, I., Siegismund, H. R., Pagani, L., Quail, M. A., Hvilsom, C., Mudakikwa, A., Eichler, E. E., Cranfield, M. R., Marques-Bonet, T., Tyler-Smith, C., Scally, A.

Prioritizing Candidate Genetic Variants Driving Local Adaptations in Human Populations

Under review

Szpak, M., Mezzavilla, M., Ayub, Q., Chen, Y., Xue, Y., Tyler-Smith, C.

Whole genome sequencing coupled to imputation discovers genetic signals for anthropometric traits

Under review

Tachmazidou, I., Süveges, D., Min, J. L., Ritchie, G. R., Steinberg, J., Walter, K., Iotchkova, V., Schwartzentruber, J., Huang, J., Memari, Y., McCarthy, S., Crawford, A. A., Bombieri, C., Cocca, M., Farmaki, A. E., Gaunt, T. R., Jousilahti, P., Kooijman, M. N., Lehne, B., Malerba, G., Männistö, S., Matchan, A., Medina-Gomez, C., Metrustry, S. J., Nag, A., Ntalla, I., Paternoster, L., Rayner, N. W., Sala, C., Scott, W. R., Shihab, H. A., Southam, L., St Pourcain, B., Traglia, M., Trajanoska, K., Zaza, G., Zhang, W., Artigas, M. S., Bansal, N., Benn, M., Chen, Z., Danecek, P., Lin, W. Y., Locke, A., Luan, J., Manning, A. K., Mulas, A., Sidore, C., Tybjaerg-Hansen, A., Varbo, A., Zoledziewska, M., Finan, C., Hatzikotoulas, K., Hendricks, A. E., Kemp, J. P., Moayyeri, A., Panoutsopoulou, K., **Szpak, M.**, Wilson, S. G., Boehnke, M., Cucca, F., Di Angelantonio, E., Langenberg, C., Lindgren, C., McCarthy, M. I., Morris, A. P., Nordestgaard, B. G., Scott, R. A., Tobin, M. D., Wareham, N. J., SpiroMeta consortium, GoT2D consortium, Burton, P., Chambers, J. C., Davey Smith, G., Dedoussis, G., Felix, J. F., Franco, O. H., Gambaro, G., Gasparini, P., Hammond, C. J., Hofman, A., Jaddoe, V. W., Kleber, M., Kooner, J. S., Perola, M., Relton, C., Ring, S. M., Rivadeneira, F., Salomaa, V., Spector, T. D., Stegle, O., Toniolo, D., Uitterlinden, A. G., arcOGEN consortium, Understanding Society Scientific Group, UK10K consortium, Barroso, I., Perry, J. R., Walker, B. R., Butterworth, A. S., Xue, Y., Durbin, R., Small, K. S., Soranzo, N., Timpson, N. J., Zeggini, E.