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

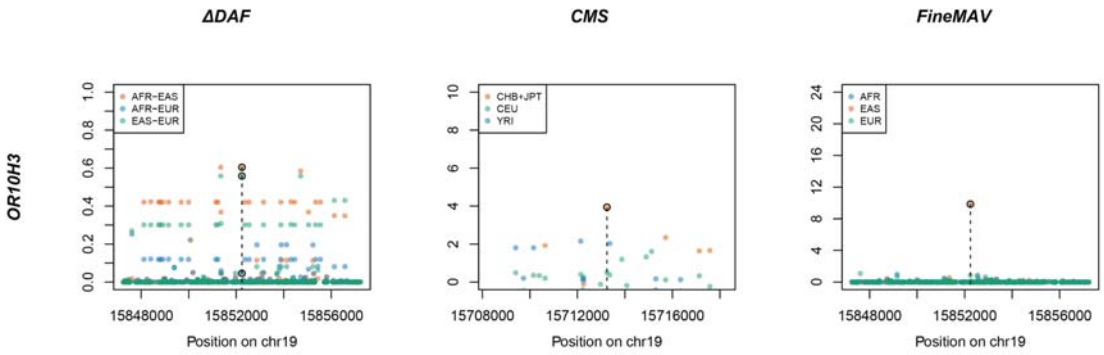

Figure 23. Signal of selection in *OR10H3* according to three different approaches. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic window spanning 10 kb around the variant of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). Normalised *CMS* 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)	-

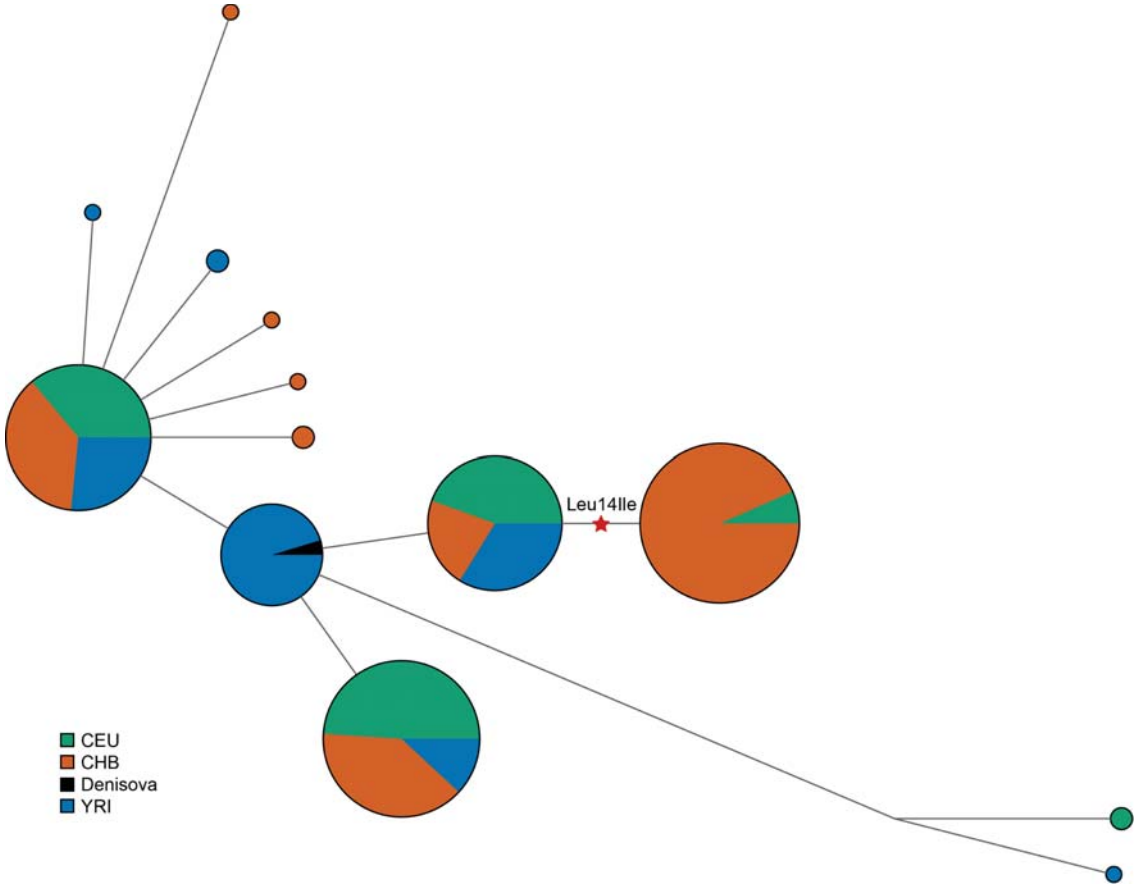

Figure 24. Haplotype network of *OR10H3*. The median-joining haplotype network was generated using Network 5.0.0.0 (430) and sequences of 270 individuals from the 1000 Genome Project, Phase 1 (85 CEU, 97 CHB, 88 YRI) (159) and the high-coverage Denisova genome (246). Each circle represents a distinct 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.

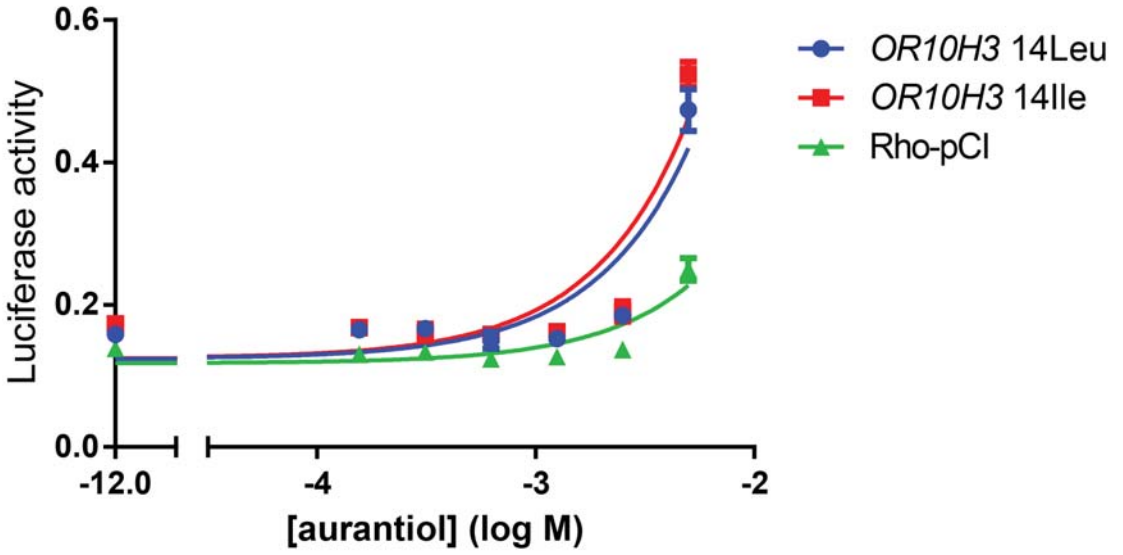

Figure 25. Dose-response curve of the derived (14Ile) and ancestral (14Leu) version of the *OR10H3* receptor to the aurantiol odorant. X-axis shows the concentration of aurantiol in Log Molar units. Y-axis shows the luciferase response (each concentration was tested in triplicate, error bars indicates ± 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-

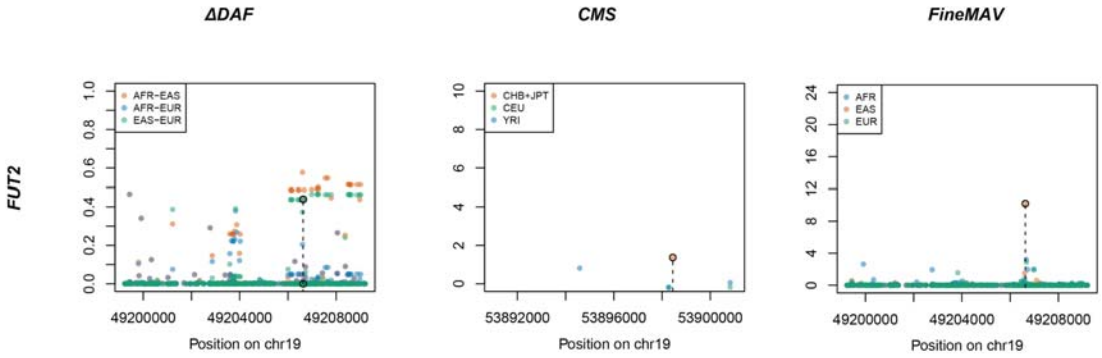

Figure 26. Signal of selection in *FUT2* according to three different approaches. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic window spanning the *FUT2* gene. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). Normalised *CMS* scores (123) were calculated 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

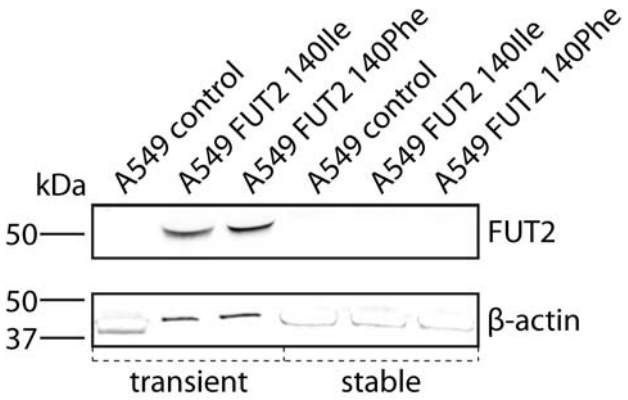

Figure 27. Western Blot analysis cropped to show regions of interest. Lysates of transient and stable transfectants of A549 cells carrying Myc-tagged FUT2, ancestral (140Ile) or derived (140Phe) vector, were separated by SDS-PAGE and transferred onto nitrocellulose membranes. Plain A549 served as a negative 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

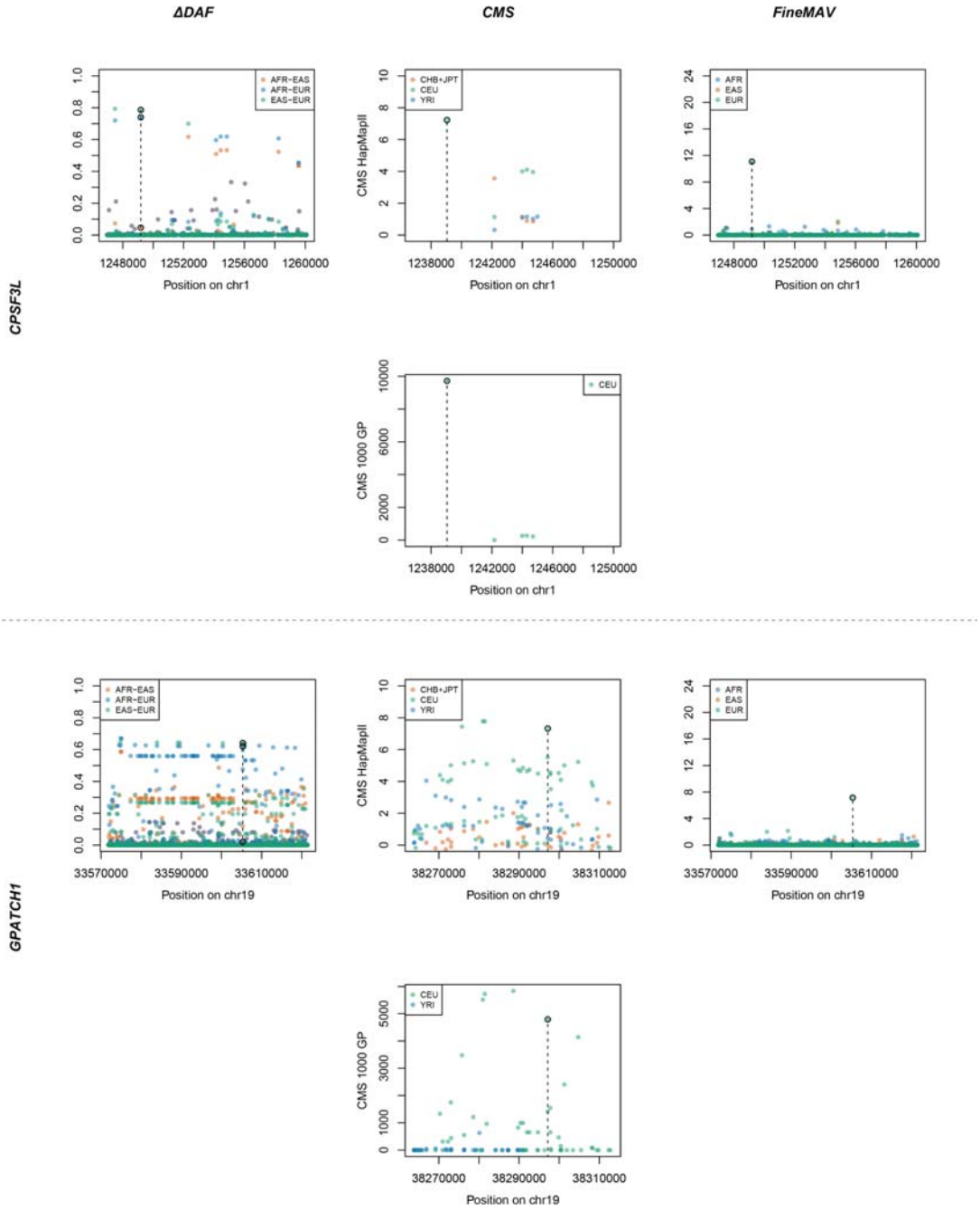

Figure 28. Comparison of three different approaches for pinpointing selected variants. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). *CMS* scores are given for booth, 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.

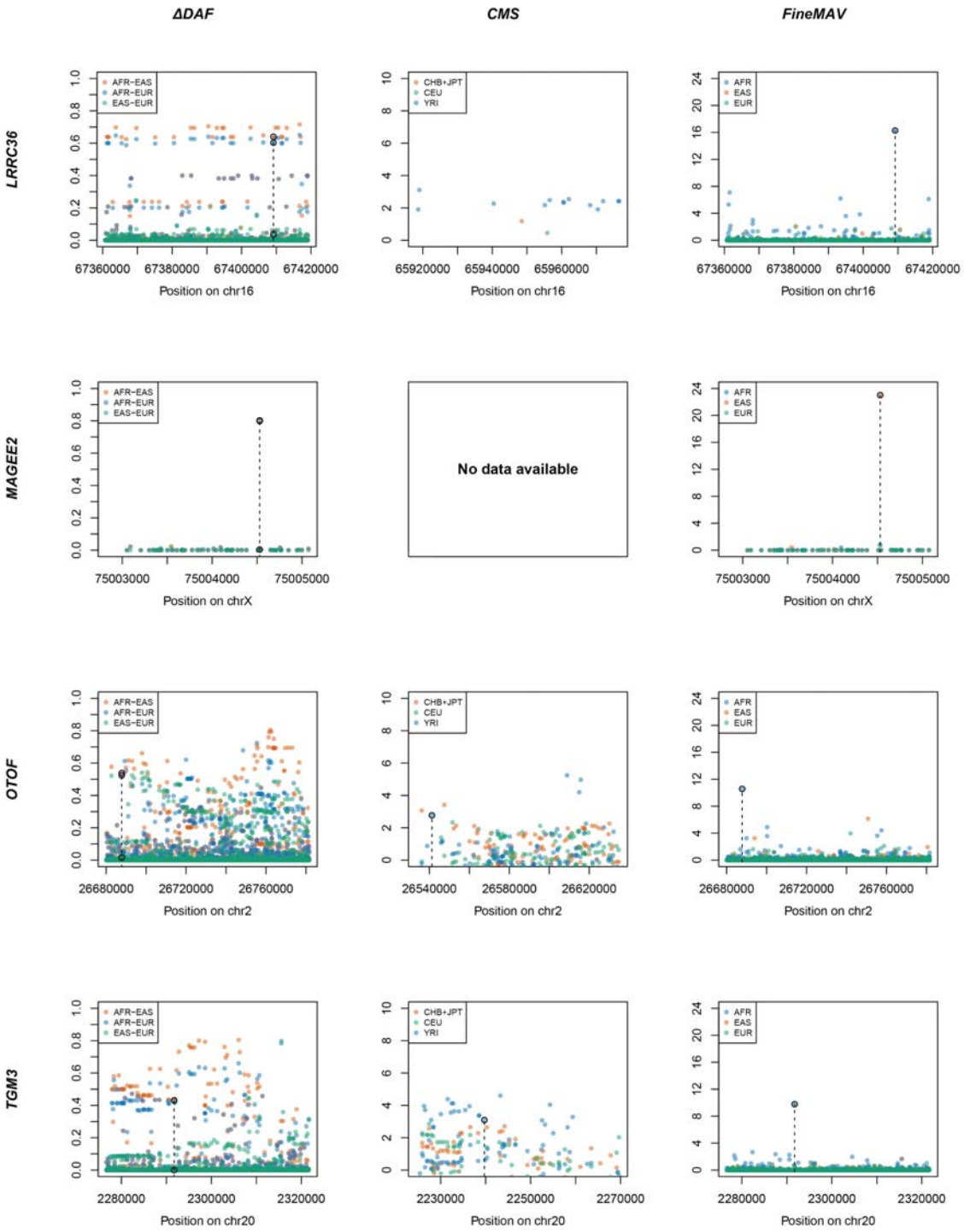

Figure 29. Comparison of three different approaches for pinpointing selected variants. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). Normalised *CMS* scores (123) 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

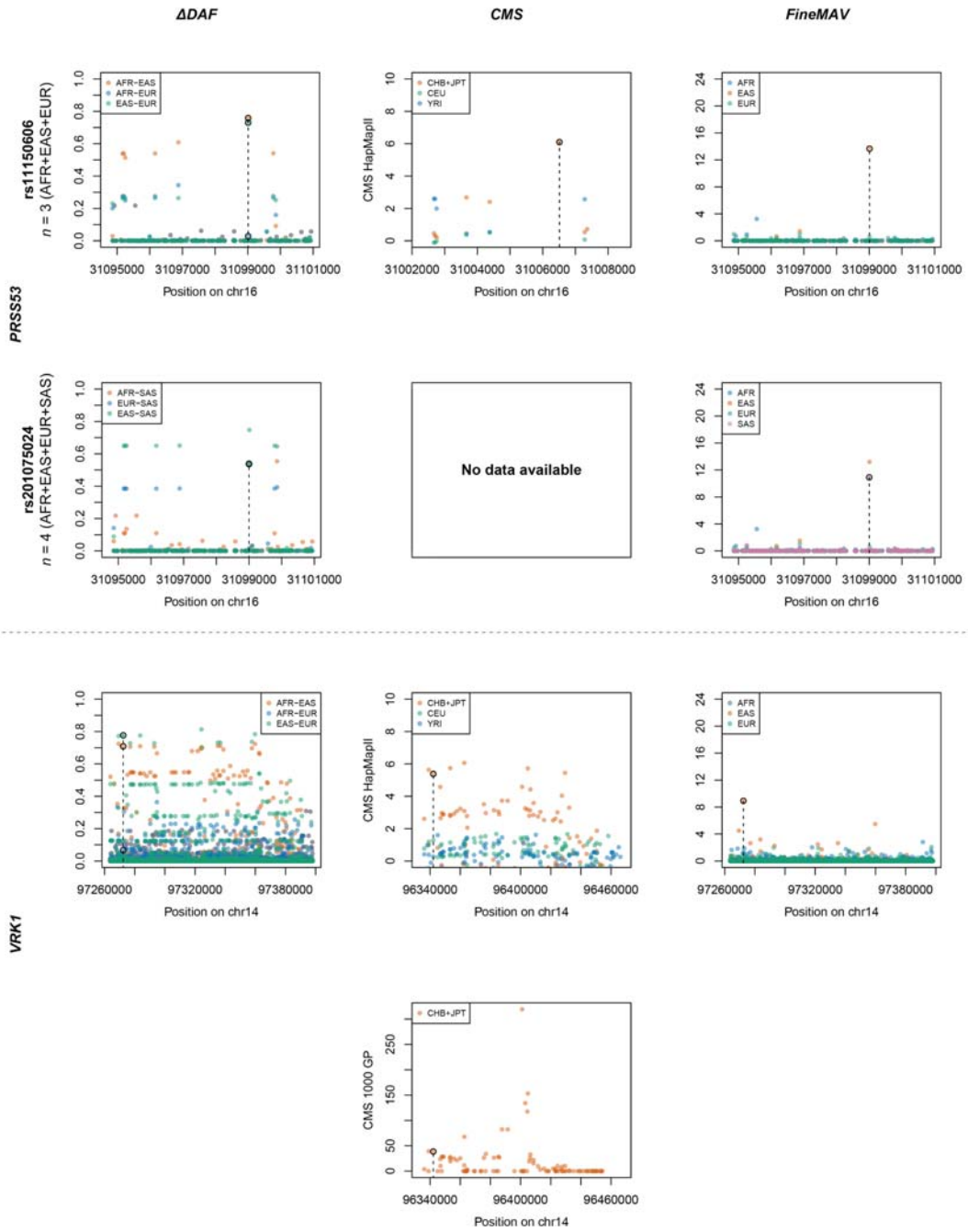

Figure 30. Comparison of three different approaches for pinpointing selected variants. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). *CMS* scores are given for the 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.

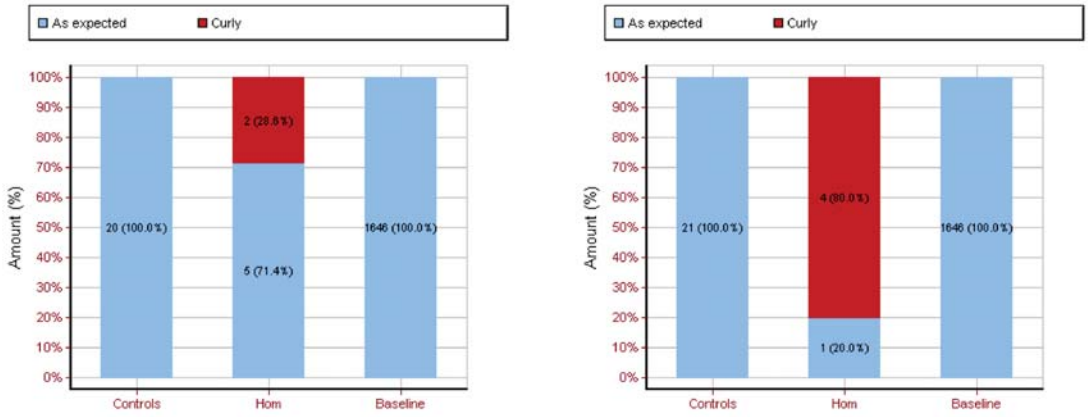

Figure 31. Vibrissae shape in *Prss53*$^{-/-}$ mice. Left panel: Females. Right panel: Males. The first number in the bar indicates the number of individuals investigated (*n*) followed by the percentage given in parentheses.

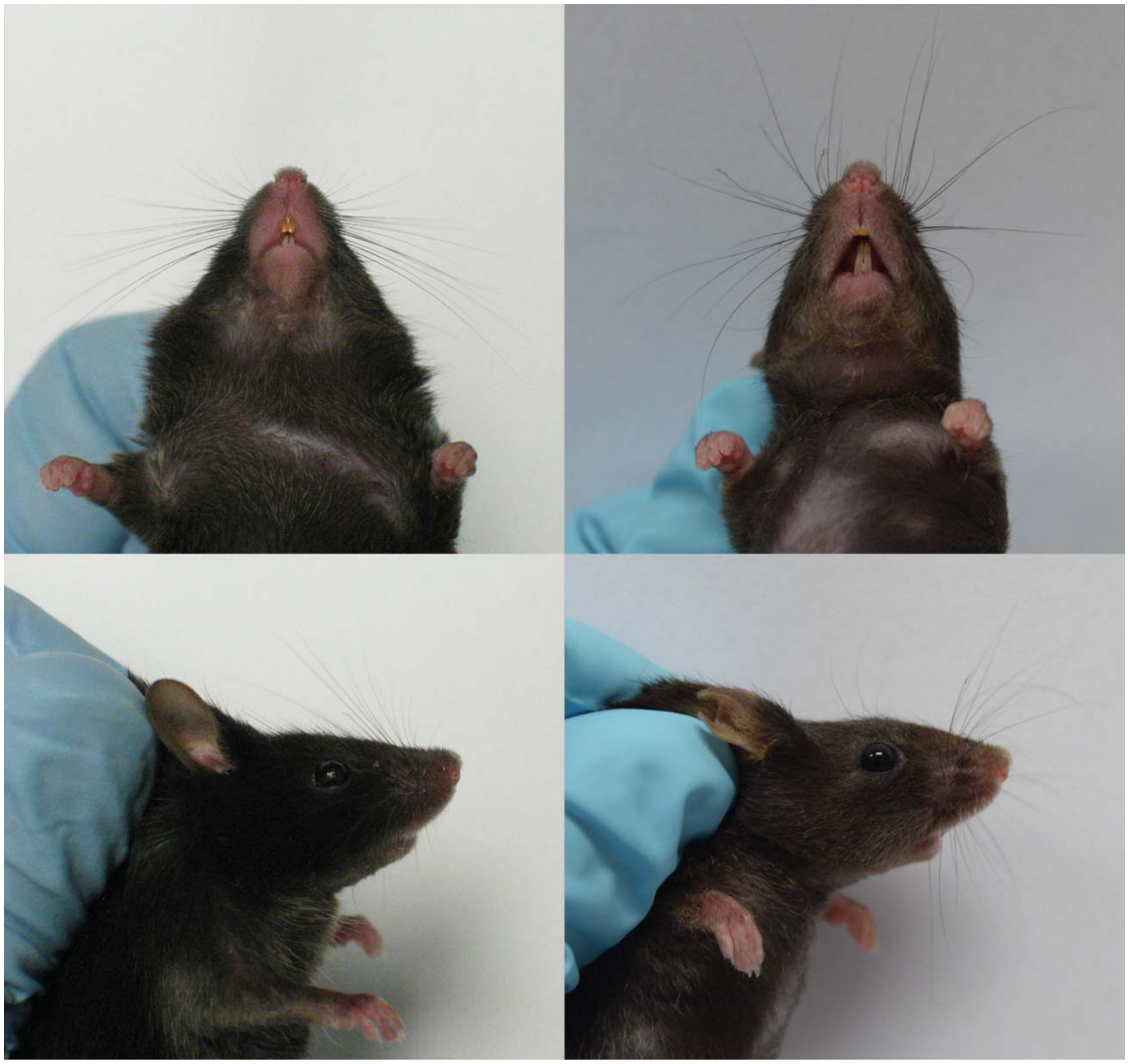

Figure 32. Vibrissae shape in *Prss53*$^{-/-}$ mice. Left panel: Control. Right panel: *Prss53* null mutant.

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

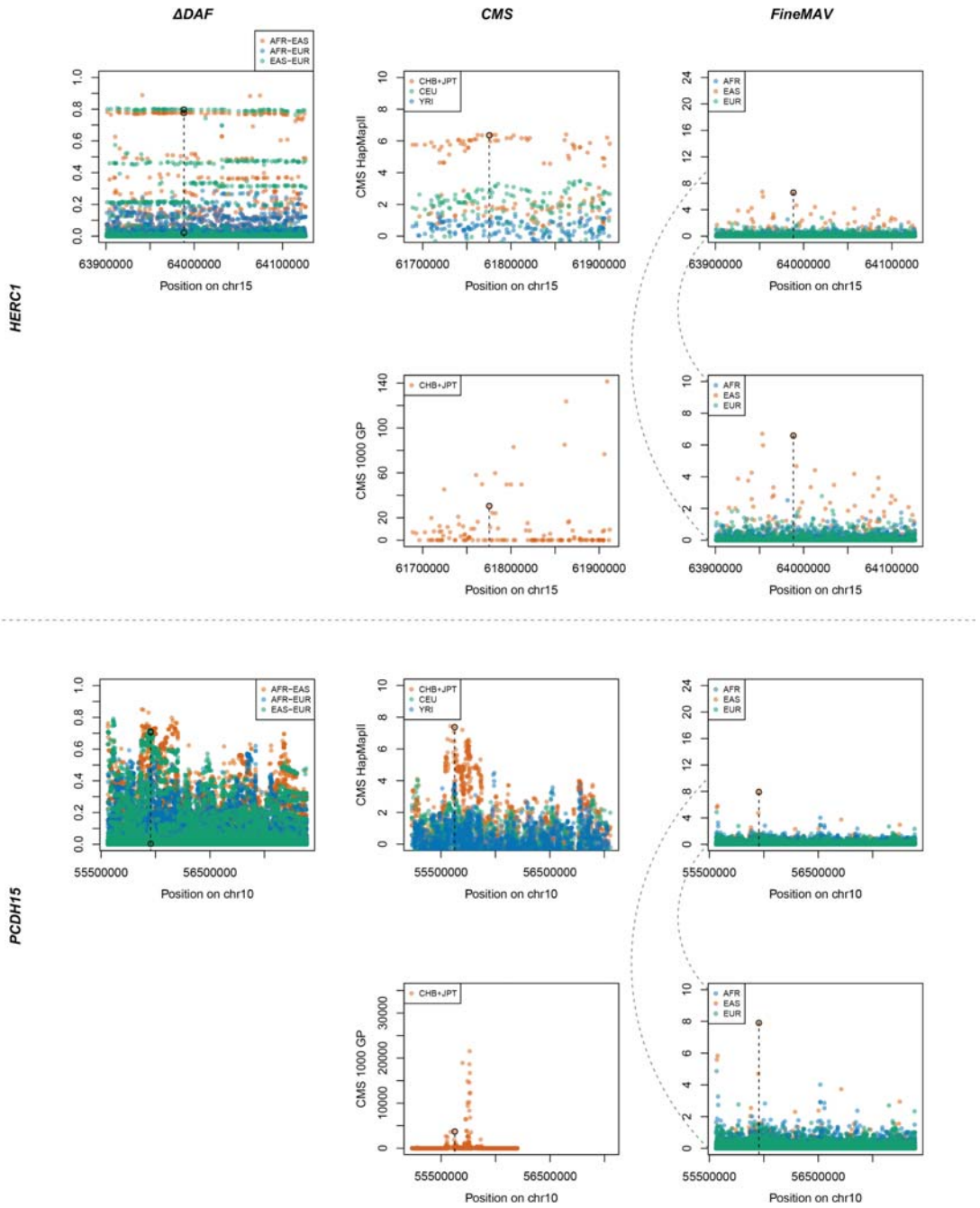

Figure 33. Comparison of three different approaches for pinpointing selected variants. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes of interest. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142). Expanded view of the FineMAV 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.

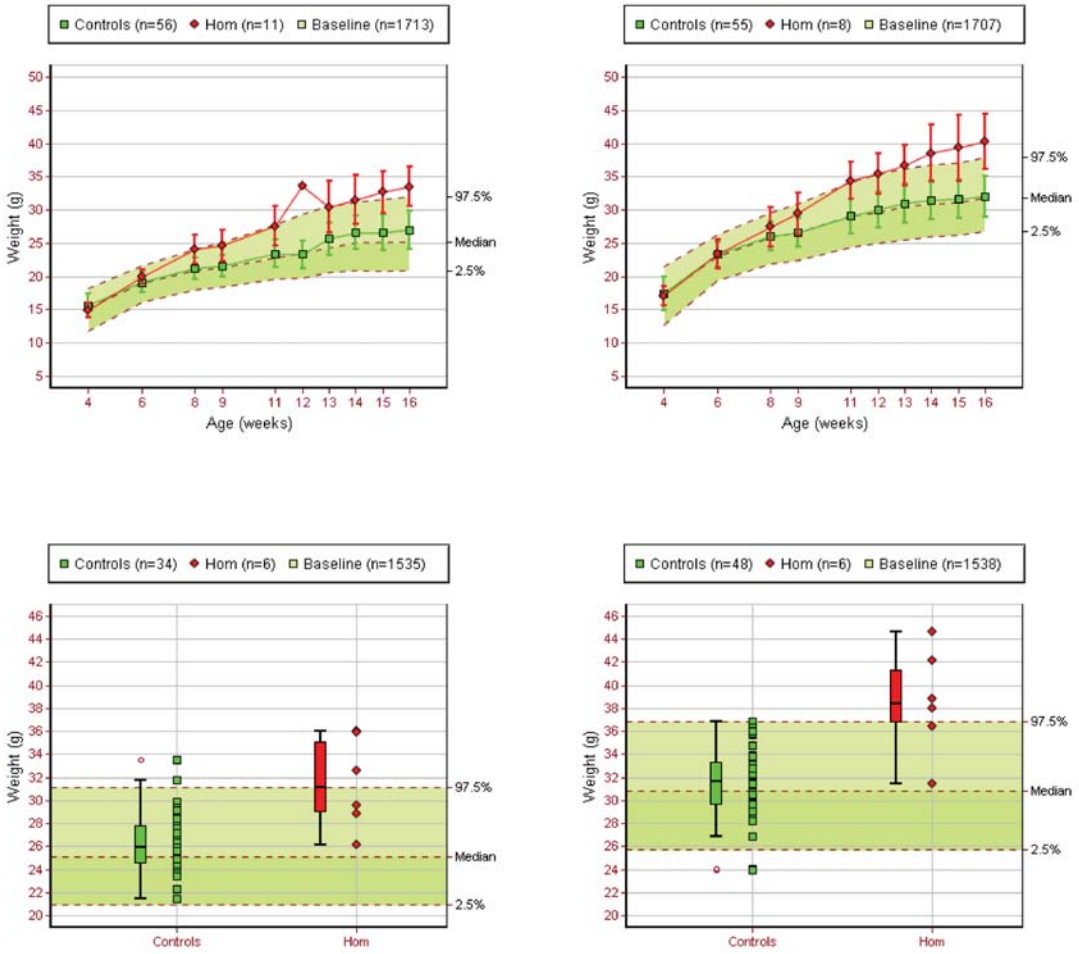

Figure 34. Increased body weight in *Herc1*$^{-/-}$ mice. Top panel: Average weight curve; Bottom panel: Body weight; Left panel: Females; Right panel: Males.

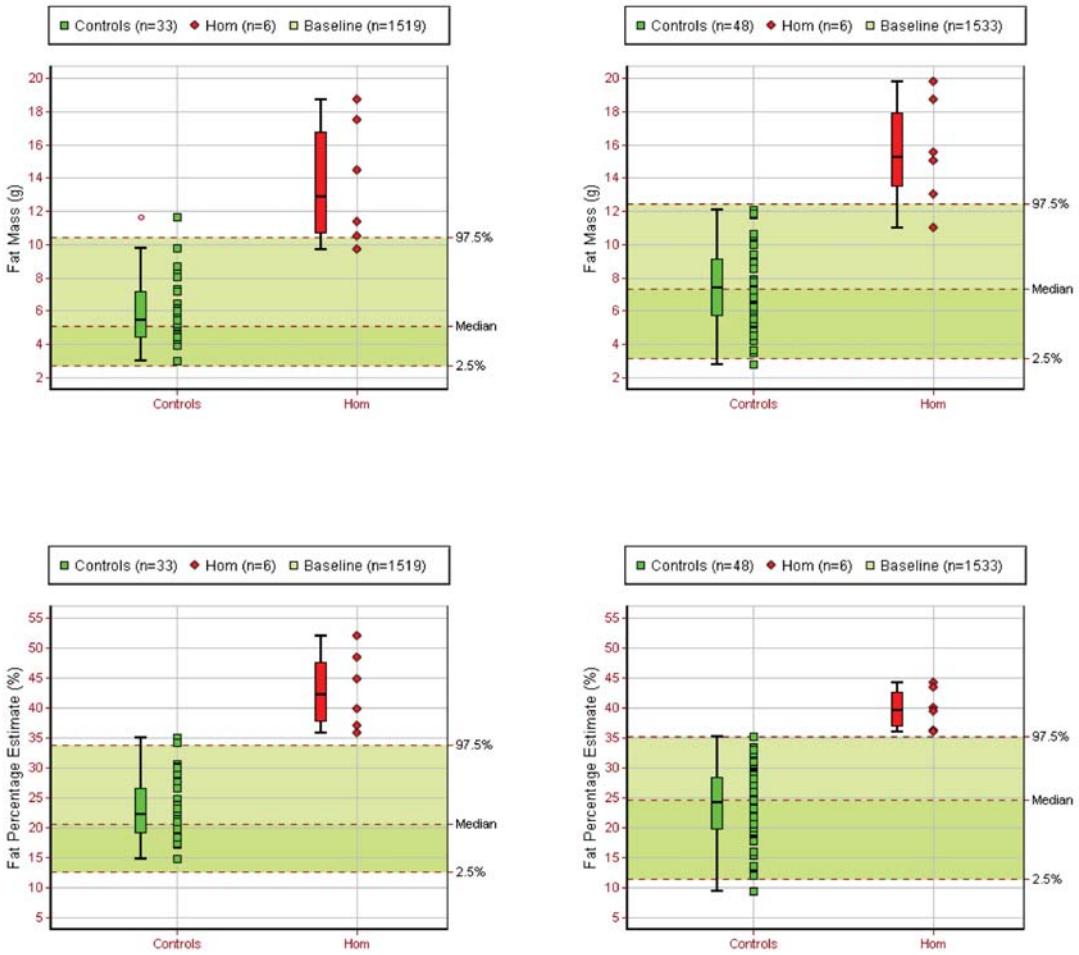

Figure 35. Increased total body fat amount in *Herc1*$^{-/-}$ mice. Top panel: Fat mass; Bottom panel: Fat percentage estimate; Left panel: Females; Right panel: Males. Body Composition was examined in anaesthetised mice using a dual energy X-ray absorptiometry machine (Lunar PIXImus II).

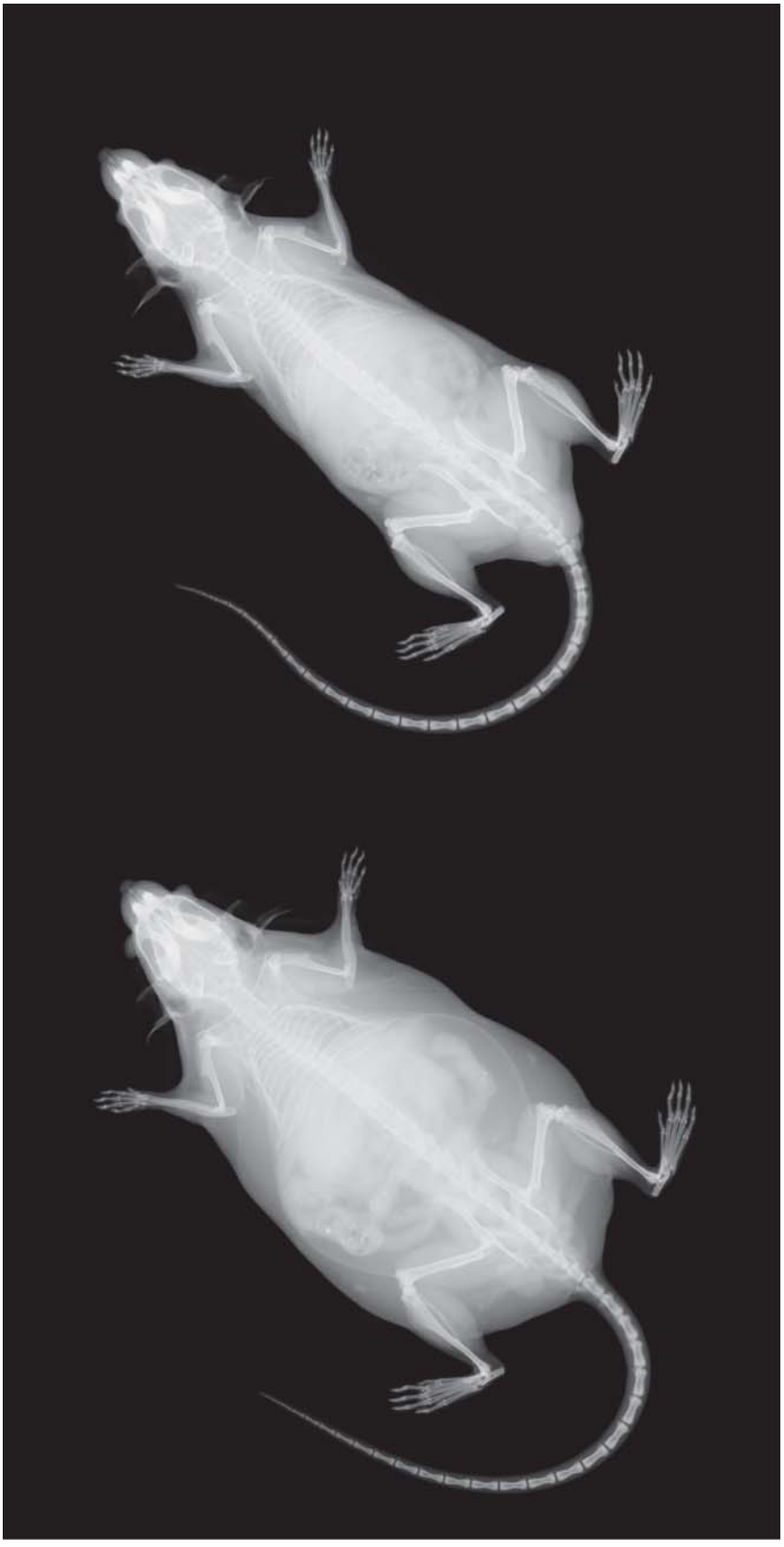

Figure 36. Body Composition X-ray imagining of a *Herc1*$^{-/-}$ mouse. Top panel: *Herc1*$^{-/-}$ male; Bottom panel: Control male. Anaesthetised mice were imaged on a dual energy X-ray absorptiometry machine (Lunar PIXImus II).

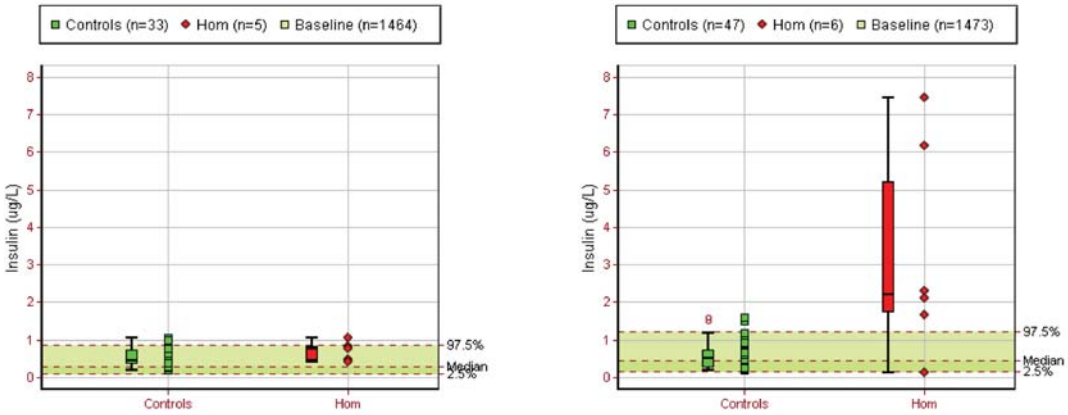

Figure 37. Increased circulating insulin level in *Herc1*$^{-/-}$ mice. Left panel: Females; Right panel: Males. Plasma insulin concentration was measured by Mesoscale Discovery (MSD) array technology in non-fasted terminally anaesthetised with ketamine/xylazine mice.

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

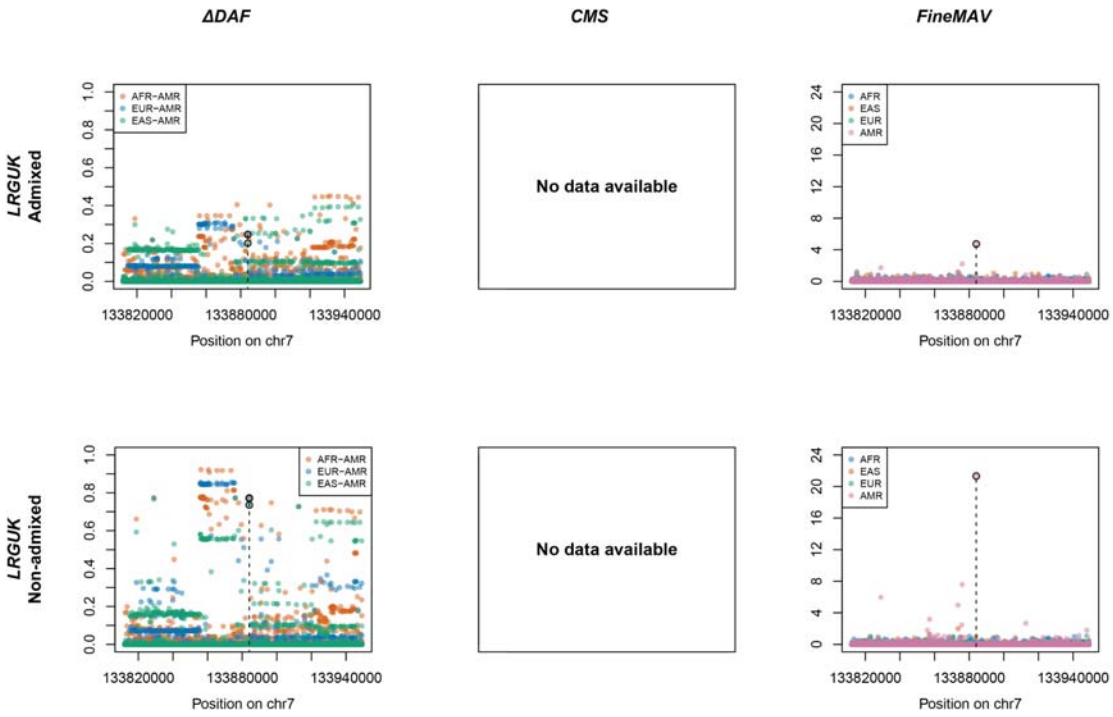

Figure 38. Signal of selection in *LRGUK* according to different approaches. *ΔDAF* and *FineMAV* scores are shown for the genomic windows spanning *LRGUK* gene. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project Phase3 dataset (142) including Admixed Americans (upper panel) and 24 Quechua 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).