4. General discussion

4.1. Summary

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

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

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

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

4.2. Next steps

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

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

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

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

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

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

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

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

4.3. Future directions

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

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

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

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

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

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

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