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

2.1. Introduction

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

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

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

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

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

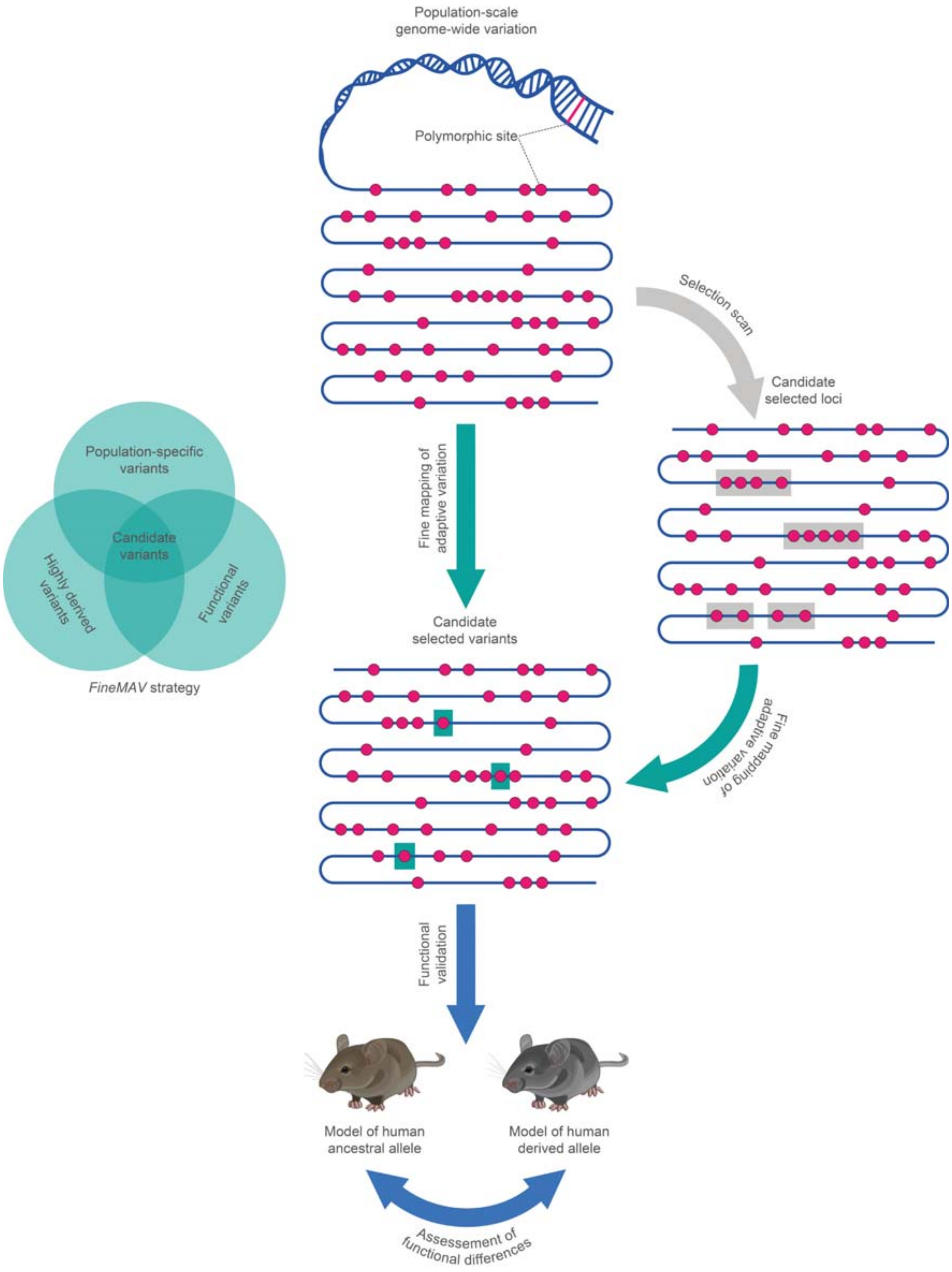

Figure 4. Workflow for prioritization of candidate variants for functional studies. The DNA molecule is represented as a blue line, with variants being red dots. Identification of the candidate causal variants from the genome-wide variation data, or the refinement of the known signal of selection to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

2.2. Meta-analysis of previous selection scans

2.2.1. Materials and Methods

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

Since one particular method of looking for evidence of selection might be more abundant in the published literature than others, its results might outweigh other methods in a simple summation of the evidence and inappropriately dominate the meta-analysis. To avoid this bias and obtain a balanced view based on all four methods, we developed a correction to control for the proportion of studies that are not independent. We first calculated a per-gene selection confidence level within each methodological category (ranging from 0 for genes not reported by any study within that category, to 1 for genes supported by all selection scans employing that detection method). We then calculated a Selection Support Index (*SSI*) by first obtaining the mean of the squares of the selection confidence levels on a per-gene basis. This would penalize genes moderately supported by several methods and promote genes strongly supported by a single approach (Equation 2). The *SSI* value was then corrected for the gene length where this strongly departed from the mean (gene length was retrieved from Ensembl (160)). The theoretical maximal *SSI* for an average-sized gene reported by all studies analysed is 1, while genes reported by all studies within one methodological category would score 0.25 (Table 2). Thus, *SSI* weighs, combines and evaluates signals of selection on a per-gene basis, starting from the results of published genome-wide selection scans of autosomal loci.

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

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

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

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

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

2.2.2. Results

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

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

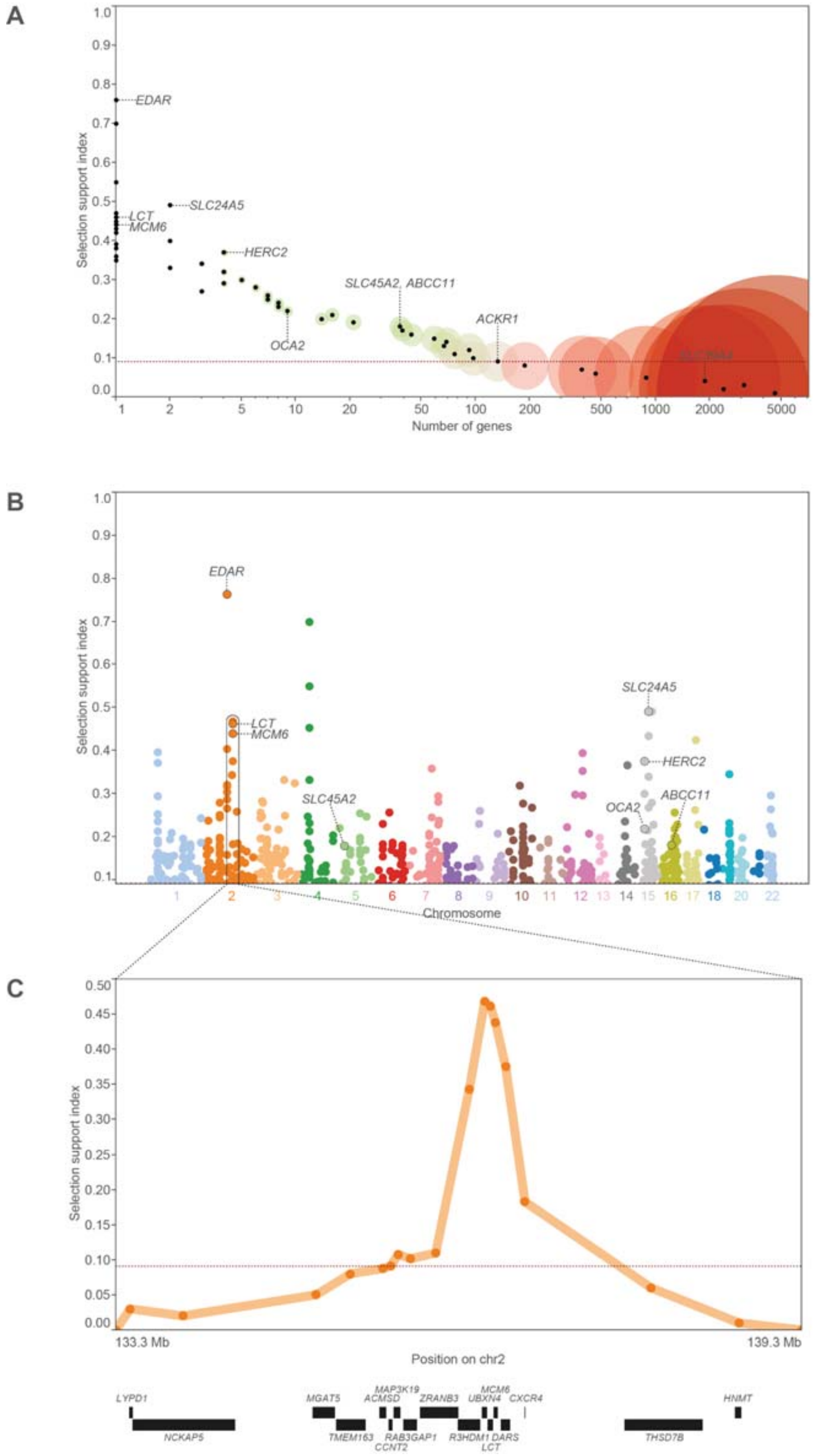

Figure 5. Meta-analysis of published genome-wide selection scans. (A) – Plot of Selection Support Index (*SSI*) scores for the positively selected genes in the published literature against the number of genes with this score; *SSI* score is also illustrated by the circle colour, and gene count by the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

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

2.2.3. Discussion

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

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

New whole-genome sequencing datasets coupled with novel methods to detect selection can outperform previous research and detect unreported candidates (as full-sequence data ensure that all potential candidate variants are evaluated). For example, the zinc uptake transporter ZIP4, known for its striking selection signature, did not show up among the top candidate genes in the meta-analysis of the published literature (Figure 5.A). ZIP4, encoded by *SLC39A4* is characterised by an extreme difference in the frequency of leucine-to-valine substitution (Leu372Val) between West Africans and Eurasians (165). The functionality of this variant was verified through *in vitro* functional experiments demonstrating differences between the human derived and ancestral alleles in surface protein expression, intracellular levels of zinc and zinc uptake (165). However, genomic scans for selection based on extended long haplotypes or deviations in the allele frequency spectrum had failed to identify ZIP4 as a candidate gene for positive selection. Such an extreme pattern of population differentiation and the absence of additional accompanying classic sweep signatures can be explained by the effect of a local recombination hotspot (165). In this scenario, *SLC39A4* should have obtained moderate support in our meta-analysis, but was missed in many studies employing population differentiation methods, as the selected SNP (or any SNP tagging it) was not included in the commonly-used Affymetrix and Illumina SNP arrays and consequently it was absent from the HGDP and Perlegen datasets (166, 167). As a result, *SLC39A4* was very weakly supported in our meta-analysis (Figure 5.A).

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

2.3. Fine-Mapping of Adaptive Variation

2.3.1. Materials and methods

2.3.1.1. *FineMAV*

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

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

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

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

2.3.1.2. Measure of population differentiation

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

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

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

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

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

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

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

2.3.1.3. Measure of allele prevalence

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

2.3.1.4. Measure of functionality

It is crucial that variant-level functional inferences are based on whole-genome level measures to ensure that all potentially selected variants are treated equally. We wanted a measure of functionality to be allele-specific and applicable to all variation, both coding and non-coding, since many signals of selection localise in regulatory elements or intergenic regions (17, 123). As proteins are usually involved in many processes through complicated interaction pathways with other proteins, amino acid change in one protein may affect many diverse traits i.e. pleiotropic phenotypes (138). In general, pleiotropic changes are thought to be disadvantageous (170), thus it is believed that a great deal of human phenotypic variation is based in regulatory variation (17, 140, 170-172). However, having a different set of annotations for coding and noncoding variation makes it challenging to compare these distinct variant categories. Thus consensus methods combining multiple annotations, each with its own weaknesses, are especially needed here for functional prioritization of variants across many functional categories (168). In our analysis we used the Combined Annotation-Dependent Depletion (*CADD* v1.2 PHRED-scaled C-score), which integrates 63 diverse genome annotations into a single measure for each variant and in theory takes a value between 0 and 99 (168).

2.3.1.5. *FineMAV* calibration

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

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

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

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

2.3.1.6. *FineMAV* calculation in 1000 Genomes Project

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

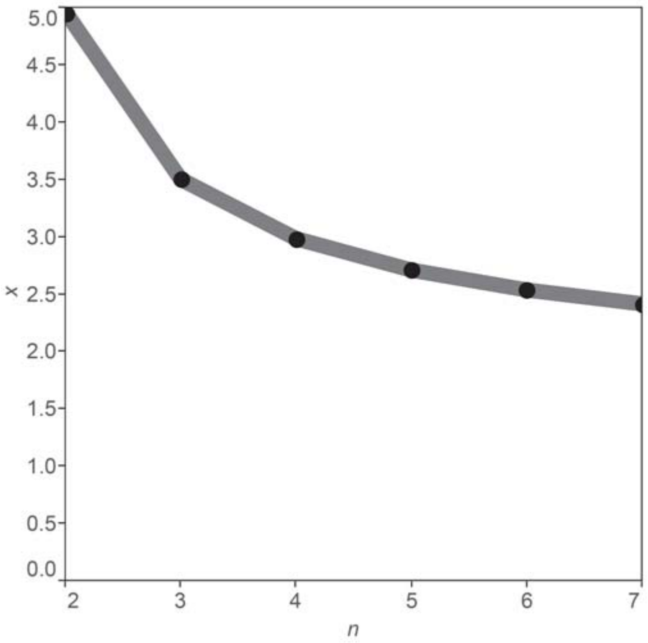

Figure 6. Recommended minimal values of *x* for given *n*. *x* – penalty parameter. *n* – number of populations being compared.

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

2.3.1.7. Simulation analysis

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

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

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

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

2.3.2. Results

2.3.2.1. *FineMAV* power analyses using simulation

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

2.3.2.2. *FineMAV* evaluation using 1000 Genomes Project

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

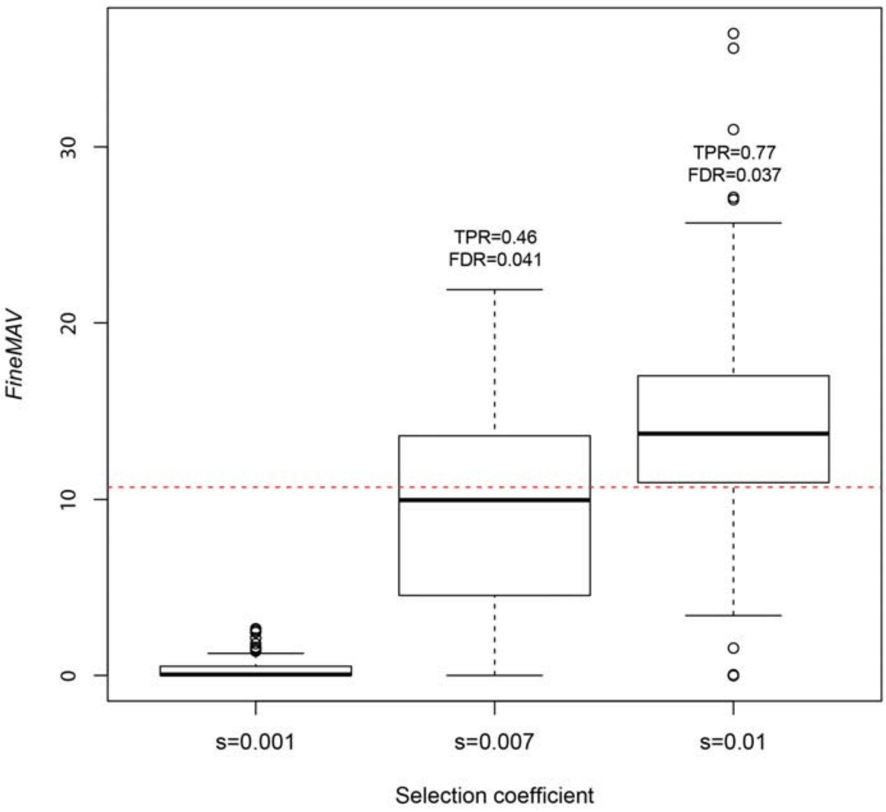

Figure 7. Simulated distribution of *FineMAV* scores for variants under selection. Three selection scenarios of varying selection strength were plotted: *s* = 0.001, *s* = 0.007 and *s* = 0.01. Distributions of *FineMAV* values for selected variants from each scenario are shown as box-plots. The red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

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

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

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

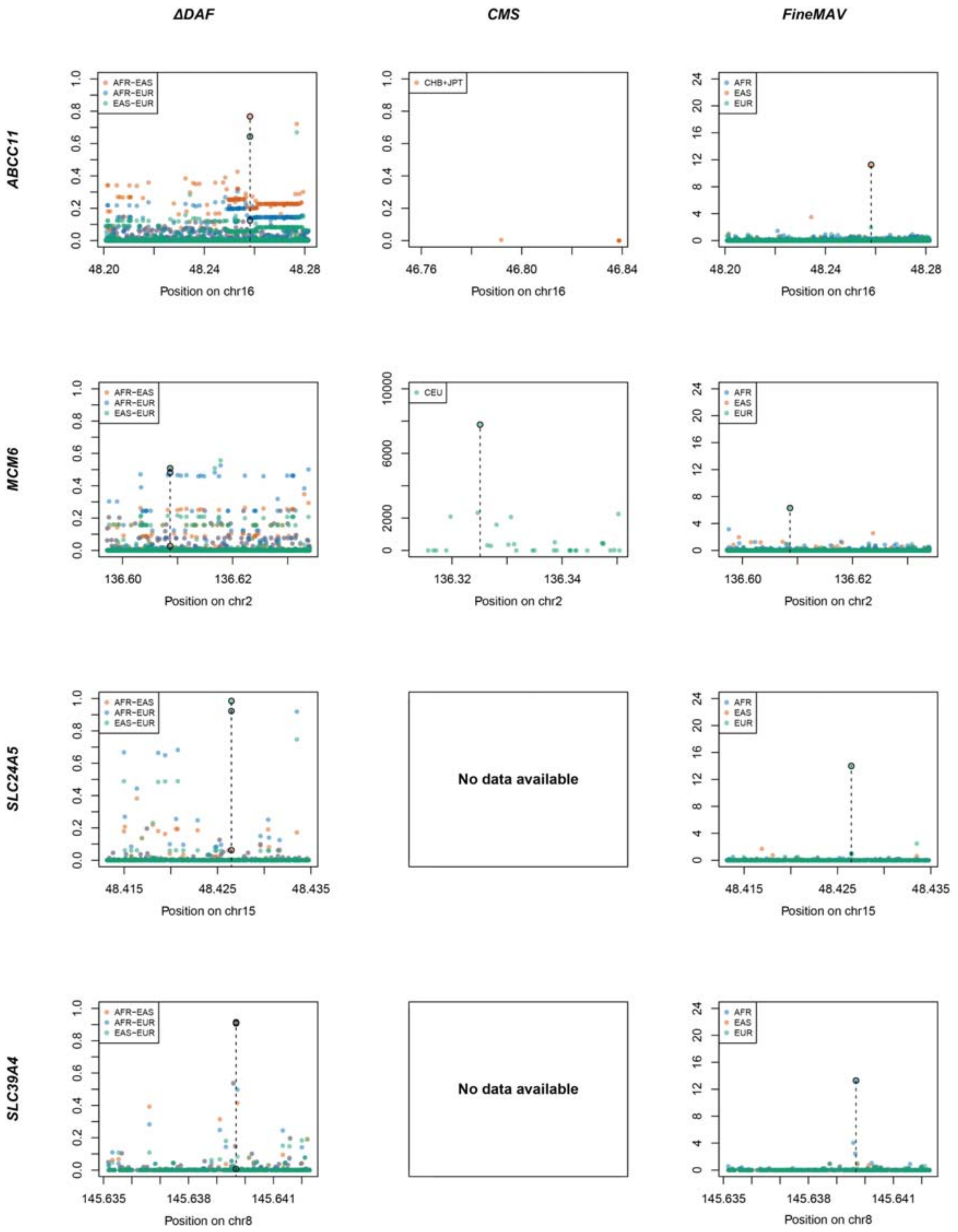

Figure 8. Comparison of three different approaches for pinpointing selected variants in the calibration set. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes from the gold standard calibration panel. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project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

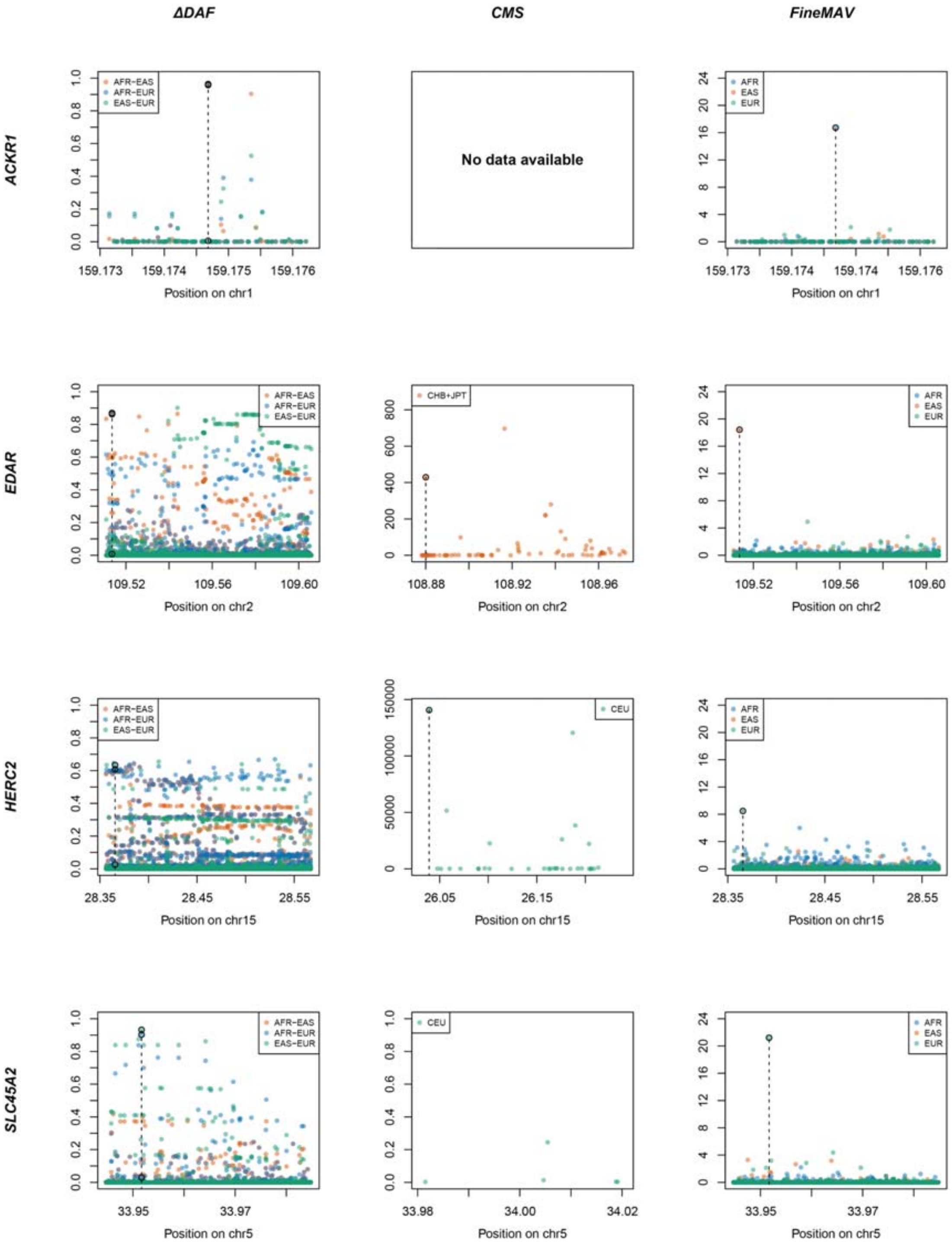

Figure 9. Comparison of three different approaches for pinpointing selected variants in the replication set. *ΔDAF*, *CMS* and *FineMAV* scores are shown for the genomic windows spanning genes from the gold standard replication panel. *ΔDAF* and *FineMAV* were calculated from the 1000 Genomes Project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

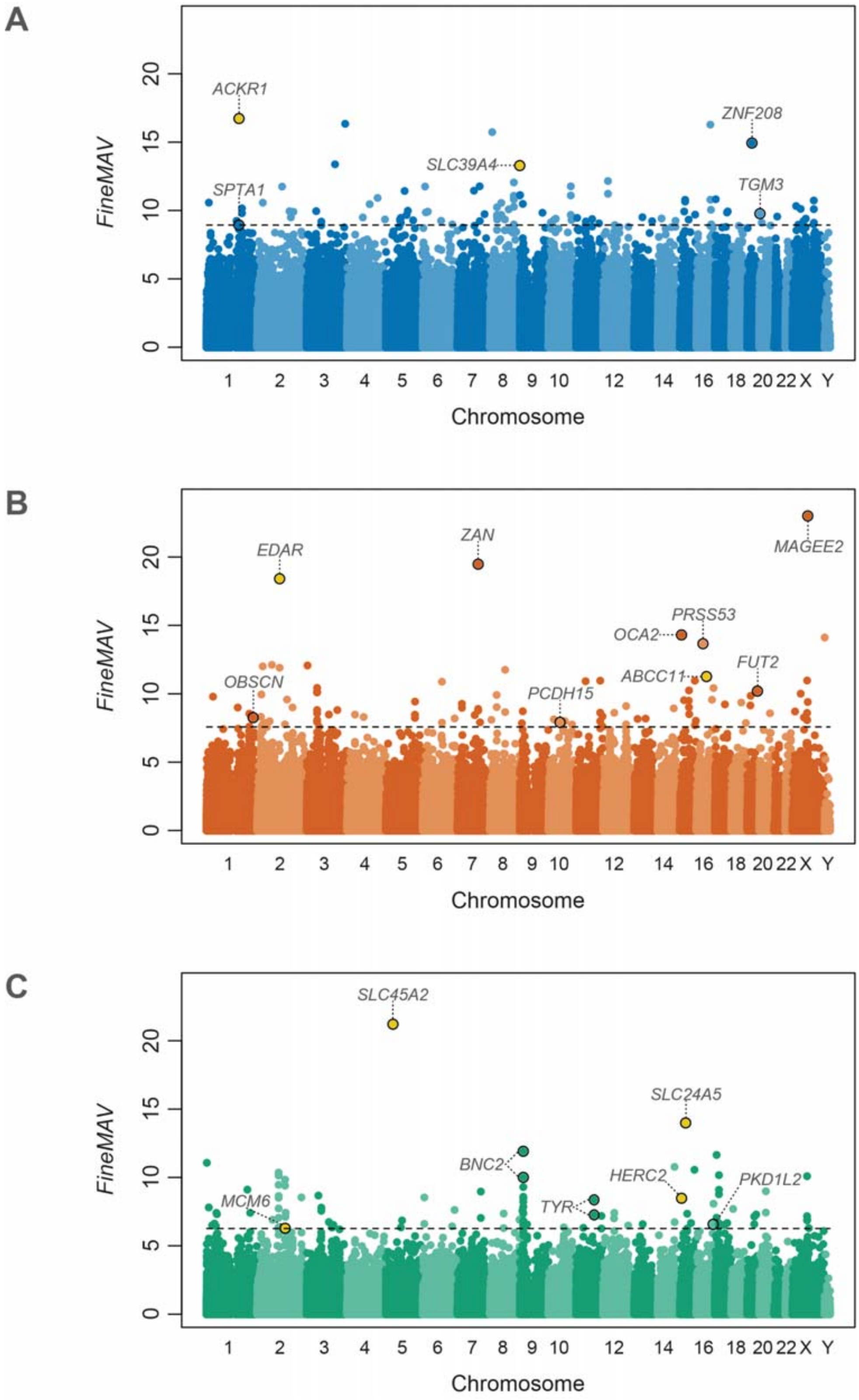

Figure 10. Manhattan plot of genome-wide *FineMAV* scores. *FineMAV* scores were calculated for genome-wide SNPs from 1000 Genomes Project Phase 3 (142) in three populations: (A) – Africans (AFR, blue); (B) – East Asians (EAS, orange); (C) – Europeans (EUR, green). Each dot in the Manhattan plots represents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

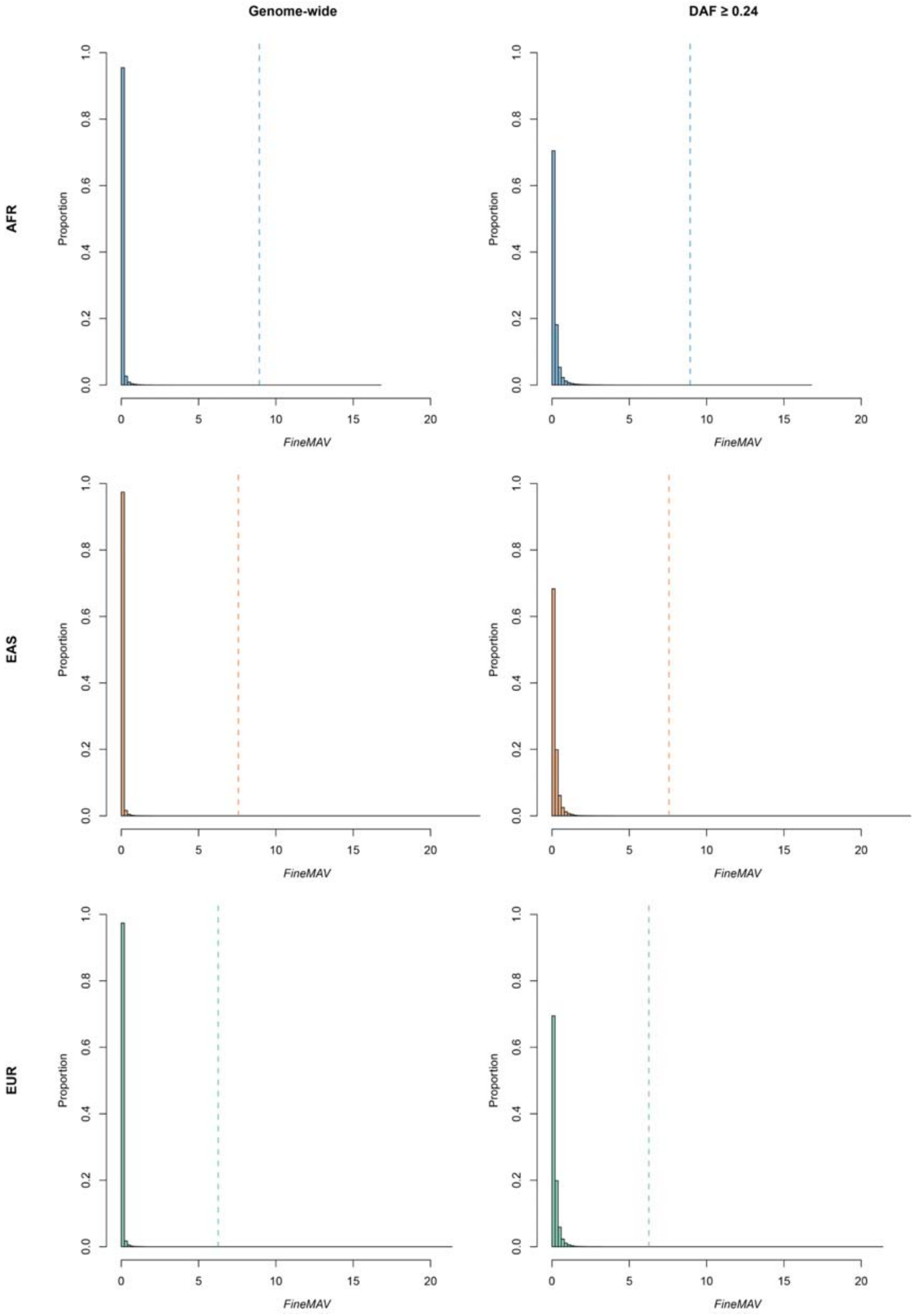

Figure 11. Distribution of *FineMAV* scores. Left: genome-wide distribution of *FineMAV* scores in each population. Right: *FineMAV* score distribution of variants matching continental derived allele frequency of our top outliers (*DAF* ≥ 0.24). Dashed vertical lines indicate *FineMAV* cutoffs to include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

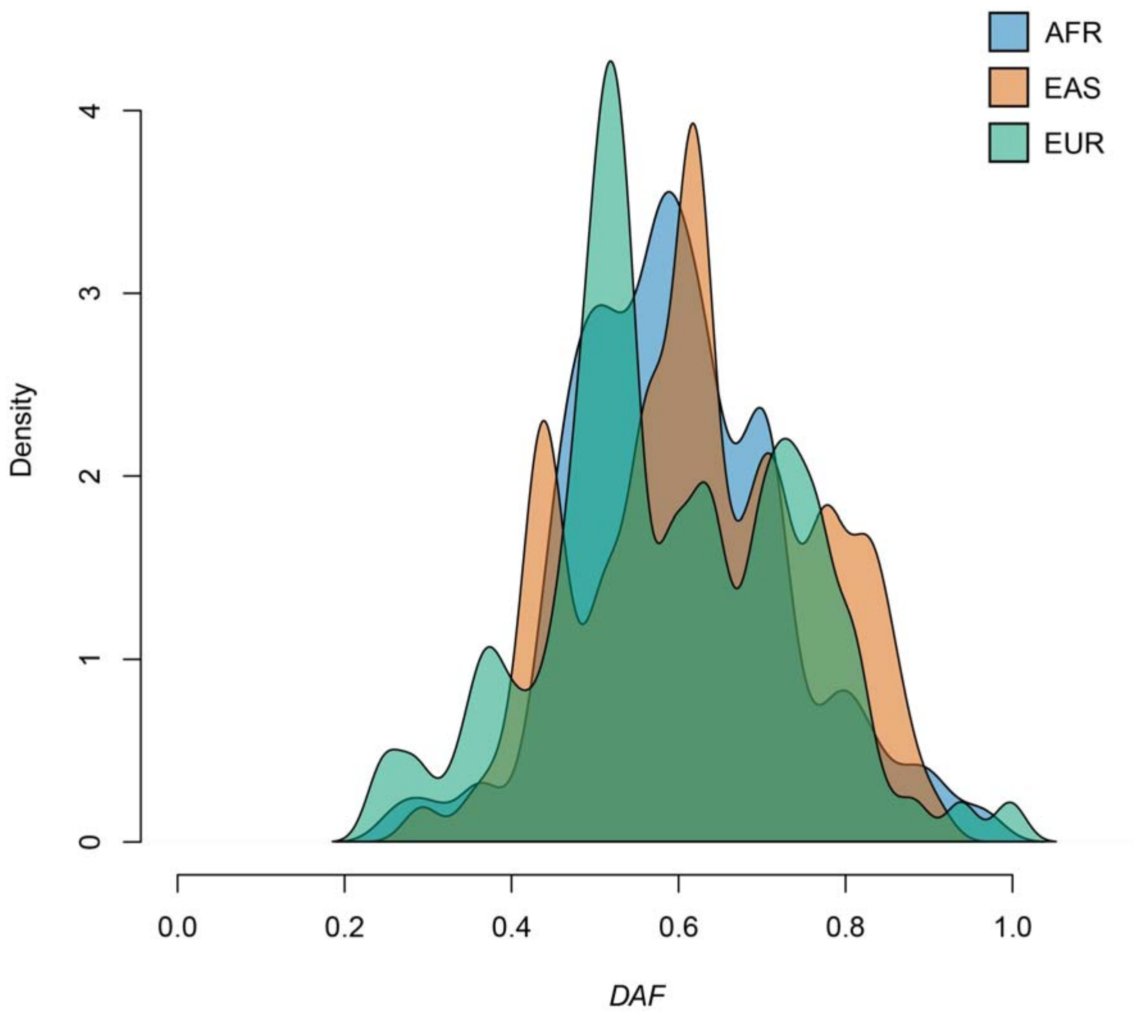

Figure 12. Derived allele frequency distribution among the top 100 *FineMAV* hits within each population.

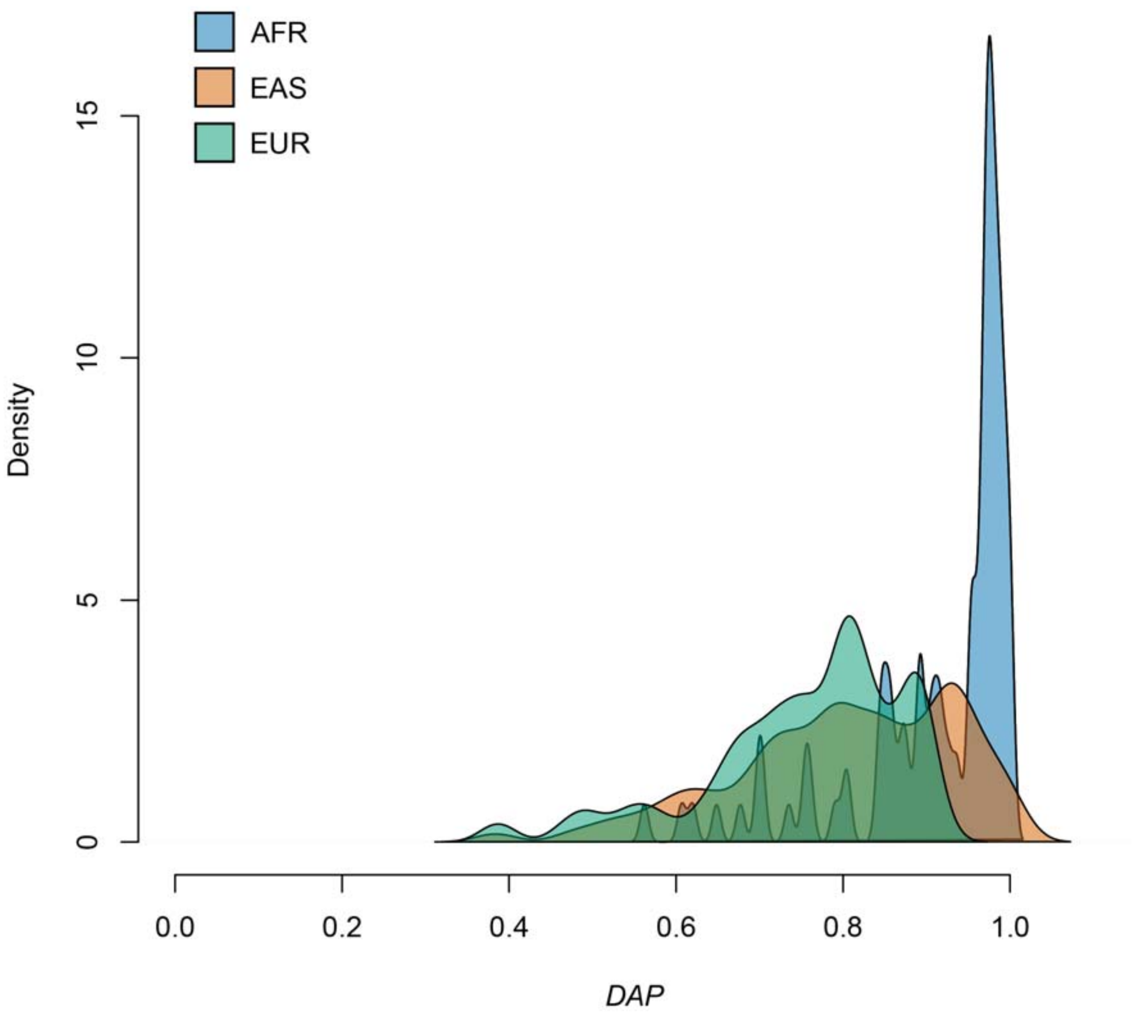

Figure 13. Derived allele purity distribution among the top 100 *FineMAV* hits within each population.

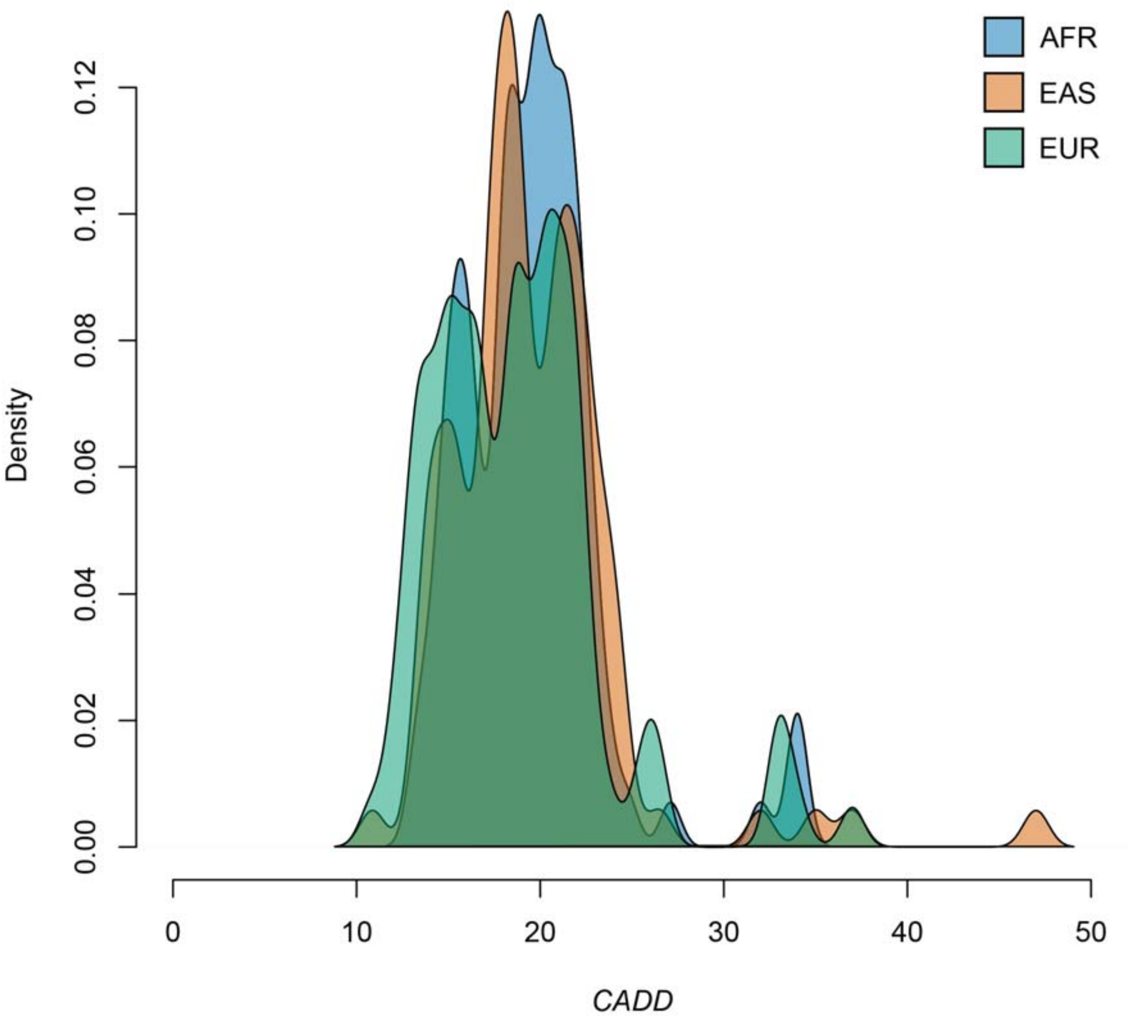

Figure 14. CADD score distribution among the top 100 *FineMAV* hits within each population.

2.3.2.2.1. Top *FineMAV* hits classification and enrichment analysis

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

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

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

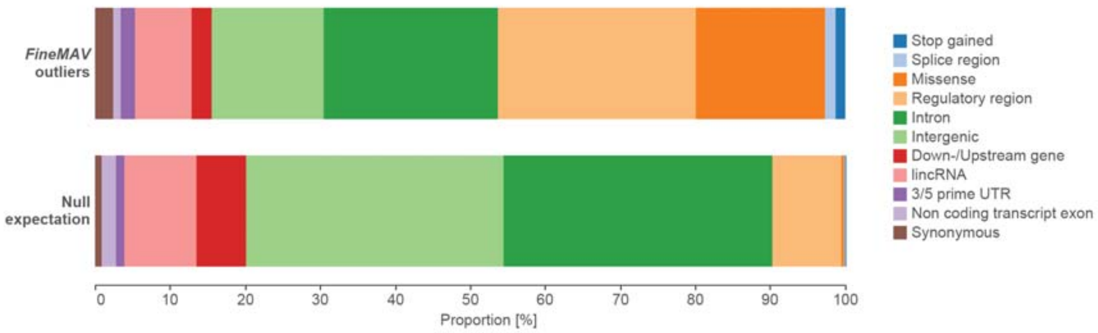

Figure 15. Functional consequences of *FineMAV* top outliers as compared to random expectation. The 100 top outliers from each population (AFR, EAS, EUR) were pooled together. The chart uses the most severe predicted consequence for each variant from Ensembl (160)

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

2.3.2.2.2. Functional validation *in silico*

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

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

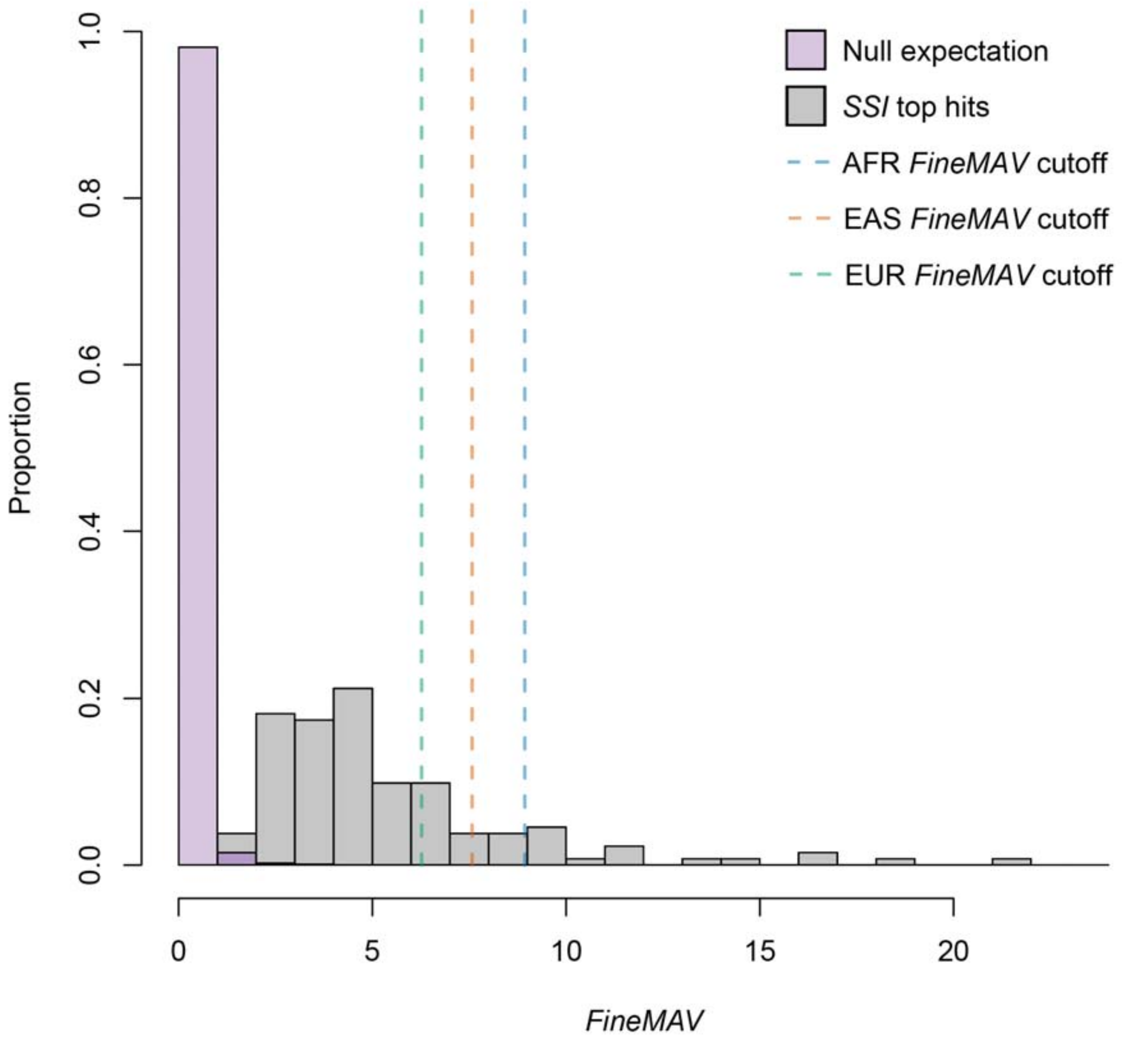

Figure 16. Distribution of *FineMAV* scores in *SSI* outlier genes. The null expectation is the distribution of *FineMAV* scores of variants matching the continental derived allele frequency of our top outliers (*DAF* ≥ 0.24) across all three populations pooled together. We then looked at the distribution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

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

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

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

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

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

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

2.3.2.3.1. Nonsense variants

We observed some high-scoring nonsense variants among our top candidates, suggesting pseudogenization of *PKD1L2* (an endogenous fatty acid synthase in skeletal muscle (197)) in Europeans, *ZNF208* (zinc finger and SRY-interacting protein (198)) in Africans, as well as *ZAN*, *OBSCN* (sacromeric signaling protein involved in myofibrillogenesis (199)) and *MAGEE2* (melanoma-associated antigen expressed in the brain (200)) in East Asians. Mice homozygous for knockout alleles of *OBSCN* and *ZAN* are viable and fertile (201, 202); *ZAN* is particularly interesting as it encodes a zonadhesin protein located in the acrosome that mediates the species specificity of sperm binding to the extracellular coat of the egg (zona pellucida) (203). Sperm from zonadhesin-null mice exhibit dramatically higher levels of inter-species gamete adhesion without alteration in fertility (202). Zonadhesin is reported to be a rapidly-evolving protein with a high level of divergence between closely-related species, but is similar in species capable of interbreeding (204, 205). The adaptive advantage of species specificity conferred by zonadhesin might be the limitation of cross-species fertilization and avoidance of sterile hybrids (205). However, polymorphism data in humans reveal a signature of positive selection on haplotypes carrying a frameshift mutation (204). We find a signal of selection at a nonsense mutation (rs2293766) present at 51% frequency in East Asians, but virtually absent elsewhere. An even higher frequency difference is observed for a stop allele at rs1343879 in *MAGEE2* on the X chromosome. Selection at this locus was previously reported by Yngvadottir *et al*., who observed lower diversity in haplotypes carrying the stop allele than in the others and concluded that, like *ZAN,* the truncated MAGEE2 conferred a selective advantage in East Asia (206).

2.3.2.3.2. Missense variants

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

Another novel signal detected in African populations falls in *SPTA1*, encoding erythrocytic spectrin, alpha 1, a principal component of the erythrocyte membrane skeleton, which is essential for the arrangement of transmembrane proteins, determining red cell membrane stability, cell shape and deformability (216-218). Variants in *SPTA1* have been associated with quantitative hematologic traits (219-221), and those causing its deficiency result in hemolytic anemias characterised by elliptically shaped erythrocytes (also seen in *Spta1*$^{-/-}$ null-mice) (222, 223). The high prevalence of such anemia in Africa (10 times higher in West Africa than in Europe or USA (224)) raised the question of a selective advantage, possibly contributing to protection against malaria (225, 226). It has been shown that decreased spectrin level inhibited malaria parasite growth *in vitro* (227) and in a mouse model (228). This evidence suggests that a functionally and structurally normal host membrane is necessary for parasite growth and development (225, 227). *FineMAV* pinpointed rs7547313 (Ile>Val) as a likely selected variant present at 0.37 frequency in Africans but absent elsewhere. Furthermore, this variant was reported to be an eQTL associated with lower expression of *ACKR1* [MIM: 613665] (also known as *DARC*); p-value = 0.000017 (200). It is worth saying that rs7547313 is not in LD with the known Duffy O allele (rs2814778); r^2=0.000228497. However, the functional effect of this missense variant on the protein level and malaria parasite growth remains uncertain.

2.3.2.3.3. Regulatory variants

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

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

2.3.3. Discussion

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

FineMAV was calibrated and tested using a gold standard panel of the eight best examples of experimentally-validated causal variants underlying signals of positive selection in humans, and was able to identify the known functional candidate in all cases (Figure 8 and Figure 9). Using the complete 1000 Genomes Project dataset (142), we then ranked all genome-wide SNPs based on their *FineMAV* value and identified extreme outliers in the upper tail of the empirical genome-wide distribution in Africa, Europe and East Asia. *FineMAV* rediscovered other known variants with strong prior evidence for being causal of positive selection signals, but which were not part of the positive control set which provides additional support for our method. We also identified potential functional variants in other genes reported to be under strong positive selection in the literature (with strong *SSI* score) where the causal mutation has not been confirmed yet, including *LPP*, *PCDH15* and *PRSS53*. The selection signal in *PCDH15* and *PRSS53* was attributed to a single missense variant (rs4935502 and rs11150606 respectively), replicating the results obtained by *CMS* (155, 196).

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

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

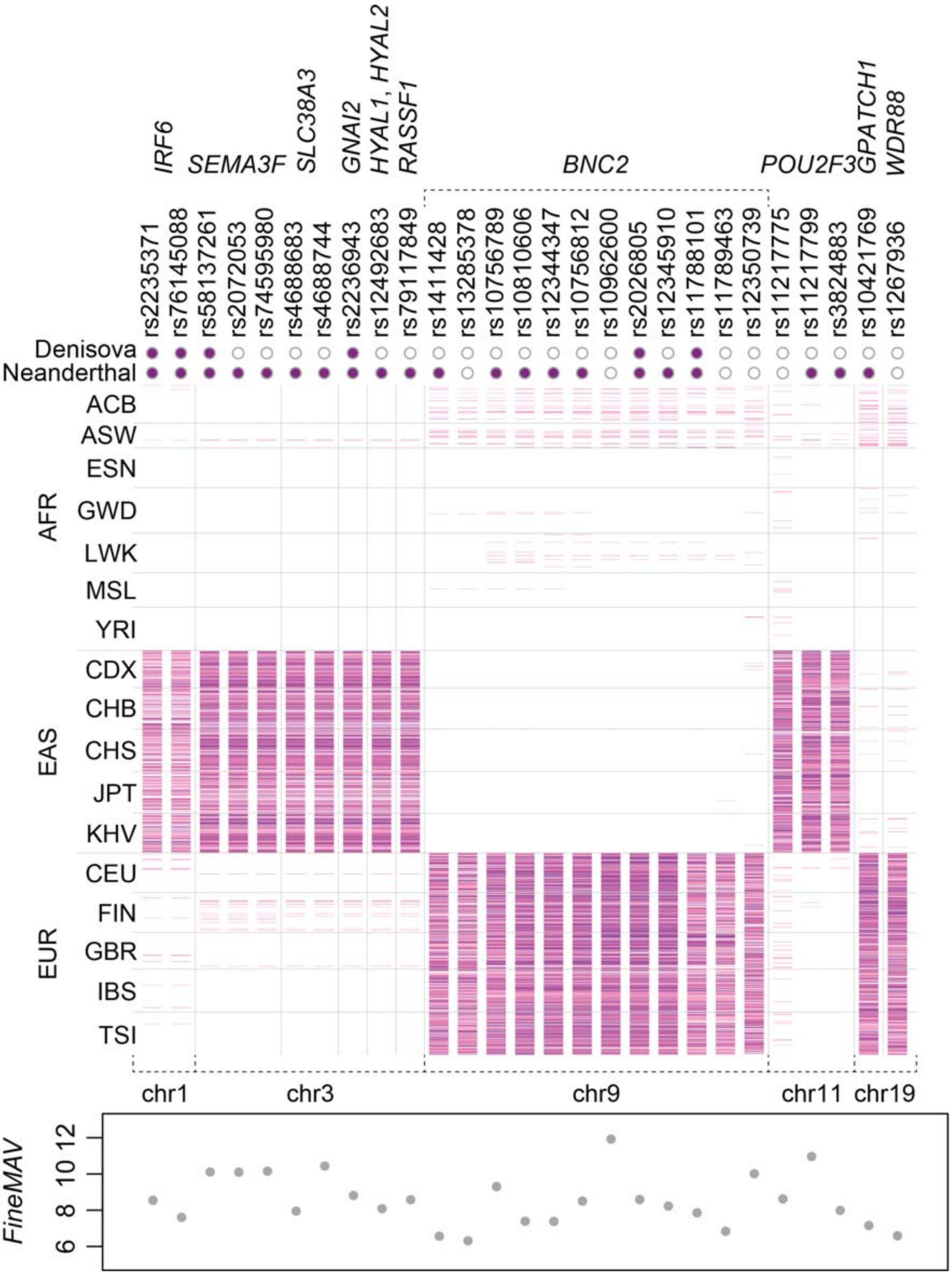

Figure 17. Genotypes of putatively introgressed SNPs identified by *FineMAV*. Rows represent individuals from Phase 3 of the 1000 Genomes Project (142) grouped by population, columns indicate variant sites picked up by *FineMAV* and falling in regions putatively introgressed from archaic hominins (30,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

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

2.4. *FineMAV* application to various populations

2.4.1. Materials and methods

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

2.4.1.1. Admixed Americans and South Asians

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

2.4.1.2. Non-admixed Native Americans

We searched for local adaptations in non-admixed Native Americans (nAMR) using a dataset comprised of unpublished low coverage whole-genome sequences from 24 Quechua from Peru generated at WTSI. In total, 29 Quechua were sequenced on either an Illumina Genome Analyzer II using 108 bp paired-end reads or HiSeq 2000 with 100 bp paired-end reads with insert size of 300-500 bp. Reads were aligned to GRCh37 (hg19/NCBI37) for general sequencing QC and yielded average coverage of 4-6x. The 29 BAMs were then merged with a subset of the 1000 Genomes Phase 1 and 2 samples using the varpipe tool. Variants and genotypes were called in the merged dataset by Luca Pagani and Petr Danecek (Wellcome Trust Sanger Institute) using Samtools and the procedure described in (247). Samples showing more than 5% European ancestry in ADMIXTURE analysis using common variants were excluded from subsequent analysis leaving a total of 24 individuals. *FineMAV* analysis in nAMR were performed using 3 reference populations from the subset of 1000 Genomes Project: AFR (Americans of African Ancestry, Southwest USA [AWS], Luhya in Webuye, Kenya [LWK], Yoruba in Ibadan, Nigeria [YRI]), EAS (Han Chinese in Bejing, China [CHB], Southern Han Chinese [CHS]), EUR (Utah Residents with European Ancestry, USA [CEU], Iberian Population in Spain [IBS], Toscani in Italia [TSI]); n = 4; x = 2.98. *DAF*, *DAP* and *FineMAV* values were calculated as described earlier. Common variants failing Hardy–Weinberg equilibrium and not called in 1000 Genomes Project, Phase 3 data release (142) were excluded.

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

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

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

2.4.2. Results

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

2.4.2.1.1. AMR and SAS from 1000 Genomes Project

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

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

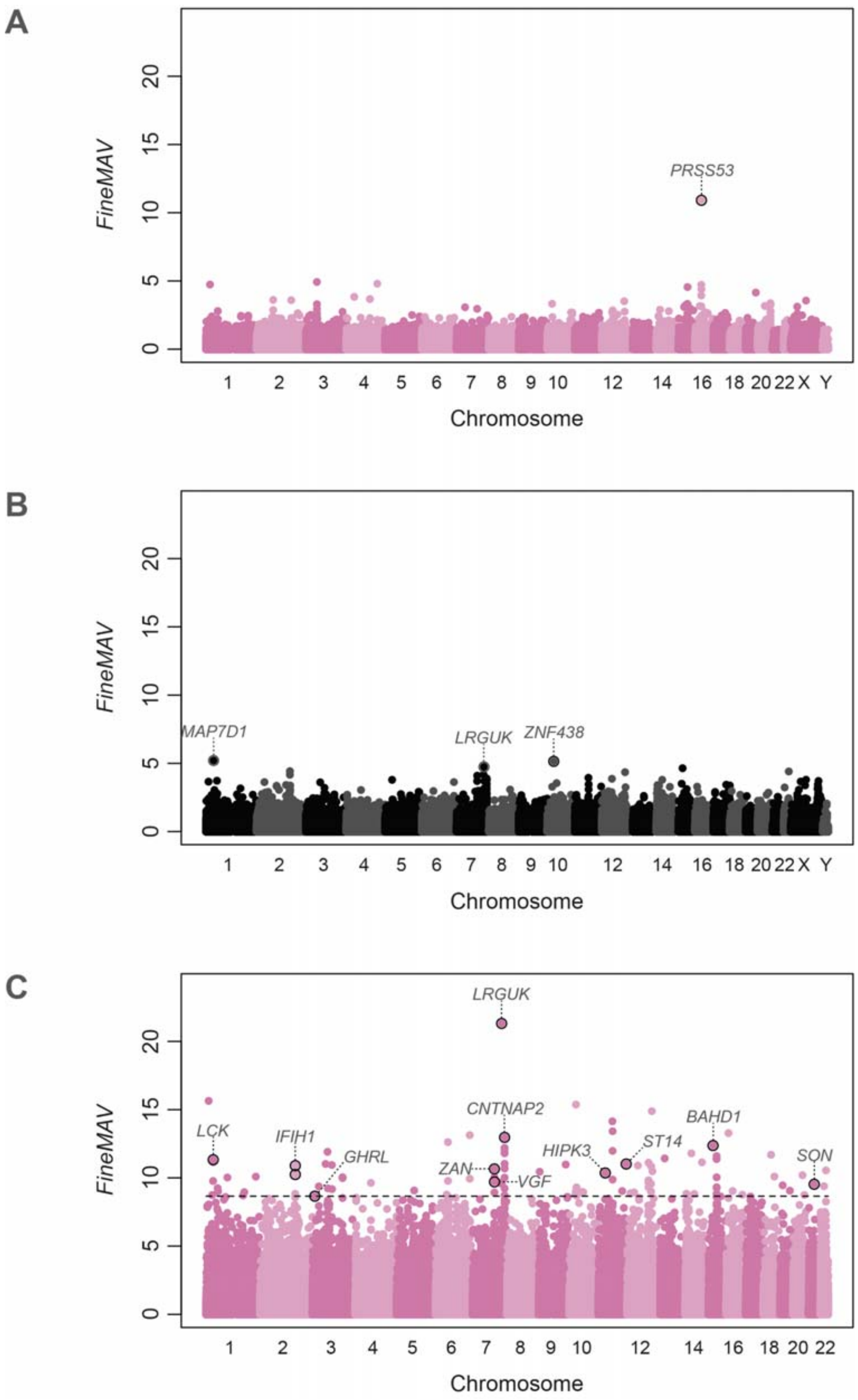

Figure 18. Manhattan plot of genome-wide *FineMAV* scores in Native Americans and South Asians. *FineMAV* scores calculated for genome-wide SNPs in: (A) – South Asians (SAS); (B) – Admixed Americans (AMR); (C) – Non-admixed Native Americans (nAMR; the threshold (dashed line) was set to include the top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

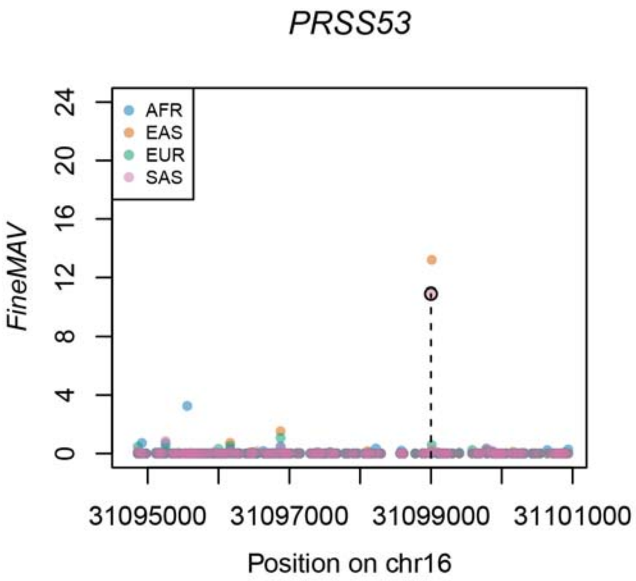

Figure 19. Signal of selection in the *PRSS53*. *FineMAV* scores of variants in the genomic window spanning *PRSS53* are plotted as dots. Genomic positions are given in bp according to GRCh37. The selected variant in South Asians (rs201075024) is marked with a dashed line with the variant selected in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

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

2.4.2.1.2. Non-admixed Native Americans

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

Other interesting variants include missense rs62621285 in *ST14* and a stop gained rs2293766 in *ZAN*, present at 56% and 79% respectively. This nonsense mutation in *ZAN* (involved in sperm species specificity (202, 203)) has been introduced in the previous section as one of the top variants selected in East Asians, yet its frequency in Native Americans is even higher. *ST14* is known for playing an important role in hair development and growth and its deficiency in mice causes brittle, thin, uneven, and sparse hair, or even a complete absence of erupted pelage hairs and vibrissae in null animals (249-254), which is interesting considering the reduced body hair in Native American populations (255, 256). Furthermore, *ST14* is required for skin keratinization, formation and maintenance of the epithelial and epidermal barrier and integrity (250, 253, 254, 257-263). It seems that this gene has pleiotropic functions affecting the development of the epidermis, hair follicles, and cellular immune system (254) as it has been shown that the *ST14* protein product (matriptase) is also an influenza virus-activating protease supporting multicycle viral replication in the human respiratory epithelium (264-266). The influenza genome does not encode any proteases and relies on host proteases for the cleavage activation of the surface receptor proteins in order to fuse with the host cell membrane (264-266). Knockdown of matriptase in human bronchial epithelial cells significantly blocked influenza virus H1 subtype replication (264-266).

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

Furthermore, the *SON* protein product was shown to regulate ghrelin receptor (*GHSR*) transcription in the brain by repressing its promoter activity (300). Ghrelin (encoded by *GHRL* and acting via GHSR) is a pleiotropic hormone secreted by the stomach that promotes food intake, weight gain and fat storage by reducing fat utilization (beta-oxidation), but also decreased glucose tolerance and decreased insulin sensitivity in mice and rats (301-305). Knockout mice display increased utilization of fat as an energy source on a high fat diet, reduced food intake, weight gain and adiposity, increased energy expenditure and locomotor activity, decreased circulating glucose level, improved glucose tolerance, increased circulating insulin level and secretion (304, 306, 307). It seems that the absence of ghrelin protects from diet-induced obesity and type 2 diabetes (306, 307). On the other hand, the ghrelin circulating level was shown to increase during fasting and it was suggested that it prolongs survival in starved humans but may also play a role in fetal adaptation to intrauterine malnutrition, while its absence impairs fasting tolerance (301, 302, 305, 308-310). It seems that ghrelin plays an important role in the metabolic adaptation to nutrient availability and determines the type of substrate (fat or carbohydrate) that is used for maintenance of energy balance (304, 306). Interestingly, one of the high-scoring variants in the Quechua population is a missense variant (rs4684677) falling in *GHRL*. *GHRL* encodes preproghrelin, which is a precursor of two peptides ghrelin and obestatin. Obestatin is ghrelin's antagonist involved in satiety and decreased appetite contributing to decreased body weight gain (311) and the variant we picked up (Gln to Leu substitution in position 90 of the ghrelin/obestatin prepropeptide; rs4684677) was shown to impact obestatin function. Gln90Leu was slightly more efficient than native obestatin in inhibiting ghrelin-induced food intake (312).

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

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

2.4.2.2. GLEE

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

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

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

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

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

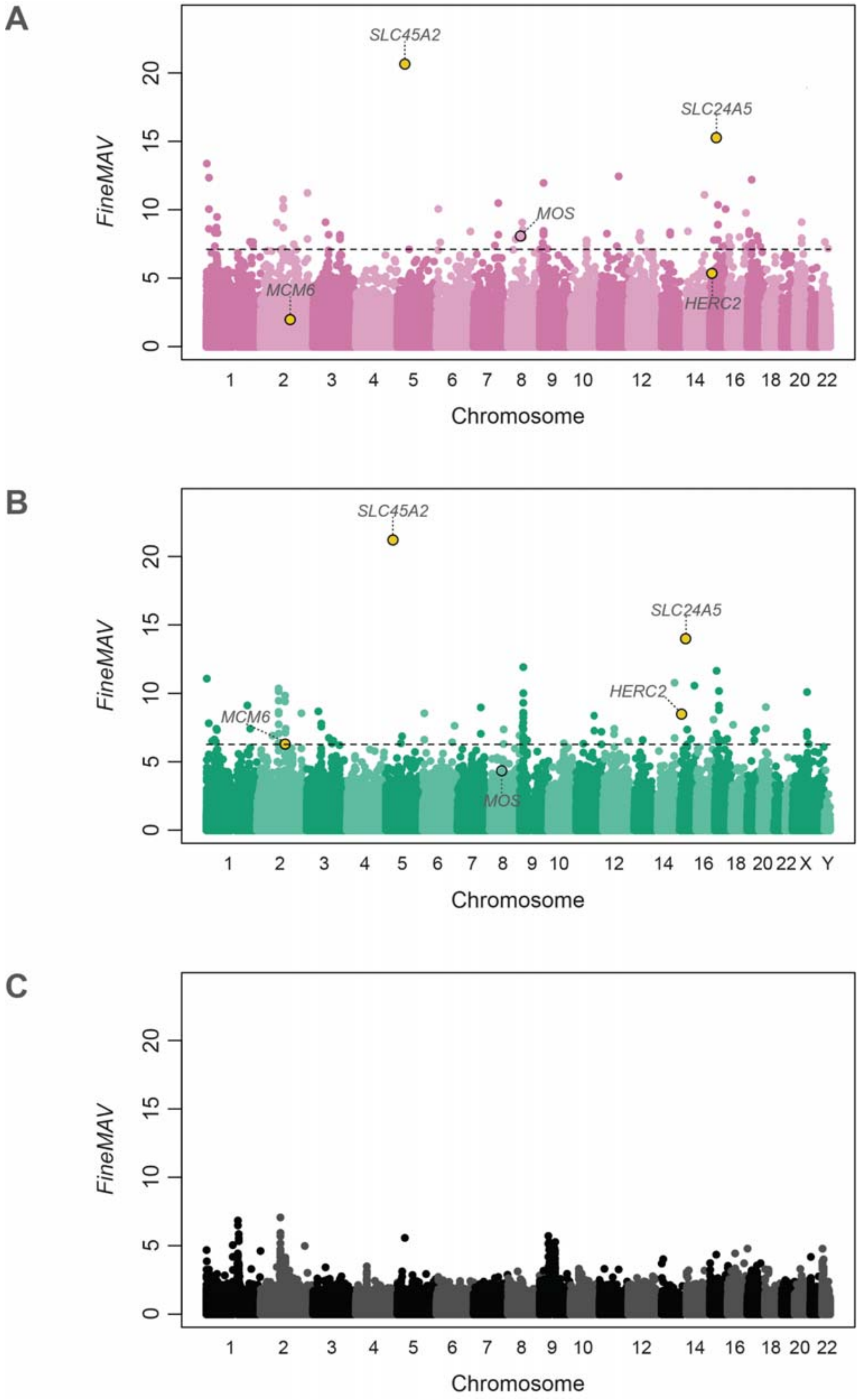

Figure 20. Manhattan plot of genome-wide *FineMAV* scores in Greeks. *FineMAV* scores calculated for genome-wide SNPs in: (A) – GRK (run together with CHB and YRI); (B) – EUR from the full 1000 Genomes Project Phase 3 calculated in the previous section; (C) – GRK (run together with CEU, CHB and YRI).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

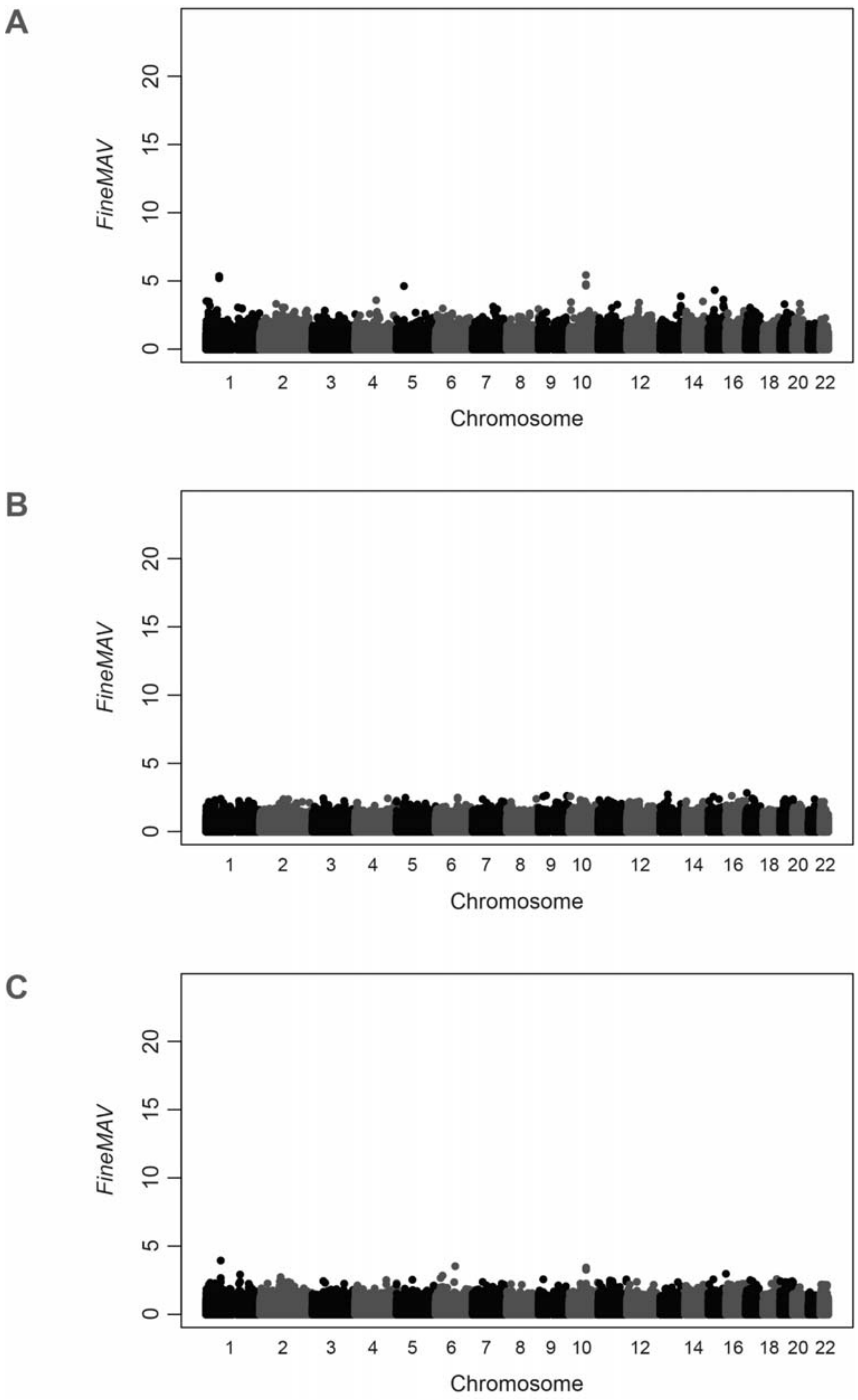

Figure 21. Manhattan plot of genome-wide *FineMAV* scores in Lebanese. *FineMAV* scores calculated for genome-wide SNPs in: (A) – LEB general population (run together with CEU, CHB and YRI); (B) – LEB Christians (run against LEB Muslims); (C) – LEB Muslims (run against LEB Christians). Each dot in the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

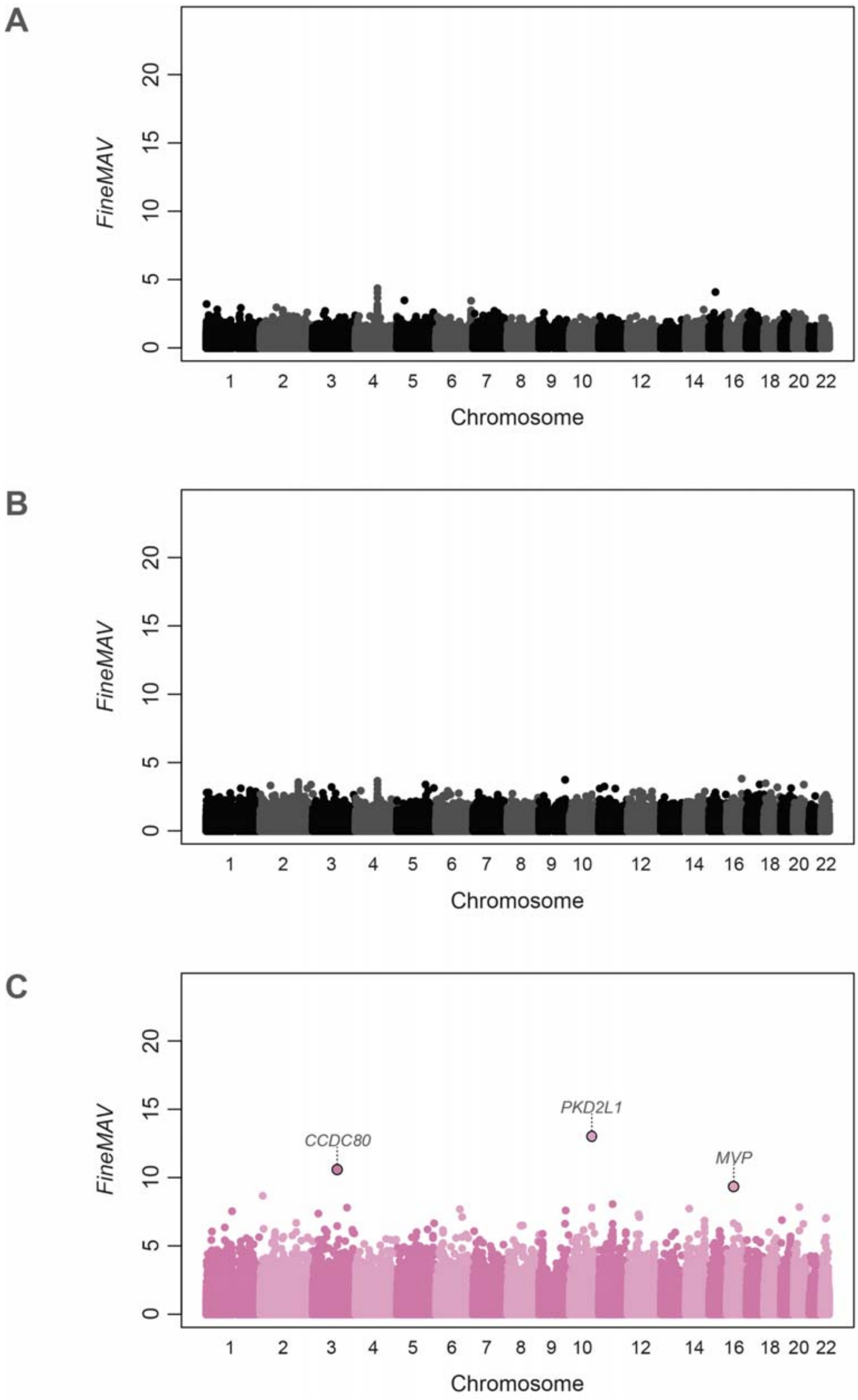

Figure 22. Manhattan plot of genome-wide *FineMAV* scores in Egyptians and Ethiopians. *FineMAV* scores calculated for genome-wide SNPs in: (A) – EGP (run together with CEU, CHB and YRI); (B) – ETP (run together with CEU, CHB and YRI); (C) – Gumuz (run together with CEU, CHB and YRI). Each dot in the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

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

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

2.4.3. Discussion

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

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

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

Finally, we found a signal of geographically restricted selection in energy metabolism genes in both Quechua and Gumuz. Widespread obesity and an elevated risk of developing type 2 diabetes and cardiovascular diseases have been reported for many indigenous communities including Native Americans (363-365). Such an observation has been linked to the so-called 'thrifty gene' hypothesis suggesting that decreased resting metabolic rate and increased energy storage was favoured in populations historically facing feast-famine cycles (44, 45, 366, 367). Adaptation to food scarcity may predispose to metabolic syndrome in a non-traditional lifestyle with continuous food supply (44, 45, 366, 367). Furthermore, the traditional diet of aboriginal Americans and Ethiopians was estimated to be high in carbohydrates (~70-80%) and low in fat (8-12%), while adoption of a modern lifestyle resulted in much higher fat-intake (35% fat) (368-370). Urban Peruvians and rural-to-urban migrants showed a higher prevalence of obesity and cardiovascular diseases compared with the rural population (although environmental factors play an important role) (371-373), and a general high incidence of hypertension and obesity was reported in Peru among both cosmopolitan and Andean Peruvians (374-381) with nearly a quarter of the adult population at an increased risk of diabetes (382). Similar trends in the prevalence of cardiovascular diseases linked to urbanisation were reported in Ethiopia (370, 383-390). Furthermore, a previous selection scan in indigenous Ethiopian population of Wolaita has also reported a recent positive selection on genes involved in immunity and energy metabolism during prolonged food shortage that were linked to diabetes and obesity susceptibility (370). Apart from diet, high-altitude hypoxia that promotes lipid storage and carbohydrate oxidation might have contributed to metabolic adaptation (370, 391, 392). However, we did not replicate the high-altitude adaptation signals reported previously for Ethiopian highlanders and Andean Quechua (370, 393), although there is no information whether the populations analysed in this study were residing at high-altitude.