

# 1 Sparse haplotype-based fine-scale local 2 ancestry inference at scale reveals recent 3 selection on immune responses

4 Yaoling Yang<sup>1,2,✉</sup>, Richard Durbin<sup>3</sup>, Astrid K. N. Iversen<sup>4</sup>, and Daniel J. Lawson<sup>1,2,✉</sup>

5 <sup>1</sup>Department of Statistical Sciences, School of Mathematics, University of Bristol, Bristol, UK

6 <sup>2</sup>MRC Integrative Epidemiology Unit, Population Health Sciences, University of Bristol, Bristol, UK

7 <sup>3</sup>Department of Genetics, University of Cambridge, Cambridge, UK

8 <sup>4</sup>Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, University of Oxford, Oxford, UK

9 ✉Corresponding authors: Yaoling Yang ([yaoling.yang@bristol.ac.uk](mailto:yaoling.yang@bristol.ac.uk)) and Daniel J. Lawson

10 ([dan.lawson@bristol.ac.uk](mailto:dan.lawson@bristol.ac.uk))

## 11 Abstract

12 Increasingly efficient methods for inferring the ancestral origin of genome regions are needed to gain  
13 new insights into genetic function and history as biobanks grow in scale. Here we describe two near-  
14 linear time algorithms to learn ancestry harnessing the strengths of a Positional Burrows-Wheeler  
15 Transform (PBWT). SparsePainter is a faster, sparse replacement of previous model-based ‘chromo-  
16 some painting’ algorithms to identify recently shared haplotypes, whilst PBWTPaint uses further ap-  
17 proximations to obtain lightning-fast estimation optimized for genome-wide relatedness estimation.  
18 The computational efficiency gains of these tools for fine-scale local ancestry inference offer the pos-  
19 sibility to analyse large-scale genomic datasets in completely novel ways. Application to the UK  
20 Biobank shows that haplotypes better represent ancestries than principal components, whilst linkage-  
21 disequilibrium of ancestry identifies signals of recent changes to population-specific selection for many  
22 genomic regions associated with immune responses, suggesting new avenues for understanding the  
23 pathogen-immune system interplay on a historical timescale.

## 24 Introduction

25 Modern human populations are complex mixtures between ancient contributing source groups<sup>1</sup>. Ge-  
26 netic admixture is the process of mixing groups that were genetically distinct due to genetic drift,  
27 which can create new distinct populations<sup>2,3</sup>. The process is ubiquitous and spans scale in space and  
28 time, from the admixture with Neanderthals around 50,000 years ago when modern humans migrated  
29 out of Africa<sup>4</sup>, to native Americans mixing with primarily European and African immigrants over the  
30 last 500 years to form the majority of United States ancestry<sup>5</sup>, and the fine-scale geographical regional-  
31 isation within a single country such as the UK<sup>6</sup>. The identification of chromosomal regions originating  
32 from a specific population is known as local ancestry inference (LAI)<sup>7</sup>, which can be used to map  
33 disease loci<sup>8</sup>, investigate the relationships between modern populations, improve association studies<sup>9</sup>,  
34 and study demographic histories<sup>10</sup>.

35 Genome-wide association studies (GWAS) have identified single nucleotide polymorphisms (SNPs)  
36 associated with human complex traits and diseases<sup>11</sup>, but the SNP frequencies are likely to be associ-  
37 ated with particular ancestries. Local ancestry may then either be viewed as a confounder of the SNP  
38 effect<sup>9</sup>, or treated as a predictor as in ‘Ancestral GWAS’<sup>12</sup>. In this framing, local ancestry inference  
39 examines the ancestral origin of risk loci in terms of a population and a time — for instance, risk alleles  
40 associated with multiple sclerosis originated from pastoralists dwelling on the Pontic Steppe, which  
41 were brought into Europe by the Yamnaya-related migration around 5,000 years ago<sup>12</sup>. Other examples  
42 include the relationship between platelet count in Hispanics and an Amerindian-origin variant of the  
43 ACTN1 gene<sup>13</sup>, a link between quantitative red blood cell traits and African- and Amerindian-origin  
44 loci in the HBA1/2 gene<sup>14</sup>, and kidney disease in African-origin variants of the APOL1 gene<sup>15</sup>.

45 It is hard to perform LAI accurately and efficiently. Various LAI software have been developed  
46 since the 21st century, and the majority<sup>16</sup> are based on the Li and Stephens hidden Markov model  
47 (HMM)<sup>17</sup>, including HAPMIX<sup>7</sup>, ChromoPainter<sup>18</sup>, LAMP-LD<sup>19</sup>, MOSAIC<sup>3</sup> and FLARE<sup>20</sup>. HAP-  
48 MIX pioneered this application but is limited to modelling two ancestries. In comparison, Chro-  
49 moPainter enables the accurate analysis of admixtures from multiple groups but is slow. LAMP-LD is  
50 faster but can be unstable<sup>16</sup>. The distinctive feature of MOSAIC is that the knowledge of the intricate  
51 connections between reference haplotypes and ancestral mixing groups is not required<sup>3</sup>. Recently,  
52 through the on-the-fly compression of reference panels, saved checkpoints and composite reference  
53 haplotypes, FLARE greatly improves the computational performance compared with the previous LAI  
54 software<sup>20</sup>. Other approaches for local ancestry inference are also possible, among which PCAdmix,  
55 a Principal Components-based algorithm<sup>21</sup>, and RFMix<sup>22</sup>, which employs a discriminative modelling  
56 strategy, are popularly used.

57 Our technical contribution is providing two algorithms that fulfil different use cases. Both are  
58 significantly faster than anything previously reported, especially for identifying fine-scale population  
59 structure. The most rapid is orders of magnitude faster, opening the application to hundreds of thou-  
60 sands or even millions of samples as presented by the most challenging modern biobanks and associa-  
61 tion studies. These approaches avoid storing the entire genotype information in memory, instead using  
62 the Positional Burrows-Wheeler Transform (PBWT)<sup>23,24</sup> to extract only a sparse set of the longest hap-  
63 lotype matches to the reference panel at each position. In PBWTpaint, only the longest *set-maximal*  
64 matches are retained, which we will show is sufficient for genome-wide ancestry. In SparsePainter, we  
65 extract a richer set of haplotypes on which we show that a sparse implementation of the Li and Stephens  
66 HMM model<sup>17</sup> can be run with a negligible accuracy cost by using a Hash Map data structure<sup>25</sup>.

67 Identifying genomic features that are of biological significance from fine-scale local ancestry infor-  
68 mation is an under-explored topic and the core of our scientific contribution. Within SparsePainter we  
69 are able to efficiently compute Linkage Disequilibrium of Ancestry (LDA), LDA score (LDAS) and  
70 Ancestry Anomaly Score (AAS)<sup>12</sup> at scale. These recently proposed summary statistics of local an-  
71 cestry are predicted under recent population-specific selection, but previous implementations based on  
72 post-processing local ancestry data are only suitable for examining small sections of genome or small

73 reference datasets. LDA is the correlation of ancestries between SNP pairs, which measures whether  
74 recombination events between ancestries are more frequent than those within ancestries. LDAS cal-  
75 culates the total LDA of each SNP on the chromosome weighted by genetic distance. A lower LDAS  
76 indicates the haplotype inherited from the reference population is shorter than expected. We identify  
77 two mechanisms that generate low LDAS and both involve a change in selection between the pre-  
78 existing and admixed population. The first involves selection on a nearby locus, leading to balancing  
79 selection at the level of haplotypes. The second is against a locus that was high frequency in at least one  
80 contributing population. AAS is the degree of difference between the estimated average ancestry prob-  
81 abilities and the genome-wide average, which detects signals of recent selection for loci experiencing  
82 changes in ancestry frequencies.

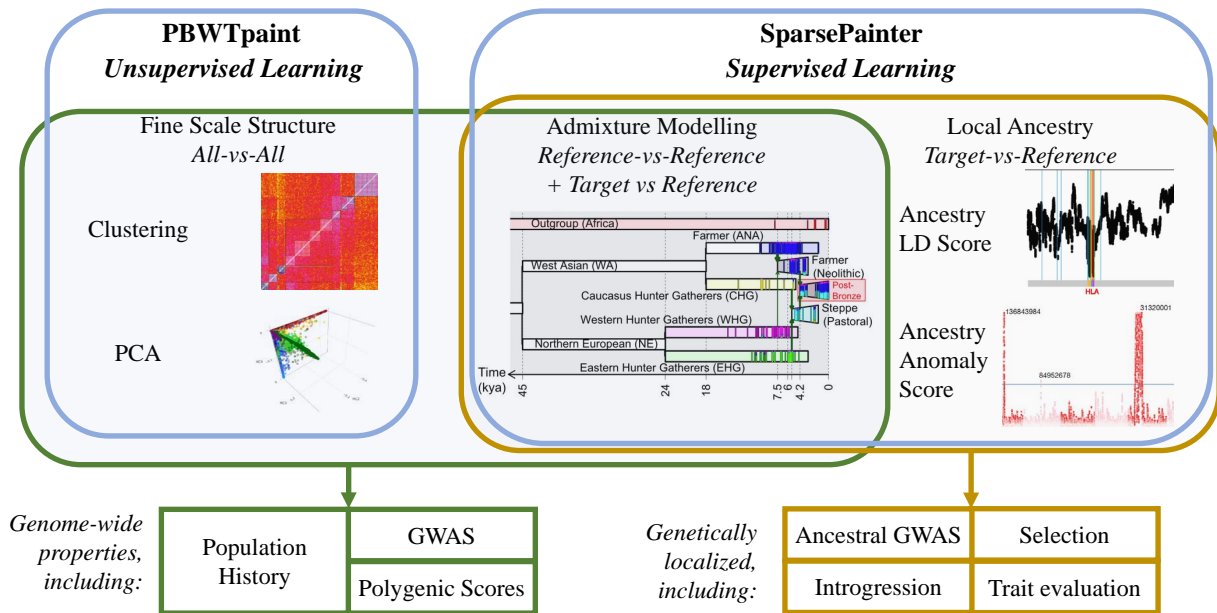
83 We benchmarked SparsePainter against ChromoPainter and FLARE, which demonstrates that Sparse-  
84 Painter is faster both empirically and in scaling at fine scale, i.e. as the number of reference populations  
85 grows. PBWTPaint is faster than all methods by orders of magnitude in identifying genome-wide hap-  
86 lotype structure within a single dataset, which is its specific capability.

87 In exploring population structure within the UK Biobank (UKB) with PBWTPaint, we construct  
88 haplotype principal components (HCs) which we compare to the widely-used SNP-based principal  
89 components (PCs). HCs are better associated with birthplace and seem to capture more nuanced ge-  
90 netic variation than PCs, revealing distinct ancestral patterns among ethnic backgrounds and significant  
91 regional distinctions within the UK and Ireland, suggesting potential for more refined population strat-  
92 ification in genetic studies. Using 1000 Genomes Project (1000GP) Data<sup>26</sup> as reference, we can apply  
93 the LDAS and AAS statistics to identify genes that show signals of recent changes to population-  
94 specific selection. This approach, applied genome-wide, identifies a number of genes that are almost  
95 entirely immune-related, pointing to population-specific immune response as a central driver of selec-  
96 tion acting on historical timescales.

## 97 Results

### 98 Method Overview

99 There are two main approaches to ancestry inference. The first is *unsupervised learning*, which ad-  
100 dresses the goal of learning fine-scale population structure. Examples include clustering<sup>18</sup>, unsuper-  
101 vised admixture models<sup>27,1</sup>, or dimensionality reduction such as Principal Component Analysis (PCA)  
102 based on either genotype<sup>28,29</sup> or haplotype data<sup>18</sup>. Here, the data are not typically curated and we  
103 aim to form the largest dataset possible for the analysis. The second approach is *supervised learning*,  
104 in which target individuals are compared to carefully curated reference populations, and recently ad-  
105 mixed individuals (which are the majority of individuals) are not directly used. The goal of supervised  
106 learning divides into *ancestry estimation* which can be used analogously to unsupervised genome-wide  
107 ancestry profiles<sup>30</sup>, or *local ancestry estimation* in which the ancestry of particular sections of DNA is  
108 inferred.



**Fig. 1: An overview of the scientific use-cases of SparsePainter and PBWTPaint.** PBWTPaint performs *all-vs-all* painting, for use in fine-scale structure estimation via unsupervised learning approaches, such as clustering (plot from Lawson et al. (2012)<sup>18</sup>) and PCA. SparsePainter performs supervised learning which can be separated into *reference-vs-reference* painting for admixture estimation and population history modelling, and *target-vs-reference* painting for local ancestry inference, such as LDAS and AAS (plots from Barrie et al. (2024)<sup>12</sup>).

109 These goals are met by two tools that facilitate a completely new scale of haplotype-based ancestry  
 110 analysis, as described in Fig. 1. The first of these is PBWTPaint, a direct extension of the PBWT<sup>23</sup>  
 111 which rapidly identifies long matches. This uses two innovations to achieve extreme computational  
 112 performance for unsupervised learning of a single dataset, comparing each individual to every other in  
 113 *all-vs-all* painting. First, PBWTPaint only considers a very limited subset of possible matches repre-  
 114 senting the maximally shared haplotypes at any locus (called *set-maximal*). Further, the Li & Stephens  
 115 model is replaced with an approximation that only considers overlapping set-maximal matches, run-  
 116 ning in linear-time so that mega-scale analyses are straightforward. Larger datasets uncover longer,  
 117 more recent matches, and any inaccuracies due to modelling approximations average out over the  
 118 whole genome for genome-wide analyses.

119 The second tool is SparsePainter, which is designed to perform accurate local ancestry inference  
 120 efficiently. Whilst SparsePainter can perform *all-vs-all* painting, it is optimised for either painting a  
 121 reference panel against itself (*reference-vs-reference* painting), or painting target individuals using a  
 122 reference panel (*target-vs-reference* painting). There are two primary outputs of SparsePainter. The  
 123 first is local ancestry estimates, which are the probabilities that a haplotype at a particular chromosomal

124 location is inherited from each ancestral individual or population. By efficient representation of this  
125 we can efficiently compute the selection statistics LDA, LDAS and AAS.

126 Chromosome Painting is different to identifying haplotypes identical-by-descent; it assigns every  
127 position of the genome to the most-recent common ancestor in the reference, without allowing overlaps  
128 or conditioning on length and hence expected age of a sharing event. This facilitates fine-scale ancestry  
129 estimation or ‘Admixture Modelling’<sup>18,1</sup> using the expected fraction of the total genome shared most  
130 recently between a target and each reference ancestral individual or population. As we use a leave-  
131 one-out scheme to make individuals from the reference and target datasets *exchangeable*, i.e. receive  
132 the same ancestry inference if they share the same ancestry (see Methods for a formal definition), this  
133 allows population history reconstruction without assuming perfect references<sup>1</sup>.

134

### 135 **PBWTpaint**

136 Storing the genotype information of all the samples in memory is a problem for large datasets. The  
137 Positional Burrows-Wheeler Transform (PBWT)<sup>23</sup> is a data structure to transform a binary matrix  $X_{ik}$   
138 (with  $2N$  haplotypes and  $K$  SNPs) into a sequence of run-length compressed arrays per SNP, in each  
139 of which the haplotype values at the SNP are sorted according to the reversed haplotype prefixes pre-  
140 ceding the SNP. From a PBWT, long matches can be efficiently extracted using the *ReportMatches*  
141 algorithm, and set-maximal matches with the *ReportSetMaximalMatches* algorithm, in  $O(NK)$  oper-  
142 ations for all haplotypes at the same time. Our models are built on these matches.

143 For each target individual  $i$ , PBWTpaint iterates through the  $M(k)$  matches at a locus  $k$  (which  
144 are typically very sparse, and sparse by construction for set-maximal matches). For each matched  
145 reference haplotype  $j$  we extract the start  $s_{jk}$  and end  $e_{jk}$  positions of the maximal exact match to  $j$   
146 covering  $k$ , i.e.  $s_{jk}$  is the location just after the first upstream mismatch, and  $e_{jk}$  is the location just  
147 before the first downstream mismatch. From these, we compute a weight  $w_{jk} = (k - s_{jk})(e_{jk} - k)$ , i.e.  
148 the weight increases linearly with distance from each end of the match, and quadratically with the total  
149 length of the match for positions at the midpoint of the match. This is normalised over matches  $j$  to  
150 give a local ancestry score  $p_{jk} = w_{jk} / \sum_{l=1}^{2N} w_{lk}$ , which we sum over loci  $k$  to produce a genome-wide  
151 ancestry estimate  $p_j$ . We also provide estimates of the total number of recombination events, as well  
152 as regional bootstraps, to enable clustering with FineSTRUCTURE<sup>18</sup>.

153

### 154 **From PBWT to an accurate Sparse Data Matrix**

155 For local ancestry inference, the longest haplotype matches at the target locus are the most important,  
156 since short matches appear within any ancestry due to statistical noise and incomplete lineage sorting,  
157 i.e. ancient structure shared across ancestries rather than recent genealogical relationships. As such,  
158 short matches provide little useful information for tracing local ancestry.

159 Whilst the original PBWT algorithm finds long matches only within the same database, it has been  
160 extended to report long matches between different haplotype sets<sup>24</sup>. For accurate and efficient local an-  
161 cestry inference we detect all matches longer than some threshold  $L$ , but there may be no genome-wide

162 ‘correct’  $L$ . Some target haplotypes will only have short matches if they diverged a long time ago, and  
163 few or even no matches are longer than  $L$ . Other target haplotypes will share extremely long segments  
164 of DNA with many reference haplotypes leading to many matches being retained, the shorter of which  
165 (also longer than  $L$ ) are not helpful for inferring ancestry. To address this barrier, we revisited the ‘long  
166 match query’ algorithm of PBWT and proposed the Algorithm ‘ReportLongestMatches’ which aims  
167 to find at least  $Q$  longest matches at each position for a target sample  $i$  (Methods). With this algorithm,  
168 we maintain a particular sparsity level at each location while also preserving the longest matches to  
169 guarantee accuracy.

170

### 171 **Using Hash Map to perform HMM Forward-Backward Algorithm in sparse form**

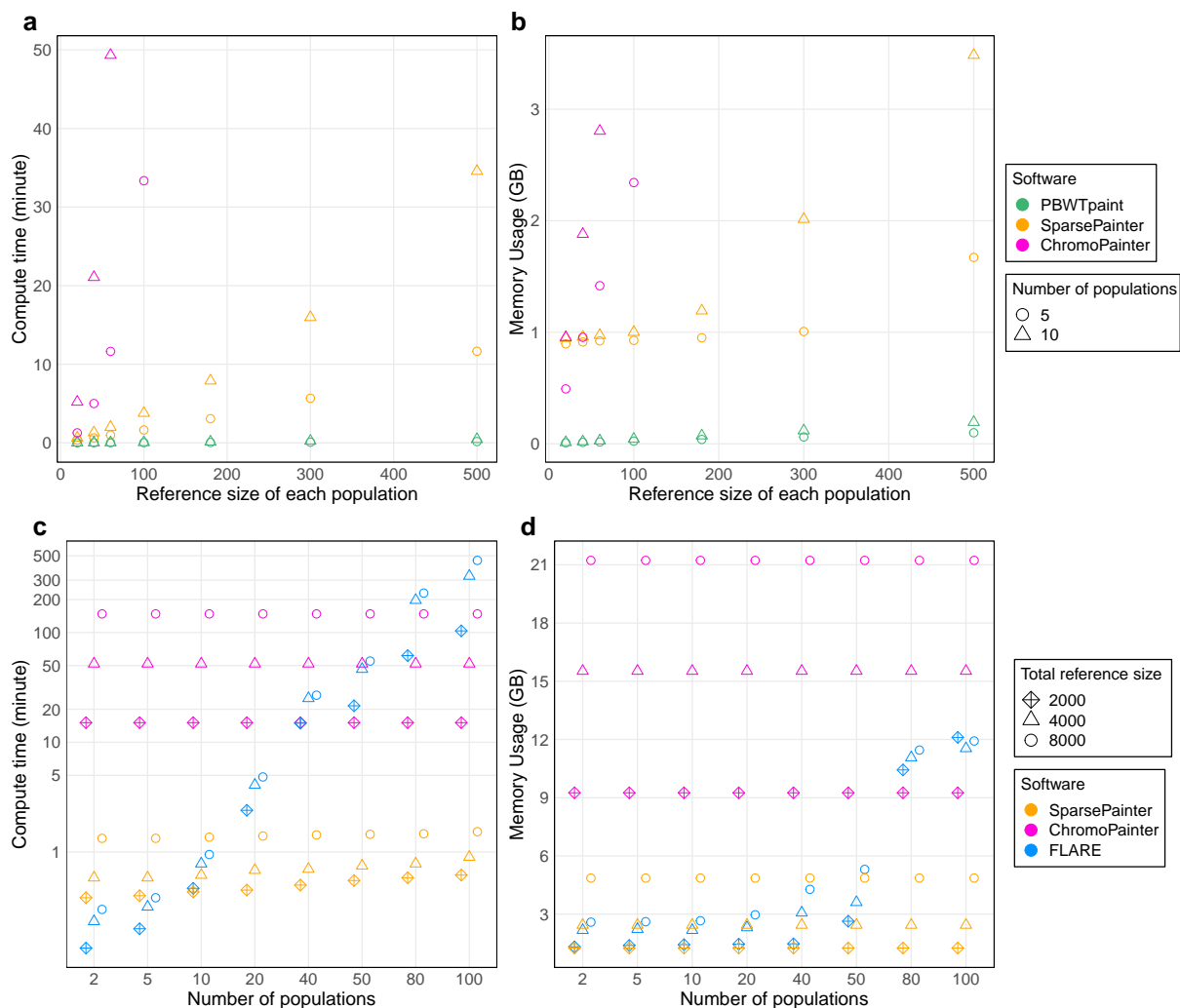
172 SparsePainter stores haplotype matches in a Hash Map data structure that implements an associative  
173 array abstract data type for efficient key-value storage and retrieval<sup>25</sup>, facilitating  $O(1)$  storage and  
174 lookup of values (here painting probabilities) based on unique identifiers or keys (here haplotype in-  
175 dices). We then employ a sparse approximation to Li and Stephen’s<sup>17</sup> model by vectorising the forward  
176 and backward probabilities and assuming a vanishing mutation rate (Methods). The forward and back-  
177 ward computation is only required within the  $Q$  longest matches to the target haplotype at each locus,  
178 allowing efficient computation of the local ancestry probabilities and the expected genome shared.  
179 Compared with computing and storing the probabilities at all  $N$  haplotypes, our approach reduces  
180 both memory usage and compute time from  $O(N)$  to  $O(Q)$ .

181

### 182 **Simulation overview**

183 We used SLiM 3.7.1<sup>31</sup> to simulate genetic data on 20 megabases throughout 3000 generations, aiming  
184 to compare the accuracy, speed and memory utilization of SparsePainter, ChromoPainter, FLARE and  
185 PBWTpaint in terms of local ancestry and genome-wide estimates. Here we focused on comparing  
186 tools which are useful for large reference panels. Therefore, we excluded MOSAIC and RFMix which  
187 are not sufficiently scalable<sup>20</sup>. For this comparison, we used four distinct simulation models with  
188 20k SNPs (noting that all methods have linear compute and memory requirements in the genome size  
189 analysed; see Methods for details):

- 190 • **Simulation 1:** A hierarchical model designed to assess the speed, memory usage, and accuracy of  
191 PBWTpaint, SparsePainter, and ChromoPainter for within reference (supervised or unsupervised)  
192 painting;
- 193 • **Simulation 2a:** An evolutionary process that generates from 2 to 100 different populations, to inves-  
194 tigate the scaling of SparsePainter and ChromoPainter in *target-vs-reference* painting;
- 195 • **Simulation 2b:** A less-separated version of Simulation 2a with limited populations, to assess the  
196 accuracy of *target-vs-reference* painting for SparsePainter, ChromoPainter, and FLARE;
- 197 • **Simulation 2c:** A larger-scale version of Simulation 2b to investigate how SparsePainter balances  
198 accuracy against speed and memory utilization in *target-vs-reference* painting.

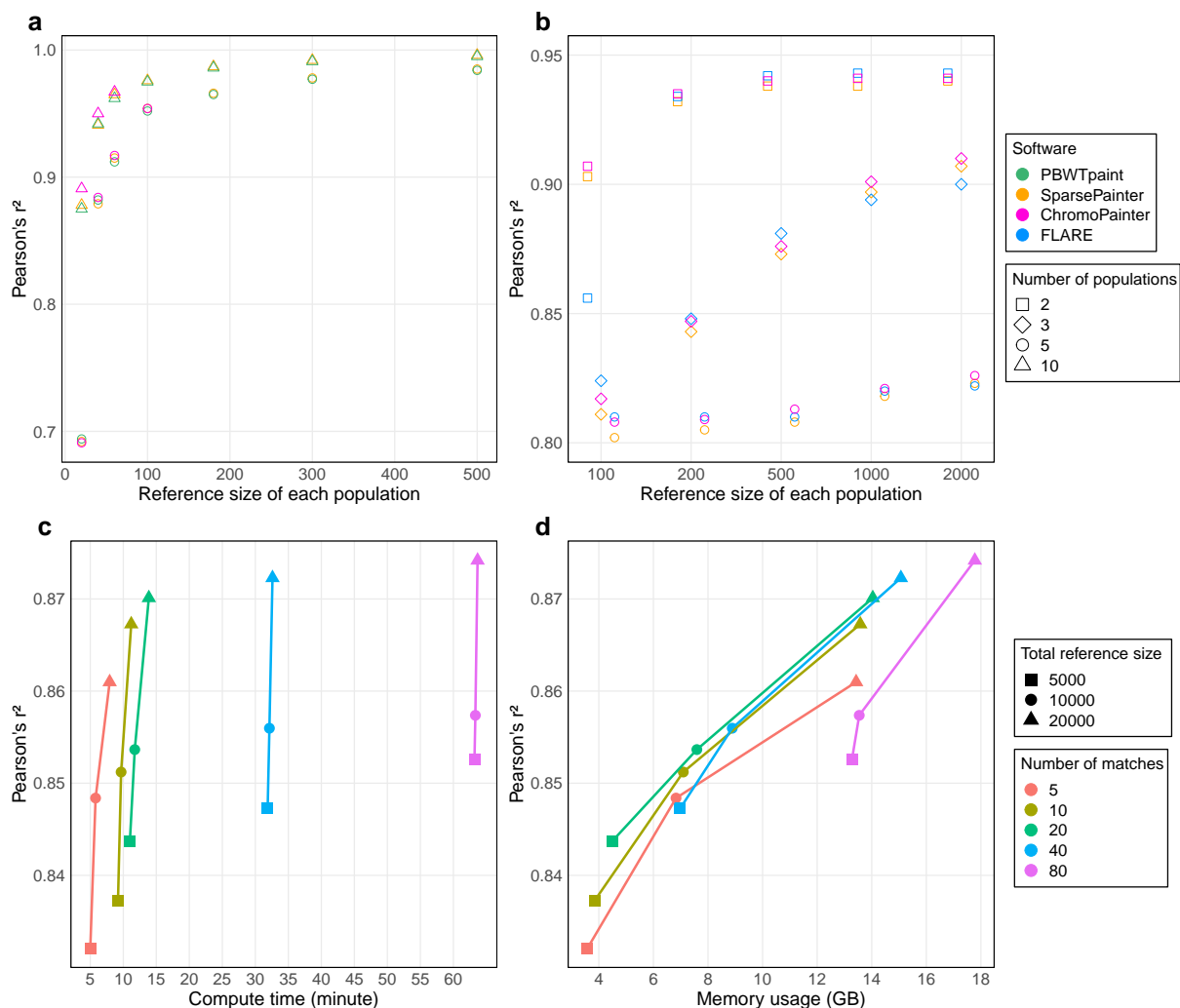


**Fig. 2: Speed and memory comparison between software.** a-b, Speed and memory of admixture estimates for *reference-vs-reference* painting between software with 5 or 10 populations and different reference sizes with 20k SNPs (Simulation 1). c-d, Speed and memory of painting 50 target individuals between software with different numbers of populations and reference sizes with 20k SNPs (Simulation 2a).

### 199 Within-sample performance comparison

200 We first compared the efficiency of PBWTpaint (using *all-vs-all* painting) and SparsePainter and ChromoPainter (using *reference-vs-reference* painting) under Simulation 1. FLARE is excluded as it can  
 201 neither perform within-sample (i.e. reference-vs-reference), nor genome-wide, comparisons. Performance is measured using the *recovery rate* of an individual's own population ancestry fraction using  
 202 squared Pearson's correlation coefficient (denoted as  $r^2$ ) with the truth (Methods).  
 203  
 204

205 Fig. 2a-b illustrates that both in theory and practice, ChromoPainter has a quadratic cost as a function of panel size, so scales poorly to larger reference sizes. SparsePainter is close to linear in both  
 206 speed and memory efficiency regardless of reference sizes. Whilst PBWTpaint also scales linearly, it  
 207



**Fig. 3: Accuracy of software and the trade-off between accuracy and computational cost in SparsePainter.** a, Self-recovery rate for *reference-vs-reference* painting with 20k SNPs (Simulation 1). b, Accuracy of local ancestry estimates with 20k SNPs and 50 target individuals sampled 13 generations after admixture (Simulation 2b). c-d: compute time and memory usage of SparsePainter for painting with different sparsity and reference sizes under a 5-way admixture model (Simulation 2c) with 20k SNPs.

208 is orders of magnitude faster, and only introduces a minor trade-off in terms of accuracy (Fig. 3a).  
 209 Notably, PBWTpaint only retains accuracy for genome-wide estimation, as its simple model with set-  
 210 maximal matches isn't suitable for estimating local ancestries (Methods).

211

### 212 **Target-vs-reference speed and memory comparison for LAI**

213 As PBWTpaint neither can paint target samples against different reference panels, nor perform local  
 214 ancestry estimates, we restricted our speed and memory comparison to SparsePainter, ChromoPainter  
 215 and FLARE with Simulation 2a. As all those software are based on the Li and Stephen's hidden



216 Markov model, their computational costs for genome-wide and local ancestry estimates are expected  
217 to be similar.

218 The speed and memory of SparsePainter and ChromoPainter remain largely unaffected by the num-  
219 ber of true populations. Conversely, whilst FLARE demonstrates impressive speed and efficient mem-  
220 ory usage with few populations ( $n_{pop} \leq 5$ ), its efficiency dramatically diminishes compared to Sparse-  
221 Painter when handling 20 or more populations (Fig. 2c-d). When painting with 100 populations,  
222 SparsePainter is over 100 times faster and 10 times more memory-efficient than FLARE.

223 A recent study<sup>30</sup> decomposed the UK Biobank into 129 distinct fine-scale reference ancestries. We  
224 replicated their analysis with the 4334 reference individuals from non-restricted data sources (i.e. all  
225 except POPRES) spanning 129 populations. For 1000 target individuals on chromosome 21 which  
226 comprises 9522 SNPs, SparsePainter is dramatically faster and requires minimal memory (6 minutes  
227 and 1.5GB) compared with ChromoPainter (272 minutes and 10.2GB) and FLARE (338 minutes and  
228 14.2GB).

229

### 230 ***Target-vs-reference accuracy comparison for LAI and admixture estimation***

231 We have illustrated the circumstances when SparsePainter has superior speed and memory use than  
232 FLARE and ChromoPainter, but it is crucial to maintain accuracy. In Simulation 2b we examined  
233 the accuracy of local ancestry estimated by both the squared Pearson's correlation coefficient and  
234 the proportion of accurate local ancestry predictions (Methods). Across all software, the accuracy of  
235 local ancestry estimation consistently improves with increased reference sizes and reduced number of  
236 populations (Fig. 3b and Supplementary Fig. 1).

237 The accuracy of SparsePainter and Flare is always comparable. Also as anticipated, SparsePainter  
238 displays a negligible accuracy drop compared to ChromoPainter, given that SparsePainter is essentially  
239 a sparse implementation of ChromoPainter.

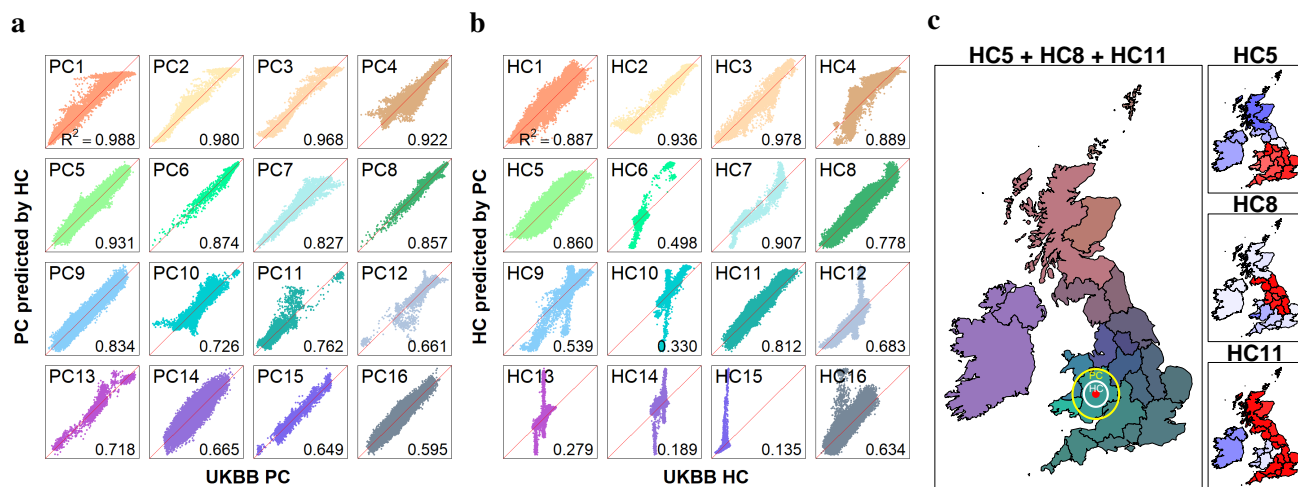
### 240 **Sparsity in SparsePainter**

241 To investigate SparsePainter's tradeoff between sparsity and accuracy, we varied the reference size of  
242 a 5-way admixture model (Simulation 2c). Fig. 3c-d shows that a larger reference size substantially  
243 boosts accuracy, whilst increments in the number of matches only marginally elevate it, and larger  
244 reference samples dilute the accuracy's sensitivity to sparsity. Conversely, computational time and  
245 memory demands surge considerably as match density escalates. This indicates that if large reference  
246 datasets are available, opting for a constant number of matches (so diminished match proportion) yields  
247 significant computational savings, at a negligible compromise in accuracy.

248

### 249 **Haplotype Principal Components Analysis for the UK Biobank**

250 The UK Biobank (UKB)'s principal components (PCs) are widely used for correctly inferring the pop-  
251 ulation structure. We inferred the (sparse) genome-wide pairwise coancestry of  $N = 406,733$  UK  
252 Biobank individuals via PBWTPaint from  $L = 569,200$  SNPs, taking 41 CPU hours (which is paral-  
253 lizable and scales as  $O(NL)$ ). We summarised these ancestries into the top 150 haplotype components



**Fig. 4: Comparison between UK Biobank PCs and HCs and the decomposition of HCs.** a, the coefficient of determination for predicting the first 16 UKB PCs (y-axis) using the first 150 HCs (x-axis) with linear regression models ( $n=406,733$  individuals), which shows strong correlations. b, the coefficient of determination for predicting the first 16 UKB HCs (y-axis) using the first 150 PCs (x-axis) with linear regression models ( $n=406,733$  individuals), which shows strong correlations on only 12 of the first 16 HCs. c, Visualisation of the average of the 5th, 8th and 11th HC stratified by birthplaces within the UK and Ireland ( $n=347,532$  individuals), corresponding to the red, green, and blue channels, respectively, in the composite plot (left), and the right plot shows the decomposition of each HC. We have also shown the median prediction error range of the birthplace of HCs (white circle, radius 39.7km) and PCs (yellow circle, radius 77.5km) of an east Wales location (red point).

254 (HCs) (Methods), and compared their informativeness with PCs in several ways. First, we can accu-  
 255 rately predict the first 16 PCs with the first 150 HCs using linear regression models (Fig. 4a), especially  
 256 for the first 9 PCs which have a coefficient of determination ( $R^2$ ) exceeding 80%. Conversely, when  
 257 using the first 150 PCs to predict the first 16 HCs, some of the HCs are poorly explained (Fig. 4b).  
 258 This observation indicates that HCs might encapsulate additional ancestry information beyond that  
 259 conveyed by PCs.

260 To investigate consistency across chromosomes, we split the SNPs from the odd and even chromo-  
 261 somes and then computed the top 150 PCs and HCs from the even chromosome set. Subsequently, we  
 262 used 150 HCs/PCs from one set to predict each of the top 50 HCs/PCs from the other set. HCs are well  
 263 explained with  $R^2 > 90\%$  for the majority of them (Extended Data Fig. 3), indicating HCs capture  
 264 ancestry information that is shared in all the chromosomes. By contrast, few PCs can be predicted  
 265 from different chromosome sets, which corresponds to the previous finding that all PCs except the top  
 266 few of them are related to specific genetic regions<sup>32</sup>.

267 HCs are associated with self-reported ethnicity (Extended Data Fig. 1): the 2nd and 3rd HCs ef-  
 268 fectively differentiate within white and black backgrounds, respectively, whereas the 4th and 6th HCs  
 269 reflect variations associated with South Asian ancestry. HCs are also associated with geography: fil-

270 tering to the 347,532 individuals with white, British or Irish ethnicity born in the UK or Ireland, we  
271 plotted the average HC for 23 UK regions (Extended Data Fig. 2). The 5th HC represents the variation  
272 between Scotland, Irish, and the rest of the UK, while the 11th HC differentiates Ireland and Wales  
273 from the other regions. By mapping the 5th, 8th, and 11th HCs onto the geography of the UK and  
274 Ireland, we created a colour-coded depiction (Fig. 4c) which uniquely identifies each county. Further,  
275 predicting birth location using HCs has a median error of 39.7km, whilst PCs give a nearly double  
276 error of 77.5km in out-of-sample individuals (see Methods). This is a surprisingly high accuracy as  
277 these individuals were not filtered for having ancestry from a single location, so prediction accuracy is  
278 bounded by migration since people need not be born where their ancestors came from.

279

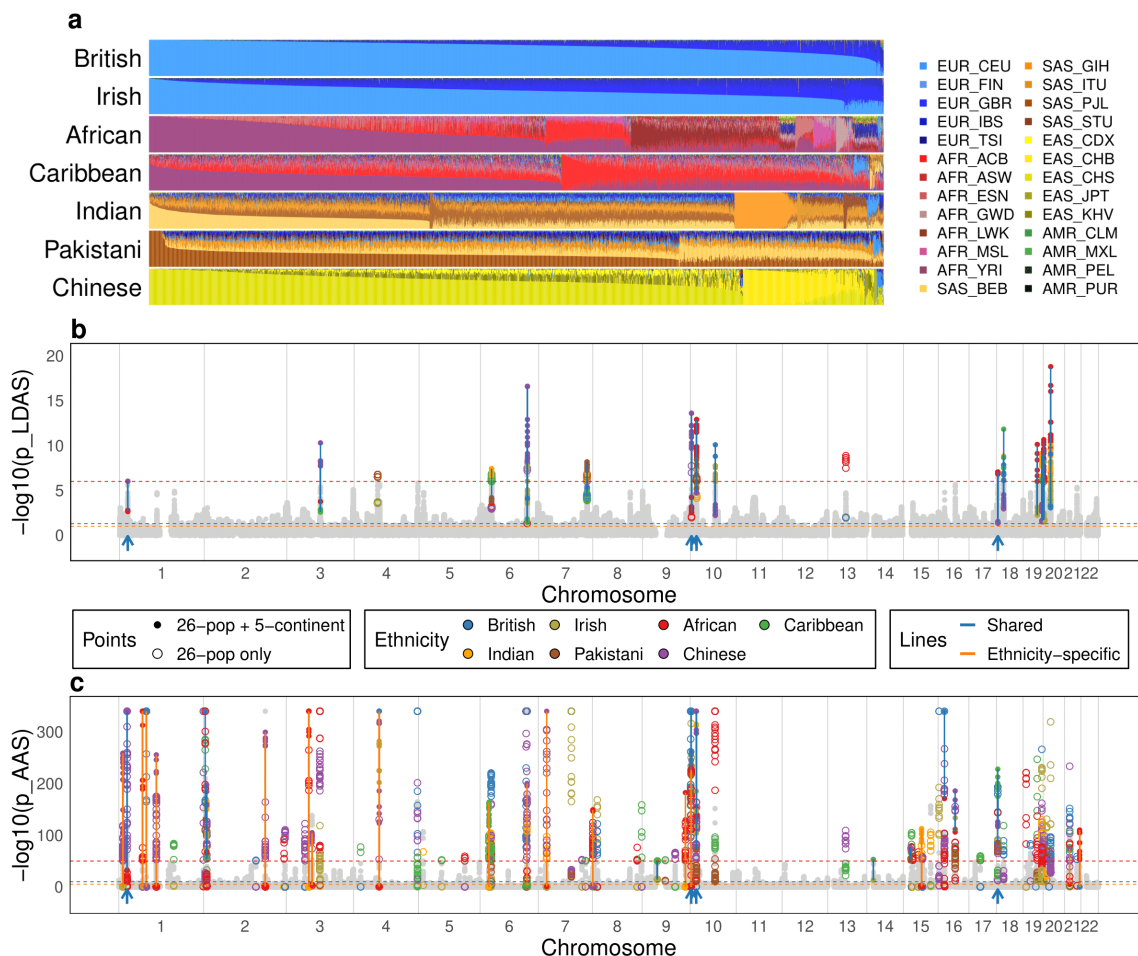
### 280 **Ethnicity-specific selection in the UK Biobank compared to the 1000 Genomes populations**

281 To demonstrate the scientific value of SparsePainter, we inferred the local ancestry of 487,409 UK  
282 Biobank<sup>33</sup> individuals using the 2504 individuals spanning 26 populations from the 1000 Genomes  
283 Project (1000GP)<sup>26</sup> as reference. From this, we evaluated selection using LDA score, which quantifies  
284 genomic regions with particularly short ancestry segments, compared to the base recombination rate,  
285 as well as an Ancestry Anomaly Score (AAS), which quantifies regions of unusual ancestry, compared  
286 to genome-wide (see Methods). We report results that replicate over 7 primary self-reported ethnic  
287 backgrounds (hereafter ethnicities) within the UK Biobank: British, Irish, Indian, Caribbean, African,  
288 Pakistani, and Chinese. The LDAS, AAS and average probabilities of 26 1000GP populations for each  
289 SNP analysed within each ethnicity are available in Supplementary Tab. 2-8.

290 Our goal is to demonstrate applications of local ancestry at scale outside of population history and  
291 admixture estimation (Fig. 5a, Methods). We look for signals of ‘putative selection’ in the form of low  
292 LDAS and unusual AAS that are shared, i.e. identified in every UKB primary ethnicity, after extensive  
293 quality control (Methods). As a sensitivity analysis, we further painted the UK Biobank with 1000GP  
294 data using 5 continental ancestries (EUR, AFR, SAS, EAS and AMR). The LDAS and AAS results  
295 of different UKB ethnicities are illustrated in Fig. 5b-c. These are mapped to genes, with shared  
296 significant low LDAS and AAS signals visualised in Fig. 6 and investigated in detail in Supplementary  
297 Note 1. Genes with ethnicity-specific AAS signals are reported in Supplementary Tab. 1.

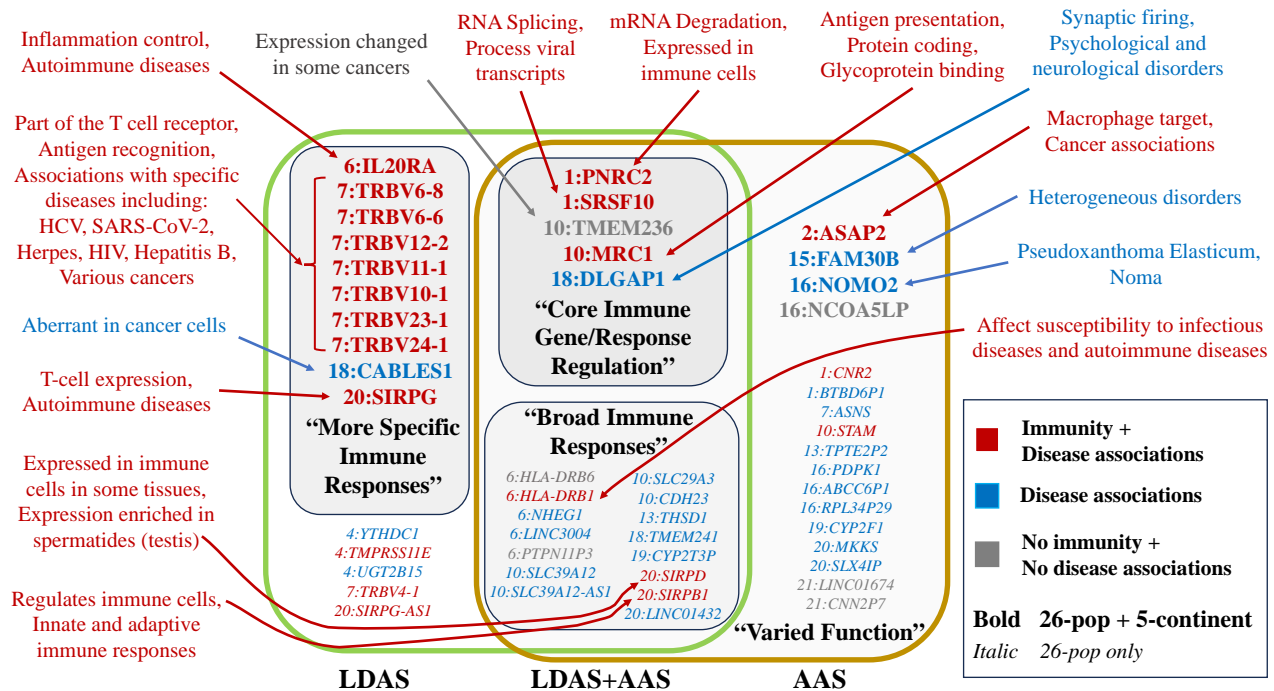
298 To aid in interpreting these signals, we extended simulations for low LDAS from Barrie et al.  
299 (2024)<sup>12</sup> (Methods, Extended Data Fig. 5-6). Two scenarios produce significantly low LDAS and  
300 extreme AAS, and both imply a change in selection following admixture. One scenario is single-locus  
301 negative selection in the admixed population, following non-negative selection in the pre-existing pop-  
302 ulations. The second scenario is multi-locus positive selection in the admixed population, while those  
303 loci are either absent or present at low frequency in some of the pre-existing populations. Selection  
304 under these scenarios is not detected by iHS, iHH12 or nSL as calculated using selscan<sup>34</sup>, showing that  
305 extreme LDAS SNPs are not expected to be previously reported as under selection.

306 Extreme AAS signals in all 7 UKB ethnicities (Extended Fig. 7) include LINC01432 from chro-  
307 mosome 20 (linked to retroperitoneum carcinoma and early-onset androgenetic alopecia) which has an



**Fig. 5: Modelling of 7 UK Biobank self-reported ethnicities using 26 1000GP populations.** a, Overall ancestry inference stratified by UKB ethnicities. For each ethnicity, the column shows ancestry decomposition for a single individual, with colours representing different 1000GP reference populations, named as regions followed by local population in standard abbreviation<sup>26</sup>. b, Linkage-Disequilibrium of Ancestry Score (LDAS), reporting  $-\log_{10}$  of the p-value of low LDAS (normality test). c, Ancestry Anomaly Score (AAS) as a function of genome position, reporting  $-\log_{10}$  of the p-value of AAS (chi-squared test,  $-\log_{10}(p)$  is capped at 340 in the plot). All plots describe the analysis of  $n=487,409$  individuals on 569,200 SNPs. In b-c, the non-light-grey points (light grey points) represent the SNPs' maximum and minimum values that exhibit significant (insignificant) scores in both (either) paintings with 1000GP 26 populations and 5 continents, respectively (Methods), and blue (orange) lines connect the maximum and minimum values at each SNP that are shared (ethnicity-specific) across the 7 ethnicities in both paintings. The thresholds used to determine significance are depicted as horizontal lines in dashed red, blue and orange, respectively.

308 exceptionally high Japanese ancestry (EAS\_JPT) across all UKB ethnicities. Similarly, in the genes  
 309 PNC2 and SRSF10 on chromosome 1, the Puerto Rican ancestry (AMR\_PUR) is over-represented,  
 310 particularly within European and Asian ethnicities. Notably, LINC03004 (highly expressed in testis



**Fig. 6: Summary of previous findings for genes with low LDAS and AAS signals shared between 7 UK Biobank self-reported ethnic backgrounds.** Genes with low LDAS and AAS signals in both 26-pop and 5-continent paintings include those with core immune gene or response regulation, while those in 26-pop painting only include many broad-impact immune genes. Genes with LDAS-only signals in both 26-pop and 5-continent paintings more typically affect responses to specific infections, and genes with AAS-only signals have varied functions and disease associations. Classification (colour) and category summaries (bold quoted text) are based on heuristic features of previous work; see Supplementary Note 1 for details.

311 and the gall bladder) and its nearby gene PTPN11P3 on chromosome 6 are predominantly represented  
 312 by African ancestry across all ethnicities, a striking example of which is seen in Chinese ethnicity,  
 313 where LINC03004 is almost completely African.

314 We observed that the different selection patterns of genes associated with the immune system were  
 315 related to distinct hierarchies of control of immune response, from control of gene expression to T cell  
 316 receptor recognition and inflammation. At the core were genes with low LDAS and AAS signals in  
 317 both the 26 population ancestries and the 5 continental ancestries. These genes affect RNS degradation  
 318 (PNRC2) and RNA splicing (SRSF10), and include a receptor that binds high-mannose structures on  
 319 the surface of potentially pathogenic viruses, bacteria, and fungi (MRC1). The product of these genes  
 320 affects immune responses (Supplementary Note 1.1), but their function is also central to non-immune  
 321 pathways, and mutations in these genes can give rise to, for example, various congenital disorders and  
 322 neurological and metabolic diseases.

323 The second level of control is broad-impact immune genes with low LDAS and AAS signals only

324 in the (more recently separated) 26 population ancestries. The product of these genes affects antigen  
325 presentation and the strength of receptor signalling. One of the genes (HLA-DRB1) presents antigens  
326 to T cells and helps regulate immune responses. Over 2000 variants of DRB1 have been identified<sup>35</sup>,  
327 some of which are associated with certain diseases or conditions (autoimmune diseases and susceptibil-  
328 ity or protection infection). Whilst HLA-DRB6 is a pseudogene with, as of now, no known function,  
329 SIRPB1 encodes a signal-regulatory-protein that interacts with TYROBP/DAP12, a transmembrane  
330 adaptor protein on natural killer (NK) cells, peripheral blood monocytes, macrophages, dendritic cells,  
331 osteoclasts, and microglia. Through this interaction, SIRPB1 is involved in regulating both adaptive  
332 and innate immune responses and other pathways.

333 The least-central control level primarily affects responses to specific infections (T cell recognition,  
334 signalling) or localized responses that occur at the site of infection (inflammation), and have low LDAS  
335 scores but no AAS signals. Among them are eight less-commonly expressed TRBV genes, which  
336 are noteworthy for well-established associations with globally widespread and ancient herpesviruses,  
337 bacteria, and old pathogens such as hepatitis virus B and C, and influenza<sup>36</sup>. The TRBV genes encode  
338 part of the beta chain, which, together with the alpha chain (encoded by TRAV), form the T cell  
339 receptor's antigen binding site. Notably, 8 TRBV but no TRAV genes are identified. SIRPG is a  
340 signal-regulatory protein (SIRP) member and is involved in the negative regulation of receptor tyrosine  
341 kinase-coupled signalling processes. It affects the signal regulatory protein gamma (SIRP $\gamma$ ) expression  
342 on T-cells and helps regulate immune responses, cell adhesion, and phagocytosis. IL20RA mediates the  
343 pro-inflammatory effects of IL-20 cytokines, helps to regulate immune responses, tissue homeostasis,  
344 and inflammation, and is a central player in the immune system. TMPRSS11E affects epithelial barrier  
345 function, inflammation and wound healing. Conversely, only two genes out of the 16 with only an AAS  
346 signal are associated with the immune system, as the CNR2 gene product has anti-inflammatory effects,  
347 among other non-immune related functions, and PDK1 is a key regulator of immune cell development  
348 and function.

## 349 Discussion

350 Local ancestry inference is fundamental to understanding the genetic history of admixed populations,  
351 and fundamentally all populations are admixed. Our study presents efficient tools for performing an-  
352 cestry inference that substantially enhance computational efficiency while retaining inference accuracy.  
353 This achievement comes from the observation that in large panels, relatively few matches are required  
354 to describe local ancestry, even in the presence of sequencing error, facilitating fine-scale haplotype  
355 analyses for large-scale projects that aim to paint thousands or even millions of individuals, such as the  
356 UK Biobank and the larger biobanks of the future.

357 Our tools are extensions of chromosome painting to describe genome-wide ancestries, and are not  
358 specifically designed for local ancestry inference. Scenarios where they might not be the optimal tool  
359 include local ancestry estimation with few populations, for which FLARE offers an edge in terms of

360 speed and memory with comparable accuracy. When reference panels are themselves very admixed,  
361 MOSAIC offers a two-stage HMM that allows the reconstruction of ancestries from imperfect ref-  
362 erence panels. Conversely, the efficiency of PBWTPaint for genome-wide ancestry estimation under  
363 *reference-vs-reference* painting makes it inaccurate at the level of local ancestry.

364 This work's broad implications extend beyond just technical improvement. The haplotype compo-  
365 nents (HCs) computed using PBWTPaint allow robust prediction of principal components (PCs) and  
366 may capture subtle genetic variations that PCs overlook - e.g. we found improved birthplace predic-  
367 tion performance within the UK Biobank. Haplotype summaries have other desirable properties such  
368 as not being associated with particular genomic regions, so replacing PCs with HCs is likely to result in  
369 a similar improvement as with ancestry components (ACs)<sup>30</sup>, which require comparison to a reference  
370 panel as SparsePainter is designed for. We therefore left a thorough examination of this task to future  
371 work and focused on the visualisation of population genetic structure using HCs at scale.

372 We presented a more in-depth exploration of two measures of selection at the ancestry level - LDAS  
373 which identifies ancestry segments that are too short (or too long), and AAS which identifies regions  
374 with unusual ancestry patterns. We have been careful to treat these as 'putative selection' when inter-  
375 preting them because there are other reasons for these anomalies to occur. LDAS and AAS would be  
376 sensitive to SNP density, long repeats, regions with many low-quality reads, or other structural issues.  
377 AAS is particularly sensitive to the makeup of the reference panel, which must be 'less admixed' than  
378 the target individuals on average to obtain a signal. LDAS is also sensitive to recombination map de-  
379 tails (though the recombination rate for each ethnicity is separately normalised). Although (as we have  
380 attempted) such issues are typically removed in quality control or by post-hoc considerations (low data  
381 volume regresses to the prior genome-wide ancestry), we know of no other methods that can confirm  
382 these types of selection on this timescale.

383 AAS has previously linked infection in admixed Scottish wildcat *Felis silvestris* to selectively retain  
384 an immune response developed in domestic cats *Felis catus*<sup>37</sup> over just 10 generations. Here, without  
385 looking specifically for it, we found many strong signals for core immune genes for all ethnicities using  
386 LDAS and AAS signals in the UK Biobank, which can be explained if there was a change in selection  
387 when these modern populations were formed as a mixture from older populations. Dating each would  
388 be very valuable - the admixture is only hundreds of years old for the African and European admixture  
389 seen in Afro-Caribbeans, and the last few thousand years for established populations described by 26  
390 inter-continental populations from the 1000 Genomes Project. This historical timescale is consistent  
391 with the continued expansion of populations and their pathogens around the globe and implies a 'melt-  
392 ing pot' of diverse diseases that evolved locally, likely related to environmental and cultural factors<sup>38</sup>  
393 and spread into global impact. For example, two selected immune genes (MRC1 and STAM) which  
394 have higher South Asian ancestries than expected facilitate the entry of the dengue virus, which is es-  
395 timated to have evolved approximately 500-1000 years ago and first became endemic in parts of South  
396 and South-East Asia<sup>39,40</sup>. Today, it is widespread globally, and its range continues to expand as global  
397 warming increases the mosquito habitat that carries the dengue virus. It remains to be seen if the signal

398 we see is this or some older virus that affects a related immune response.

399 It is hard to obtain ground truthing for selection statistics, and LDAS being relatively new and  
400 population-specific by design is no exception. We have attempted to rule out the most obvious con-  
401 founders - beyond the usual quality control, we removed low or heterogeneous SNP density regions,  
402 which preferentially removes regions near centromeres, telomeres and indels, as well as testing for  
403 GC bias and structural variation (Methods). The strongest evidence is the clear interpretability of the  
404 signal as being immunity-associated in all 7 ethnicities. Additional evidence is needed before coming  
405 to firm conclusions, but we believe that this strongly motivates more widespread investigation of local  
406 ancestry outside of the reconstruction of individual and population histories.

407 Our analysis suggested that varying genetic selective patterns prevailed at different levels of control  
408 of a hierarchical complex biological system such as the immune system. Using these methods with  
409 carefully constructed reference panels targeting particular admixture times, and the analysis of specific  
410 contact events, could eventually build the pathogenic landscape around the world, and bring insights  
411 into more diseases and traits selected in our recent ancestors.

## 412 **Methods**

### 413 **Modes of SparsePainter and PBWTpaint**

414 As in Figure 1, there are three modes of SparsePainter and one mode of PBWTpaint as below. The  
415 painting with a leave-one-out strategy (as required for GLOBETROTTER<sup>1</sup> and related methods) is  
416 classified as panel painting, which is only possible for SparsePainter.

417 (1) *all-vs-all*. Under this mode, we paint each individual against all the other individuals, i.e. only the  
418 individual itself is left out. This is for clustering, computing HCs, or similar tasks. PBWTpaint has the  
419 best performance of speed and can only operate in this mode.

420 (2) *reference-vs-reference* painting with  $n_{pop}$  populations. For exchangeability between a target and  
421 the reference, one individual is left out of each other population and oneself is left out from the own  
422 population. Then we paint a reference panel against itself. This ‘panel painting’ makes a palette for  
423 each of the  $n_{pop}$  populations as required for GLOBETROTTER<sup>1</sup>, NNLS, and related admixture esti-  
424 mation methods. SparsePainter is efficient for this.

425 (3) *target-vs-reference* with  $n_{pop}$  populations. We paint target individuals using a reference panel. We  
426 can optionally use leave-one-out painting (one individual is left out of each population) for admixture  
427 estimation, or without leave-one-out, we can do local ancestry inference. SparsePainter is efficient for  
428 this.

### 429 **Algorithm ‘ReportLongestMatches’**

430 The code implementation of the PBWT structure in SparsePainter drew extensively from Sanullah et  
431 al. (2021)<sup>24</sup>. We extend the ‘long match query’ algorithm of PBWT in Algorithm ‘ReportLongest-  
432 Matches’ which aims to find at least  $Q$  longest matches at each position for a target sample  $i$ , in a  
433



434 two-stage process. In stage 1 we ensure a minimum number of matches, by storing only matches of  
435 length  $L_{min}$  or longer containing SNPs with fewer than  $Q$  matches in a set  $\{s\}$ . For efficiency, we  
436 search the longest matches first, by iteratively halving the match length  $L_q$ , beginning from  $L_0$ . For  
437 every SNP that still has fewer than  $Q$  matches, all matches longer than  $L_q$  containing the SNP are  
438 added to  $\{s\}$ , until all positions have at least  $Q$  matches or the halved length falls below  $L_{min}$ .

439 Stage 2 reduces the number of matches. First, we calculate the genetic length for each match in  
440  $\{s\}$  and sort them in descending order of their genetic lengths. An empty set  $\{e\} = \emptyset$  is then populated  
441 with only the required matches. The algorithm traverses through the sorted  $\{s\}$ , adding a match to  $\{e\}$   
442 if any of its positions have fewer than  $Q$  matches in  $\{e\}$ . The final set  $\{e\}$ , containing elements that  
443 each specify the start position, end position, and reference sample number, represents the selected long  
444 matches to the reference haplotypes for the target sample  $i$ .

---

**Algorithm 1** ReportLongestMatches—find at least  $Q$  Longest matches at each position for target sample  $i$

---

**Stage 1:** Ensuring a minimum number of matches;

Run PBWT and record all matches longer than or equal to  $L_0$  SNPs in set  $\{s\}$ .

Let  $\mathbf{r}$  be a list of SNP indices with fewer than  $Q$  matches;

Iteration  $q \leftarrow 1$  and current minimum length  $L_q \leftarrow L_0/2$ ;

**while**  $|\mathbf{r}| \neq 0 \wedge L_q \geq L_{min}$  **do**

    Run PBWT and with min length  $L_q$  ;

    Add matches containing SNPs in  $\mathbf{r}$  with length  $\in [L_q, L_{q-1})$  to set  $\{s\}$ ;

    Update  $\mathbf{r}$  with the indices of SNPs with fewer than  $Q$  matches of length  $L_q$  or longer;

    Half the minimum match length  $L_q$  subject to constraints, i.e.  $L_{q+1} \leftarrow \max(L_q/2, L_{min})$ ;

$q \leftarrow q + 1$ ;

**end while**

**Stage 2:** Reduce the number of matches;

*▷ Retain only the longest, required matches.*

Compute the genetic distance of each match in set  $\{s\}$  and store in  $\{g\}$

Sort set  $\{s\}$  in descending order of  $\{g\}$ ;

Define  $\{e\}$  as an empty set to record final selected matches;

**for**  $b \leftarrow 1$  to  $|\mathbf{s}|$  **do**

    Add match  $s[b]$  to set  $\{e\}$  if it contains SNPs with fewer than  $Q$  matches;

    If all SNPs have at least  $Q$  matches **break**;

**end for**

Report  $\{e\}$  as the selected long matches to reference haplotypes for target sample  $i$ .

---

445 The efficiency of this algorithm is reflected by the majority of the genome being processed in Stage

446 1 with few long matches, even though there are huge numbers of matches throughout the genome.  
 447 Subsequently, we only need to proceed to search relatively short sections of genomes for few relatively  
 448 short matches.

449 Note that because of the limitation of  $L_{min}$ , we may end up having fewer than  $Q$  matches or even  
 450 no matches at specific positions. The former doesn't decrease the accuracy of local ancestry inference,  
 451 and we will address the latter in Methods – hidden Markov model.

452

### 453 **Hidden Markov model in vector form**

454 Let  $N$  be the number of haplotypes in the reference panel  $K$  be the number of SNPs, and  $\mu$  be the  
 455 mutation probability per SNP.  $\lambda$  is a recombination scaling constant, proportional to effective popu-  
 456 lation size in simple demographies and called  $N_e$  in<sup>18</sup>. The reference panel  $X$  is an  $N$  by  $K$  matrix,  
 457 and a target haplotype  $y$  is an  $K$ -vector, all taking values of either 0 or 1 corresponding to whether the  
 458 reference allele is present or not. However, we can simplify this into a **match matrix**  $M$  of dimension  
 459  $N \times K$  which also takes values of either 0 or 1, with  $M_{ij} = 1$  if  $X_{ij} = y_j$  and 0 otherwise. We will  
 460 refer to the row vectors  $\mathbf{m}_j = M_{.j}$  and use the shorthand  $D(\mathbf{x}) = \text{Diag}(\mathbf{x})$  as the matrix with the vector  
 461  $\mathbf{x}$  on the diagonal. We will refer to  $D_N(x)$  as an  $N \times N$  matrix with the scalar  $x$  on the diagonal.

462 SparsePainter implements the Li and Stephen's model<sup>17</sup> in the form of ChromoPainter<sup>18</sup> in a sparse  
 463 setting. We define  $\mathbf{V}$  as the emission matrix, and the column vectors are  $\mathbf{v}_j = V_{.j}$

$$\mathbf{V}_{ij} = \begin{cases} 1 - \mu & \text{if } M_{ij} = 1 \\ \mu & \text{if } M_{ij} = 0 \end{cases} \quad (1)$$

464 The observation matrix is an  $N \times N$  matrix:

$$\mathbf{O}_j = (1 - \mu)D_N(\mathbf{m}_j) + \mu D_N(\mathbf{1}_N - \mathbf{m}_j) = D_N(\mathbf{v}_j) \quad (2)$$

465 The transition matrix from position  $j$  to position  $j + 1$  is an  $N \times N$  matrix:

$$\mathbf{T}_j = \rho_j D_N(1) + \frac{1 - \rho_j}{N} \mathbf{1}_{N \times N} \quad (3)$$

466 where  $\rho_j = \exp(-\lambda g_j)$  with  $g_j$  being the genetic distance between position  $j$  and position  $j + 1$  in  
 467 Morgans.

468 Let  $f_0 = 1/N$  be the prior probabilities for the matches. We can write the forward probabilities for  
 469  $j = 1, \dots, K$  as:

$$\mathbf{f}_j = \mathbf{f}_{j-1} \mathbf{T}_{j-1} \mathbf{O}_j, \quad (4)$$

470 where  $\mathbf{f}_j$  are row vectors ( $1 \times N$ ). With  $\mathbf{b}_K = \mathbf{1}_N$  where  $\mathbf{1}_N$  is an  $1 \times N$  row vector, the backward  
 471 probabilities for  $j = 1, \dots, K - 1$  are:

$$\mathbf{b}_j^T = \mathbf{T}_j \mathbf{O}_{j+1} \mathbf{b}_{j+1}^T. \quad (5)$$

472 However, Equation (4) and (5) can be significantly simplified due to the special form of the output  
 473 and transition matrices. We can arrive at a **vector form** for which computations are  $O(N)$  instead of  
 474  $O(N^2)$ .

475 To simplify notation, we write the marginal (partial) probabilities  $\sum_{i=1}^N f_{ij} = \tilde{f}_j$  and  $\sum_{i=1}^N b_{ij} = \tilde{b}_j$ ,  
 476 the total number of matches  $\tilde{m}_j = \sum_{i=1}^N m_{ij}$ , and  $\tilde{\rho}_j = \frac{1-\rho_j}{N}$ . These are all scalar properties in what  
 477 follows below. For the forward probabilities:

$$\begin{aligned} \mathbf{f}_j &= \mathbf{f}_{j-1} [\rho_{j-1} D_N(1) + \tilde{\rho}_{j-1} \mathbf{1}_{N \times N}] [(1-\mu)D(\mathbf{m}_j) + \mu D(\mathbf{1}_N - \mathbf{m}_j)] \\ &= \mathbf{f}_{j-1} \rho_{j-1} [(1-\mu)D_N(\mathbf{m}_j) + \mu D(\mathbf{1}_N - \mathbf{m}_j)] + \tilde{f}_{j-1} \tilde{\rho}_{j-1} \mathbf{1}_N [(1-\mu)D(\mathbf{m}_j) + \mu D(\mathbf{1}_N - \mathbf{m}_j)] \\ &= (1-\mu)\mathbf{m}_j \circ [\rho_{j-1}\mathbf{f}_{j-1} + \tilde{f}_{j-1}\tilde{\rho}_{j-1}\mathbf{1}_N] + \mu [\rho_{j-1}\mathbf{f}_{j-1} + \tilde{f}_{j-1}\tilde{\rho}_{j-1}\mathbf{1}_N] \circ (\mathbf{1}_N - \mathbf{m}_j) \\ &= \mathbf{v}_j \circ [\rho_{j-1}\mathbf{f}_{j-1} + \tilde{f}_{j-1}\tilde{\rho}_{j-1}\mathbf{1}_N] \end{aligned} \quad (6)$$

478 where we use the notation  $\mathbf{x} \circ \mathbf{y}$  for entry-wise vector multiplication (Hadamard product). Similarly  
 479 for the backward probabilities, using the shorthand  $\mathbf{c}_j = \mathbf{m}_j \circ \mathbf{b}_j$  and  $\sum_{i=1}^N m_{ij} b_{ij} = \tilde{c}_j$ :

$$\begin{aligned} \mathbf{b}_j^T &= [\rho_j D_N(1) + \tilde{\rho}_j \mathbf{1}_{N \times N}] [(1-\mu)D_N(\mathbf{m}_{j+1}) + \mu D_N(\mathbf{1}_N - \mathbf{m}_{j+1})] \mathbf{b}_{j+1}^T \\ &= \rho_j [(1-\mu)D_N(\mathbf{m}_{j+1}) + \mu D_N(\mathbf{1}_N - \mathbf{m}_{j+1})] \mathbf{b}_{j+1}^T + \\ &\quad \tilde{\rho}_j \mathbf{1}_{N \times N} \left[ (1-\mu)(\mathbf{m}_{j+1} \circ \mathbf{b}_{j+1})^T + \mu \mathbf{b}_{j+1}^T \circ (\mathbf{1}_N - \mathbf{m}_{j+1})^T \right] \\ &= \rho_j (1-\mu) \mathbf{c}_{j+1}^T + \rho_j \mu (\mathbf{b}_{j+1}^T - \mathbf{c}_{j+1}^T) + \tilde{\rho}_j (1-\mu) \tilde{c}_{j+1} \mathbf{1}_N^T + \tilde{\rho}_j \mu (\tilde{b}_{j+1} - \tilde{c}_{j+1}) \mathbf{1}_N^T \\ &= \rho_j (\mathbf{c}_{j+1}^T - 2\mu \mathbf{c}_{j+1}^T + \mu \mathbf{b}_{j+1}^T) + \tilde{\rho}_j (\tilde{c}_{j+1} - 2\mu \tilde{c}_{j+1} + \mu \tilde{b}_{j+1}) \mathbf{1}_N^T \\ &= \rho_j \mathbf{d}_{j+1}^T + \tilde{\rho}_j \tilde{d}_{j+1} \mathbf{1}_N^T \end{aligned} \quad (7)$$

480 where  $\mathbf{d}_j = \mathbf{v}_j \circ \mathbf{b}_j$  and such that  $\sum_{i=1}^N v_{ij} b_{ij} = \tilde{d}_j$ . Finally, the posterior probabilities are written in  
 481 the following form:

$$P(\mathbf{m}_j | \mathbf{O}) \propto \mathbf{f}_j \circ \mathbf{b}_j^T. \quad (8)$$

482 If we assume the mutation rate  $\mu \rightarrow 0$ , the forward and backward probabilities (Equation (6) and  
 483 (7)) simplify to

$$\mathbf{f}_j = \mathbf{m}_j \circ [\rho_{j-1}\mathbf{f}_{j-1} + \tilde{f}_{j-1}\tilde{\rho}_{j-1}\mathbf{1}_N] \quad (9)$$

484 and

$$\mathbf{b}_j^T = \rho_j \delta_{j+1}^T + \tilde{\rho}_j \tilde{\delta}_{j+1} \mathbf{1}_N^T \quad (10)$$

485 respectively, where  $\delta_j = \mathbf{m}_j \circ \mathbf{b}_j$  and  $\tilde{\delta}_j = \sum_{i=1}^N m_{ij} b_{ij}$ . In this case, only the forward probabilities  $\mathbf{f}_j$   
 486 for the matched samples at position  $j$  are non-zero and need to be calculated. For backward probabilities  
 487 ties, we compute different  $\mathbf{b}_j^T$  for matched samples at position  $j+1$ , with unmatched samples sharing  
 488 the same default value  $\tilde{\rho}_j \tilde{\delta}_{j+1}$  in the  $j$ th hash vector. Finally, when computing the posterior proba-  
 489 bilities  $P(\mathbf{m}_j | \mathbf{O})$  (Equation (8)), only samples with matches in SNP  $j$  or  $j+1$  require computation,  
 490 whereas the others are exactly 0.

491 Note that this assigns non-zero probability to single mutation breaks in haplotypes, provided a  
 492 match is found both to the left and the right. In conclusion, the Hash-Map-based forward and backward  
 493 algorithm reduces computational cost from  $O(N)$  (e.g., ChromoPainter<sup>18</sup>) to approximately  $O(Q)$ .

494 There are instances when few positions have no matches spanning at least  $L_{min}$  SNPs, and are  
 495 therefore interpreted as no matches, which disrupts the forward and backward algorithm because a  
 496 0-vector of  $\mathbf{f}_j$  causes all  $\mathbf{f}_t$  to become 0-vectors for any  $t > j$ . To address this issue, for each position  
 497 without matches longer than  $L_{min}$  SNPs, we find the closest SNP (in genetic distance) that has matches.  
 498 We then impute the matches from this closest SNP to the position without matches.

499 The recombination scaling constant  $\lambda$  is usually estimated by the Expectation–Maximization (E-  
 500 M) algorithm (Supplementary Note 2.2). However, the Viterbi algorithm, a dynamic programming  
 501 technique to identify the most probable sequence of hidden states in a hidden Markov model, can be  
 502 advantageously employed to improve the efficiency of estimating  $\lambda$ , compared with the E-M algorithm.  
 503 In this context, let  $N_{seg}$  represent the minimum number of contiguous segments from different reference  
 504 samples required to construct the target haplotype, and therefore  $N_{break} = N_{seg} - 1$  is essentially the  
 505 number of distinct recombination events that have been inferred. Then  $\lambda$  is estimated as

$$\lambda^* = \frac{N_{break}}{\sum_{j=1}^K g_j}. \quad (11)$$

506

507

### 508 **The normalised versions of the forward and backward equations**

509 It is helpful to work in the normalised versions of the forward and backward equations  $\check{\mathbf{f}}_j = \mathbf{f}_j / \tilde{f}_j$  and  
 510  $\check{\mathbf{b}}_j = \mathbf{b}_j / \tilde{b}_j$ . We define  $F_j$  and  $B_j$  as the normalising constant at state  $j$ .

$$\frac{\mathbf{f}_j}{\tilde{f}_{j-1}} = \mathbf{m}_j \circ [(1 - \mu) (\rho_{j-1} \check{\mathbf{f}}_{j-1} + \tilde{\rho}_{j-1} \mathbf{1}_N) - \mu (\rho_{j-1} \check{\mathbf{f}}_{j-1} + \tilde{\rho}_{j-1} \mathbf{1}_N)] + \mu [\rho_{j-1} \check{\mathbf{f}}_{j-1} + \tilde{\rho}_{j-1} \mathbf{1}_N] \quad (12)$$

511 Setting  $\mu \rightarrow 0$ ,  $\mathbf{v}_j$  shrinks to  $\mathbf{m}_j$ :

$$\begin{aligned} \mathbf{f}_j &= \mathbf{m}_j \circ [\rho_{j-1} \mathbf{f}_{j-1} + \tilde{f}_{j-1} \tilde{\rho}_{j-1} \mathbf{1}_N] \\ \check{\mathbf{f}}_j &= \frac{\tilde{f}_{j-1}}{\tilde{f}_j} \frac{\mathbf{f}_j}{\tilde{f}_{j-1}} = \frac{\tilde{f}_{j-1}}{\tilde{f}_j} \mathbf{m}_j \circ [\rho_{j-1} \check{\mathbf{f}}_{j-1} + \tilde{\rho}_{j-1} \mathbf{1}_N] \\ &= \frac{1}{F_j} \mathbf{m}_j \circ [\rho_{j-1} \check{\mathbf{f}}_{j-1} + \tilde{\rho}_{j-1} \mathbf{1}_N] \end{aligned} \quad (13)$$

512 which has the following consequences:

513 (a) Let  $s_j$  be the set of matches at SNP  $j$ :  $i \in s_j \iff m_{ij} = 1$ .

514 (b)  $\check{f}_{ij}^* = \rho_{j-1} \check{f}_{i(j-1)} + \tilde{\rho}_{j-1}$  if  $i \in s_j$  and is zero otherwise.

515 (c)  $F_j = \sum_{i \in s_j} \check{f}_{ij}^*$  and  $\check{f}_{ij} = \check{f}_{ij}^* / F_j$ .

516 (d) for a sparse algorithm, we only need to track matches and the relative sums of their probabilities.

517 For the backward algorithm with  $\mu \rightarrow 0$ ,  $\mathbf{d}_j$  shrinks to  $\mathbf{c}_j$  :

$$\begin{aligned}\mathbf{b}_j^T &= \rho_j \mathbf{c}_{j+1}^T + \tilde{\rho}_j \tilde{\mathbf{c}}_{j+1} \mathbf{1}_N^T \\ \check{\mathbf{b}}_j^T &= \frac{\tilde{\mathbf{c}}_{j+1}}{\tilde{b}_j} [\rho_j \check{\mathbf{c}}_{j+1}^T + \tilde{\rho}_j \mathbf{1}_N^T]\end{aligned}\tag{14}$$

518 which has the following consequences:

519 (a)  $\check{b}_{ij}^* = \rho_j \check{b}_{i(j+1)} + \tilde{\rho}_j \tilde{\mathbf{c}}_{j+1}$  if  $i \in s_{j+1}$  and  $\check{b}_{ij}^* = \tilde{\rho}_j \tilde{\mathbf{c}}_{j+1}$  otherwise, where  $\tilde{\mathbf{c}}_{j+1} = \sum_{i \in s_{j+1}} b_{i(j+1)}$ .

520 (b)  $B_j = \sum_{i \in s_{j+1}} \check{b}_{ij}^* + (N - n_{j+1}) \tilde{\rho}_j \tilde{\mathbf{c}}_{j+1}$  and  $\check{b}_{ij} = \check{b}_{ij}^* / B_j$ .

521 (c) Again this can be computed without explicit reference to non-matches and we need to sum over  
522 only matches.

523

### 524 Estimation of the expected length of copied chunks

525 Let  $\hat{l}_i$  denote the posterior expected length (in Morgans) of the total genome for which the sample  
526 haplotype copies from the  $i$ th reference haplotype.

$$\begin{aligned}\hat{l}_i &= \frac{1}{2 \Pr(D)} \sum_{j=1}^{K-1} g_j [f_{ij} b_{ij} + f_{i(j+1)} b_{i(j+1)}] \\ &= \frac{1}{2 \prod_{k=1}^K F_k} \sum_{j=1}^{K-1} g_j \left[ \check{f}_{ij} \check{b}_{ij} \left( \prod_{k=1}^j F_k \right) \left( \prod_{k=j}^K B_k \right) + \check{f}_{i(j+1)} \check{b}_{i(j+1)} \left( \prod_{k=1}^{j+1} F_k \right) \left( \prod_{k=j+1}^K B_k \right) \right] \\ &= \frac{1}{2} \sum_{j=1}^{K-1} g_j [w_j^l \check{f}_{ij} \check{b}_{ij} + w_j^r \check{f}_{i(j+1)} \check{b}_{i(j+1)}]\end{aligned}\tag{15}$$

where

$$w_j^l = \exp \left( \log \left( \prod_{k=1}^j F_k \right) + \log \left( \prod_{k=j}^K B_k \right) - \log \left( \prod_{k=1}^K F_k \right) \right)$$

and

$$w_j^r = \exp \left( \log \left( \prod_{k=1}^{j+1} F_k \right) + \log \left( \prod_{k=j+1}^K B_k \right) - \log \left( \prod_{k=1}^K F_k \right) \right).$$

527

528

### 529 Estimation of the expected number of chunks copied

530 Let  $\hat{c}_i$  denote the posterior expected number of chunks copied from the  $i$ th reference haplotype.

$$\begin{aligned}
 \hat{c}_i &= \frac{1}{\Pr(D)} f_{i1} b_{i1} + \frac{1}{\Pr(D)} \sum_{j=1}^{K-1} [f_{i(j+1)} b_{i(j+1)} - f_{ij} b_{i(j+1)} V_{i(j+1)} \rho_j] \\
 &= \frac{1}{\prod_{k=1}^K F_k} \check{f}_{i1} \check{b}_{i1} F_1 B_1 + \\
 &\quad \frac{1}{\prod_{k=1}^K F_k} \sum_{j=1}^{K-1} \left[ \check{f}_{i(j+1)} \check{b}_{i(j+1)} \left( \prod_{k=1}^{j+1} F_k \right) \left( \prod_{k=j+1}^K B_k \right) - \check{f}_{ij} \check{b}_{i(j+1)} \left( \prod_{j=1}^j F_j \right) \left( \prod_{k=j+1}^K B_k \right) V_{i(j+1)} \rho_j \right] \quad (16) \\
 &= \frac{1}{\prod_{k=1}^K F_k} \check{f}_{i1} \check{b}_{i1} F_1 B_1 + \sum_{j=1}^{K-1} [a_j^l \check{f}_{i(j+1)} \check{b}_{i(j+1)} - a_j^r \check{f}_{ij} \check{b}_{i(j+1)} V_{i(j+1)} \rho_j]
 \end{aligned}$$

where

$$a_j^l = \exp \left( \log \left( \prod_{k=1}^{j+1} F_k \right) + \log \left( \prod_{k=j+1}^K B_k \right) - \log \left( \prod_{k=1}^K F_k \right) \right)$$

and

$$a_j^r = \exp \left( \log \left( \prod_{k=1}^j F_k \right) + \log \left( \prod_{k=j+1}^K B_k \right) - \log \left( \prod_{k=1}^K F_k \right) \right).$$

531

532

### 533 **Non-negative Least Squares (NNLS) for admixture estimation**

534 Admixture estimation can be performed on both the reference individuals and the target individuals  
 535 via NNLS, which requires the expected total genome shared between each reference ancestry, and  
 536 each reference (or target) individual with each reference ancestry. The former is derived by paint-  
 537 ing the reference samples against themselves with one sample left out of each other population (i.e.  
 538 *reference-vs-reference* painting). We then average the expected length of copied chunks for each ref-  
 539 erence individual within each reference ancestry to provide a reference palette. When investigating  
 540 admixture estimation for target individuals, we also require painting each target sample (i.e. *target-vs-*  
 541 *reference* painting) against a reference panel, with one sample left out from every reference ancestry.  
 542 Reference (target) samples are then described as a mixture of the reference ancestries using NNLS,  
 543 calculated by the R package ‘npls’<sup>1</sup>. In detail, we fit the NNLS model by minimising  $\|Ax - b\|_2$  with  
 544 the constraints  $x \geq 0$ , where  $A$  is the reference palette and  $b$  is the expected length of copied chunks  
 545 for each reference (target) sample, and finally obtain the estimates  $x$ .

546

### 547 **Simulation details for comparison between SparsePainter, PBWTpaint, ChromoPainter and FLARE**

548 We simulated different simple models (Simulation 2a-c) for *target-vs-reference* painting, and a hierar-  
 549 chical model (Simulation 1) for *reference-vs-reference* painting. Each simulation is repeated 10 times,  
 550 and the average statistics, i.e. compute time, memory usage and accuracy, are reported.

551 The simple simulation model for *target-vs-reference* painting (Simulation 2a-c) begins with an  
 552 ancestral population of 50000 individuals that evolved for 2500 generations prior to diverging into  $n_{pop}$

553 populations with different sizes. Here we specify the population sizes for Simulation 2b-c, including  
554  $n_{pop} = 2, 3, 5$ :

- 555 (1) 5000 and 20000 for 2-way admixture model;
- 556 (2) 5000, 15000 and 25000 for 3-way admixture model;
- 557 (3) 5000, 10000, 7000, 14000 and 9000 for 5-way admixture model.

558 Following an evolutionary period of another 500 generations, these  $n_{pop}$  populations admixed into  
559 1000 modern individuals with different proportions. Again, we specify the admixture proportions for  
560 Simulation 2b-c:

- 561 (1) 50% and 50% for 2-way admixture model;
- 562 (2) 20%, 50% and 30% for 3-way admixture model;
- 563 (3) 20%, 10%, 10%, 40% and 20% for 5-way admixture model.

564 The admixed individuals had a growth rate of 5% per generation, and they were sampled 13 generations  
565 after admixture.

566 For Simulation 1, we constructed a hierarchical model that mirrors the evolutionary trajectory of  
567 real-world populations, which is used for the comparison of *reference-vs-reference* panel painting. We  
568 simulated a 5-population and a 10-population hierarchical model. Here we illustrate the 5-population  
569 ( $P_i (i = 1, 2, \dots, 5)$ ) model in detail. After an ancestral population with 10000 individuals evolved for  
570 2700 generations, it split into  $P_1$  and  $P_4$  with 7000 and 3000 individuals. After generation 2890,  $P_2$   
571 emerged from migrations originating from  $P_1$  with a population size of 3000. Moving forward to the  
572 2940th generation, 1000 people from  $P_2$  migrated to a new population  $P_3$ . A final migration occurred  
573 at the 2950th generation when 2000 individuals from  $P_4$  settled to create  $P_5$ . All the populations  
574 had a growth rate of 5% from the 2970th to the 3000th generation. At the 3000th generation, we  
575 sampled an equivalent number of individuals (20, 40, 60, 100, 180, 300 and 500) from each population  
576  $P_i (i = 1, 2, \dots, 5)$ . A similar model was constructed for simulating 10 hierarchical populations.

577 Because all methods considered are linear in genome length, all simulations (Simulation 1 and  
578 Simulation 2a-c) use 20 megabases (Mb) of genome, characterized by a mutation rate of  $1.44 \times 10^{-8}$   
579 per base pair per generation, a recombination rate of  $1 \times 10^{-8}$  Morgans per base pair per generation.  
580 Following Browning et al. (2023)<sup>20</sup>, we included gene conversion at twice the recombination rate with  
581 an average tract length of 300 base pairs, and genotype error with a proportion of 0.02%. We retained  
582 20k SNPs with Minor Allele Frequency (MAF)  $\geq 1\%$  shared between the reference and target datasets.

583 The true local ancestry is defined as 5 generations before admixture, which is derived from the  
584 recombination events recorded in the tree sequences (in SLiM) during the 500 generations before  
585 admixture. Some regions (around 10%-20%) in target haplotypes were inherited from the ancestral  
586 population and haven't experienced any recombination events during the 500 generations. As in<sup>20</sup>, to  
587 compare the local ancestry estimates we excluded the SNPs within those regions, but we emphasise  
588 that the 'true' local ancestry of these regions can only be defined in terms of a mixture of the descen-  
589 dent populations. Genome-wide ancestry estimates are obtained by summing the probabilities as in  
590 ChromoPainter<sup>18</sup>.

591 For all the simulations, we also retained 20,000 common SNPs with Minor Allele Frequency  
592 (MAF) of at least 1% from the reference and target datasets presented in the Variant Call Format  
593 (VCF). In detail, all simulations generated more than 20,000 SNPs after MAF filtering, and we sam-  
594 pled 20,000 random SNPs from them for analysis. Subsequently, we merged the reference and target  
595 datasets and phased the merged dataset with Beagle 5.4<sup>41</sup> before splitting it into the reference and tar-  
596 get datasets. FLARE requires input data in VCF format, while ChromoPainter requires phase format,  
597 and SparsePainter and PBWTPaint enable both input formats (we used the phase format). Phase format  
598 can be converted from VCF efficiently with PBWT.

599 For SparsePainter, unless otherwise stated, we ensured no more than 10 longest matches (longer  
600 than 20 SNPs) at each locus are retained. All simulations are performed on an MSI laptop with an Intel  
601 Core i7-10750H processor running at 2.60GHz on 10 CPU cores in parallel.

602 We explored a number of different parameters for Simulation 2a-c.

- 603 • **Simulation 2a:** we simulated 2-, 5-, 10-, 20-, 40-, 60-, 80- and 100-way admixture ( $n_{pop} = 2, 5, 10,$   
604  $20, 40, 50, 80$  and  $100$ ) models to compare the speed and memory of painting 50 admixed individuals  
605 between software, with varying numbers of total reference sizes (2000, 4000 and 8000) with random  
606 numbers of (at least 10) individuals per reference ancestry.
- 607 • **Simulation 2b:** we simulated 2-, 3- and 5-way admixture ( $n_{pop} = 2, 3$  and  $5$ ) models to compare the  
608 local ancestry inference accuracy of 50 admixed individuals between software, with varying numbers  
609 of reference sizes for each reference ancestry (100, 200, 500, 1000 and 2000).
- 610 • **Simulation 2c:** we drew from reference pools of 1000, 2000, or 4000 individuals for each of the  
611  $n_{pop} = 5$  reference ancestries. We then evaluated SparsePainter's efficiency in painting 1000 ad-  
612 mixed individuals under varying levels of sparsity, i.e. only the longest 5, 10, 20, 40 and 80 matches  
613 which are longer than 20 SNPs are retained at each SNP. This was manipulated via the 'nmatch'  
614 parameter in SparsePainter.

### 615 **Methods to evaluate the accuracy of local ancestry and NNLS estimates**

616 We used two different methods to assess the accuracy of local ancestry estimates. The first method is  
617 the squared Pearson's correlation coefficient (denoted as  $r^2$ ). At each SNP, we calculated the estimated  
618 dosage of each individual by averaging the posterior probabilities of both haplotypes for each reference  
619 ancestry, and the true dosage is the average true local ancestry which takes values of 0, 0.5, or 1.  
620 We computed the  $r^2$  between the estimated and actual dosages for each reference ancestry across all  
621 individuals and positions, and the unweighted mean  $r^2$  of these values is reported to measure the overall  
622 accuracy. The second method evaluates the proportion of accurate local ancestry predictions across all  
623 haplotypes and positions. For each haplotype at a specific position, a correct local ancestry inference  
624 is determined when the true local ancestry corresponds to the highest estimated posterior probability,  
625 i.e. the best-guess strategy.

626 To evaluate the accuracy of admixture estimation, we calculated the squared correlation between  
627 the NNLS-estimated coefficient (see above) and the true proportion for all the individuals, and reported



628 the unweighted mean  $r^2$  of NNLS from different populations.

629

### 630 **The accuracy of PBWTpaint for local ancestry estimation**

631 Unlike SparsePainter, PBWTpaint does not provide a calibrated estimate of local ancestry. To assess  
632 this, we compare local ancestry estimates under *reference-vs-reference* panel painting. On the simple  
633 simulation model (Simulation 2a-2c) in which the ancestries are distinct, the  $r^2$  between PBWTpaint  
634 and SparsePainter is high at 0.79. However, for complex cases in which there is uncertainty, or the  
635 true ancestry is an ancestor of extant populations (Simulation 1), the set maximal matches used by PB-  
636 WTpaint lead to over-confident or inaccurate local ancestry assignment ( $r^2 = 0.3$ ) even though these  
637 mistakes are self-averaging for the estimation of genome-wide ancestry. This illustrates that PBWT-  
638 paint is not an appropriate method for performing local ancestry estimates.

639

### 640 **Paint all UK Biobank individuals against themselves and calculate haplotype principal compo- 641 nents**

642 To infer the haplotype principal components, we painted UKB biobank individuals against themselves,  
643 i.e. *all-vs-all* painting. We first excluded related individuals as described by Bycroft et al. (2018)<sup>33</sup> and  
644 excluded withdrawn individuals. We then performed PBWTpaint (with command `pbwt -paintSparse`)  
645 on each chromosome of UK Biobank phased genotype data, which in total has 406,733 individu-  
646 als with approximately 569,200 SNPs. The total chunk length of PBWTpaint for each individual on  
647 chromosome  $i$  is  $2K_i$ , where  $K_i$  is the number of SNPs. Assume  $g_i$  is the total genetic distance for  
648 chromosome  $i$ , we weighted the chunk length for chromosome  $i$  with weight  $g_i/K_i$ . Then we summed  
649 up the sparse chunk length matrix for all the chromosomes as matrix  $A$ , such that for each individual  
650 (i.e. each row of  $A$ ), the expected lengths of copied chunks from all other individuals reached the sum  
651 of the total genetic distance  $G = \sum_{i=1}^{22} g_i$ .

652 We performed singular value decomposition (SVD) on the log-transformed sparse chunk length  
653 matrix  $\log_{10}(A + 1)$  with R package ‘`sparsesvd`’:  $\log_{10}(A + 1) = UDV^T$ , where  $D$  is a diagonal  
654 matrix of the singular values. Then we extracted the the first 150 columns of  $U\sqrt{D}$  as the top 150  
655 haplotype principal components.

656

### 657 **Prediction of birth locations with HCs and PCs**

658 We conducted an analysis to evaluate the predictive accuracy of Haplotype Components (HCs) and  
659 Principal Components (PCs) on the birth locations, i.e. the east and north coordinates, within the UK.  
660 We selected a cohort of 347,532 individuals who were born in the UK or Ireland and identified as  
661 white, British, or Irish ethnicity. This cohort was divided into two groups: a training set comprising  
662 80% of the individuals, and a test set consisting of the remaining 20%. Subsequently, with either the  
663 top 150 PCs or HCs as explanatory variables and either the east or north coordinate as the response  
664 variable, we used a 5-fold CV to determine the optimal number of boosting iterations before fitting  
665 the regression model on the training set with eXtreme Gradient Boosting (XGBoost<sup>42</sup>), and then we

666 predicted the birth coordinates of individuals in the test set. Finally, we computed the direct distance  
 667 between the predicted coordinates and the actual coordinates of each individual on the test set and  
 668 reported the median which reveals that using HCs as predictors (median error=39.7km) reduced 49%  
 669 error compared with using PCs as predictors (median error=77.4km). This indicates a notably higher  
 670 predictive accuracy of birthplaces when using HCs.

671

## 672 **Paint UK Biobank with 1000 Genomes Project**

673 We inferred the local ancestry of UK Biobank individuals using the 1000 Genomes Project (1000GP) as  
 674 the reference data, which includes 2504 individuals from 26 populations. We retained the common bi-  
 675 allelic SNPs with  $MAF \geq 5\%$  before merging these two datasets. Then we used Beagle 5.4<sup>41</sup> to phase  
 676 the merged dataset, after which it was split into the reference and target datasets. For a comparative  
 677 analysis of the genetic painting and population structure within the UK Biobank, we randomly selected  
 678 10,000 individuals with self-reported British backgrounds, and incorporated all individuals from spe-  
 679 cific self-reported ethnic backgrounds: Irish (12713), Indian (5660), Caribbean (4297), African (3203),  
 680 Pakistani (1747), and Chinese (1503).

681 We estimated the average recombination scaling constant  $\lambda = 164.2$  of all these individuals on  
 682 chromosome 19. This fixed parameter was subsequently used for painting across chromosomes 1-22.  
 683 We configured the parameters of SparsePainter to aim for finding the 50 longest matches (longer than  
 684 20 SNPs) at each position.

685

## 686 **Quality control for shared and ethnicity-specific LDAS and AAS**

687 Here we explain the method for finding shared and ethnicity-specific LDAS and AAS, and the addi-  
 688 tional Quality Controls (QC) applied. As introduced by Barrie et al. (2024)<sup>12</sup>, we compute the LDAS  
 689 of SNP  $j$  in principle as the integral of the LDA between every other position genome with  $g_j$ , over the  
 690 recombination map with length  $L_j$  consisting of the chromosome holding the  $j$ -th SNP:

$$LDAS(j) = \int_0^{L_j} LDA(g, g_j) dg. \quad (17)$$

691 In practice the pairwise LDA shrinks to almost 0 when the closest SNPs are more than 3 centiMor-  
 692 gan (cM) away, so the integral is approximated over a  $X = 4cM$  window as  $LDAS(j; X)$  by:

$$LDAS(j; X) = \begin{cases} \int_{g_j-X}^{g_j+X} LDA(g, g_j) dg & \text{if } X \leq g_j \leq L_j - X, \\ \int_0^{g_j+X} LDA(g, g_j) dg + \int_{2g_j}^{g_j+X} LDA(g, g_j) dg & \text{if } g_j < X, \\ \int_{g_j-X}^{L_j} LDA(g, g_j) dg + \int_{g_j-X}^{2g_j-L_j} LDA(g, g_j) dg & \text{if } g_j > L_j - X. \end{cases} \quad (18)$$

693 where  $g_j$  is the genetic position in centiMorgan for the  $j$ th SNP, and  $LDA(g, g_j)$  is the LDA between  
 694 position  $g$  and the target SNP at position  $g_j$ .

695 Because LDA can only be computed at discrete SNPs, in practice these integrals are approximated,  
 696 which leads to an error that must be controlled. If the SNPs present are random with respect to the

697 true recombination locations, then the lowest mean-square-error estimate of LDAS in Equation (18)  
698 integral treats LDA as a piecewise linear function:

$$\text{LDA}(g, g_j) = (1 - \alpha)\text{LDA}(g_i, g_j) + \alpha\text{LDA}(g_{i+1}, g_j),$$

699 where  $\alpha = (g - g_i)/(g_{i+1} - g_i) \in [0, 1)$  for  $g \in [g_i, g_{i+1})$ . Further, an upper bound and lower bound  
700 can be obtained by replacing the piecewise linear function with a step function. In detail, we take the  
701 larger and smaller LDA values of two neighbouring SNPs, respectively, as the fixed LDA in the genetic  
702 distance between two SNPs  $i$  and  $j$  in the integral:

$$\text{LDA}_{\text{upper}}(g, g_j) = \max \{ \text{LDA}(g_i, g_j), \text{LDA}(g_{i+1}, g_j) \}$$

703 and

$$\text{LDA}_{\text{lower}}(g, g_j) = \min \{ \text{LDA}(g_i, g_j), \text{LDA}(g_{i+1}, g_j) \}.$$

704 These estimates are substituted into Equation 18 to obtain an upper and lower bound respectively  
705 of the LDAS of SNP  $j$ . When computing  $\text{LDAS}_{\text{lower}}(j; X)$ , we assume the chromosome ends have  
706 zero LDA with the target SNP, i.e.  $\text{LDA}(0, g_j) = \text{LDA}(L_j, g_j) = 0$  for conservative estimation.

707 The maximum possible error in the LDAS estimate at SNP  $j$  is

$$\text{LDAS}_{\text{error}}(j; X) = \text{LDAS}_{\text{upper}}(j; X) - \text{LDAS}_{\text{lower}}(j; X). \quad (19)$$

708 It is necessary to account for different scales of LDAS across different ethnic backgrounds, because  
709 of different admixture times with respect to the populations in the panel. Therefore, for each ethnic  
710 background, we normalise the  $\text{LDAS}_{\text{error}}$  with the average LDAS across the genome, i.e.  $\text{LDAS}_{\text{error}}^* =$   
711  $\text{LDAS}_{\text{error}}/E(\text{LDAS})$ . Finally, in the QC we remove SNPs with large relative error, i.e.  $\text{LDAS}_{\text{error}}^* \geq \delta$   
712 where  $\delta$  is a specified threshold (we used  $\delta = 0.3$ ). This provides an implicit condition of high SNP  
713 density with respect to the recombination map.

714 A final challenge is that no LDA can be detected if SNPs are very sparse, so that  $\text{LDAS}_{\text{upper}}$   
715 is estimated near zero and the error is undefined. We therefore remove SNPs if any nearby 0.5cM  
716 region within 3cM has too few SNPs: SNP  $j$  is removed if at least one of  $n_m(j) < \theta$  for  $m =$   
717 0.5, 1, 1.5, 2, 2.5, 3, where  $n_m(j)$  is the number of SNPs that is  $(m - 0.5, m]$ cM away from SNP  $j$  and  
718  $\theta$  is a specified threshold (we used  $\theta = 10$ ).

719 In conclusion, we use two additional filters; firstly that  $\text{LDAS}_{\text{error}} < \delta$  and  $n_m \geq \theta$  ( $m=0, 0.5,$   
720 1.0, 1.5, 2, 2.5) as the quality control of SNPs, which alleviates the bias estimates due to sparsity  
721 of the painting data and therefore avoids extreme LDA scores. In practice this removes 3.5% of the  
722 genome (20,075 out of 569,200 SNPs) in 62 contiguous segments (see Supplementary Tab. 2-8 for  
723 detail). Because of the SNP selection process inherent in the UK Biobank genotyping chip, these are  
724 predominantly centromeres, telomeres, and regions that already have SNPs removed due to standard  
725 QC procedures, including where there are missing data due to e.g. indels, alignment issues, etc.

726 The computation of AAS is not affected by the discrepancy of recombination events across chro-  
727 mosomes and ethnicities, and we implemented the procedures as described in Barrie et al. (2024)<sup>12</sup>  
728 with SparsePainter.

729 As validated through simulation, we assume normality of LDAS for all ethnicities across the  
730 genome. We converted the LDAS into p-values through the one-sided normality test which aims to  
731 detect low LDAS, and we only focused on SNPs with LDAS from at least one ethnic background that  
732 is significant at  $p = 10^{-6}$ . Those SNPs are classified as shared or ethnicity-specific low LDAS if  
733 LDAS from all the other ethnic backgrounds are significant at  $p = 0.05$ , or insignificant at  $p = 0.1$ ,  
734 respectively.

735 As AAS approximately follows a Gamma distribution and produces more extreme p-values (through  
736 the one-sided Gamma test), we employed a stricter significance level,  $p = 10^{-50}$ , for filtering SNPs  
737 with significant AAS. Similarly, those SNPs are categorized as having shared or ethnicity-specific sig-  
738 nificant AAS if AAS from all the other ethnic backgrounds are significant at  $p = 10^{-10}$ , or insignificant  
739 at  $p = 10^{-5}$ , respectively.

740 Furthermore, to ensure robust results, we repainted UKB using 5 continental populations as de-  
741 linedated by the 1000GP continents (Europe, Africa, America, South Asia and East Asia) to obtain an  
742 alternate set of LDAS and AAS results. We then mapped each SNP with low LDAS and AAS signals  
743 to its gene (if the SNP overlaps with a gene) via R package ‘gprofiler2’, and visualised the results in  
744 Fig. 5 and Fig. 6.

745 To ensure the validity of LDAS and AAS signals, we evaluated their association with GC bias.  
746 Using the GC frequency reported for East Asia, Europe and Africa<sup>43</sup>, we found that all the regions  
747 with shared LDAS or AAS signals had random frequencies of G+C (Supplementary Fig. 3), which  
748 showed no evidence of association with GC bias. We also checked the association of LDAS and AAS  
749 signals with structural variation (SV). We downloaded the regions with SV in 1000GP<sup>44</sup>, and found  
750 this covers 3.91% of the whole genome. For SNPs with LDAS or AAS signals, we classified them into  
751 various small regions which are no longer than 10kb, and computed the proportion of these regions that  
752 have SV. We detected 7.45%, 7.04%, 7.48% and 8.89% of the regions with 26-pop LDAS, 5-continent  
753 LDAS, 26-pop AAS, and 5-continent AAS signals, respectively. Whilst it is plausible that selection  
754 acts on SV and LDAS jointly, we cannot rule out reverse causation of SV causing an LDAS or AAS  
755 signal without selection. Therefore those regions with SV were excluded from further analysis, though  
756 this choice does not materially affect the conclusions (see Supplementary Tab. 9 for a list of SNPs  
757 affected).

758

### 759 **Simulation for LDAS under genetic drift**

760 We assessed the robustness of the LDAS and its sensitivity to demographic changes by examining it  
761 under genetic drift across exponentially expanding population sizes over time. We simulated a genome  
762 of a 500Mb region as follows: initially, an ancient population evolves for 1000 generations, subse-  
763 quently diverging into five distinct subpopulations. Each of these subpopulations, growing at a rate of

764 2% per generation, evolves independently for 100 generations. This period of divergence is followed  
765 by a phase of admixture, forming a modern, unified population, which then undergoes evolution for an  
766 additional 30 generations at an increased growth rate of 5% per generation.

767 We computed the LDAS of 500 simulated modern individuals with 2000 simulated reference indi-  
768 viduals from each of the 5 subpopulations. After normalisation, the z-scores of the LDAS (Extended  
769 Data Fig. 4a) predominantly exhibit under-dispersion, despite some noticeable deviation on both tails.  
770 This pattern suggests that the normal distribution is a reasonable approximation for the LDAS distribu-  
771 tion. Subsequently, we calculated the p-values for low LDAS through a one-sided test for normality, as  
772 depicted in Extended Data Fig. 4b. Notably, no low LDAS signals are detected under the genetic drift  
773 model (excluding selection effects), as evidenced by the most significant SNP with  $p < 10^{-3}$  through  
774 the one-sided normality test. This outcome solidifies our conclusion that low LDAS signals are not  
775 present under this model.

776

### 777 **Simulation for comparing LDAS with statistics for positive selection**

778 Here we simulated the similar two-loci and one-locus model as used in Barrie et al. (2024)<sup>12</sup>.

779 For the two-loci selection model (Extended Data Fig. 5), we simulated a genome of 150Mb. Ini-  
780 tially, an ancient population evolved for 2200 generations before splitting into two sub-populations  
781  $P1$  and  $P2$ . After evolving 400 generations, we added mutation  $m1$  for  $P1$  and  $m2$  for  $P2$  at locus  
782 20Mb and 23Mb, respectively. These added mutations were then positively selected in the following  
783 300 generations before admixing to  $P3$  at generation 2900.  $m1$  and  $m2$  then experienced strong pos-  
784 itive selection for another 50 generations, after which we sampled 500 individuals from  $P3$  as target  
785 individuals. 500 individuals are sampled for  $P1$  and  $P2$  at generation 2899 as the reference panel.

786 For the one-locus selection model (Extended Data Fig. 6), we simulated a genome of 50Mb.  
787 The remaining difference from the above mode is that only one locus  $m0$  at 20Mb was added at  
788 generation 2601 for both  $P1$  and  $P2$ , and it was positively selected until generation 2900. In the  
789 admixture population  $P3$ , this SNP underwent negative selection until generation 2950 when the target  
790 individuals were sampled.

791 Both simulations had a mutation rate of  $1.44 \times 10^{-8}$  per base pair per generation, and a recombi-  
792 nation rate of  $1 \times 10^{-8}$  Morgans per base pair per generation.

793

### 794 **Comparison of LDAS and AAS signals with natural selection in Bronze Age Britain and archaic 795 adaptive introgression in 1000GP populations**

796 Our LDAS and AAS analyses from painting 7 UK Biobank ethnic backgrounds with 1000GP pop-  
797 ulations have detected various signals of selection (Fig. 6 and Supplementary Tab. 1), and we in-  
798 vestigated the overlaps with the other selection signals. By comparison with the genome-wide sig-  
799 nificant ( $P < 10^{-7}$ ) selection signals in the ancient British data<sup>45</sup>, we found the only overlap genes  
800 are HLA-DRB6 and HLA-DRB1 on chromosome 6. We compared loci that have been identified as  
801 exhibiting adaptive introgression from Neanderthal or Denisovan ancestries in the 1000GP popula-

802 tions<sup>46</sup>. Although none of them overlaps genes with LDAS signals, we discovered that the ADARB2  
803 gene, located on chromosome 10 overlaps with AAS signals. This gene experiences introgression from  
804 Denisovan ancestry within the 1000GP PEL population, and coincides with the AAS signals in British,  
805 Irish, Indian, Caribbean and Chinese ethnicities. Notably, the utilization of different reference panels  
806 can probably lead to the identification of distinct genes exhibiting selection signals of LDAS and AAS.

## 807 **Data availability**

808 The phased 1000 Genomes Project data build GRCh37/hg19 are available at [https://bochet.gcc.biostat.washington.edu/beagle/1000\\_Genomes\\_phase3\\_v5a/b37.vcf/](https://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5a/b37.vcf/). The genetic map build GRCh37/hg19  
809 is available from [https://bochet.gcc.biostat.washington.edu/beagle/genetic\\_maps/](https://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/). The UK Biobank  
810 data can be accessed by approved researchers through <https://www.ukbiobank.ac.uk>. We used the UK  
811 Biobank data under project 81499. The UK map data are available at <https://gadm.org>.  
812

## 813 **Code availability**

814 The C++ code for SparsePainter is available on GitHub at <https://github.com/YaolingYang/SparsePainter>,  
815 and the website for SparsePainter is at <https://sparsepainter.github.io/>. PBWTPaint is available on  
816 GitHub at <https://github.com/richarddurbin/pbwt>. The UK Biobank painting pipeline and methods  
817 to compute haplotype components (HCs) are available on GitHub at <https://github.com/YaolingYang/SparsePainter/tree/main/painting-pipeline>.  
818

## 819 **Acknowledgements**

820 We thank the participants in the UK Biobank (UKB) and 1000 Genomes Project (1000GP). Y.Y. was  
821 supported by China Scholarship Council [grant number 202108060092]. This work was carried out  
822 using the computational facilities of the Advanced Computing Research Centre, University of Bristol  
823 - <http://www.bris.ac.uk/acrc>.

## 824 **Author contributions**

825 Y.Y., R.D. and D.J.L. conceived and designed the project and methodology. D.J.L. supervised the  
826 project. Y.Y., R.D. and D.J.L. developed the methodology. R.D. and D.J.L. programmed the codes  
827 for PBWTPaint. Y.Y. programmed the codes for SparsePainter. Y.Y. did simulations and UKB data  
828 analysis under the supervision of D.J.L. A.K.N.I. analysed and interpreted genes with LDAS and AAS  
829 signals. Y.Y. wrote the initial manuscript draft. All authors wrote, reviewed, discussed and revised the

830 subsequent versions of the manuscript (led by Y.Y. and D.J.L.). Y.Y. and A.K.N.I. wrote the Supple-  
831 mentary Information. All authors agreed with the submitted manuscript.

## 832 **Competing interests**

833 All authors have declared no competing interests.

## 834 **References**

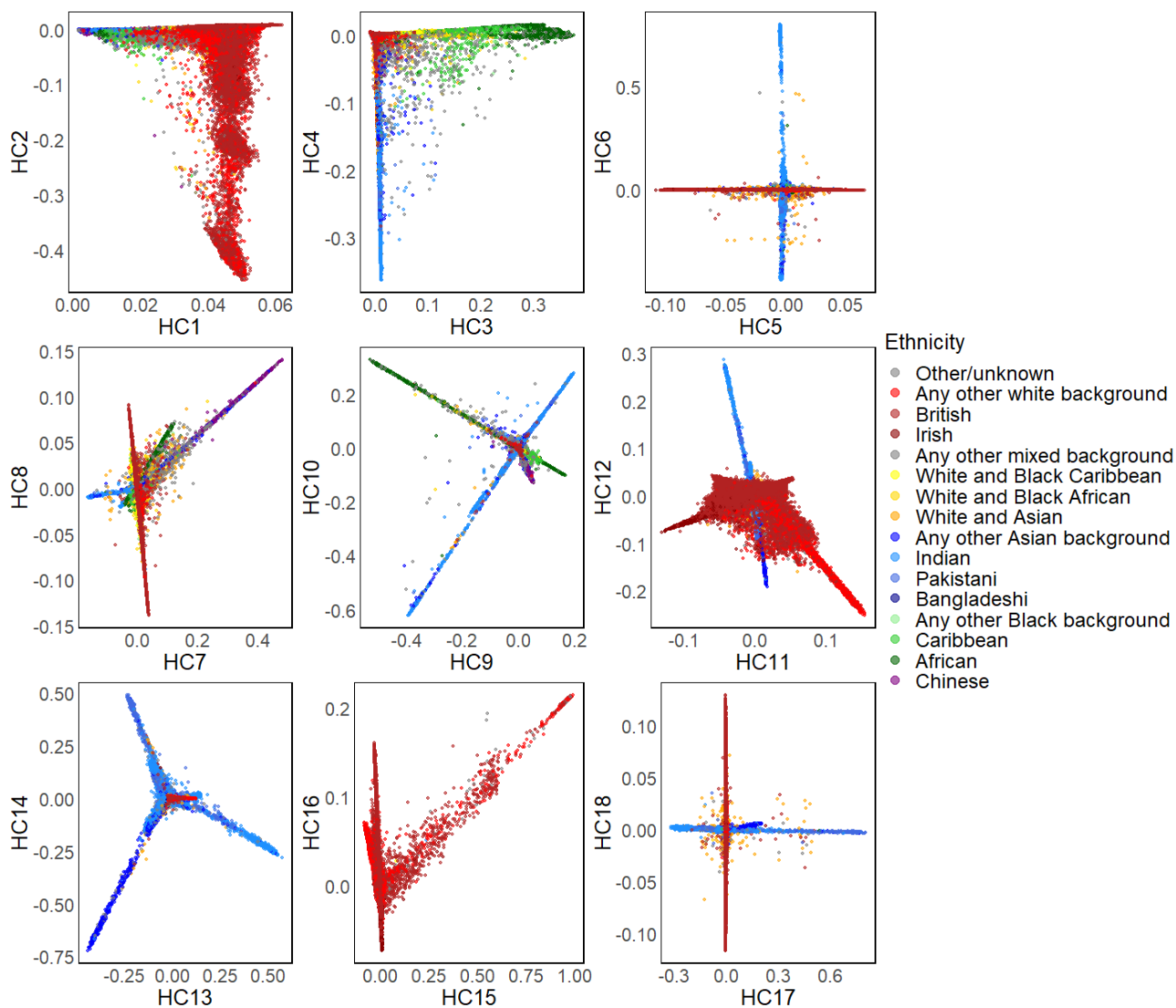
- 835 1. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
- 836 2. Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607–619 (2012).
- 837 3. Salter-Townshend, M. & Myers, S. Fine-scale inference of ancestry segments without prior knowl-  
838 edge of admixing groups. *Genetics* **212**, 869–889 (2019).
- 839 4. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
- 840 5. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry  
841 of African Americans, Latinos, and European Americans across the United States. *The American*  
842 *Journal of Human Genetics* **96**, 37–53 (2015).
- 843 6. Leslie, S. *et al.* The fine-scale genetic structure of the british population. *Nature* **519**, 309–314  
844 (2015).
- 845 7. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed  
846 populations. *PLoS Genetics* **5**, e1000519 (2009).
- 847 8. Zhang, J. & Stram, D. O. The role of local ancestry adjustment in association studies using  
848 admixed populations. *Genetic Epidemiology* **38**, 502–515 (2014).
- 849 9. Gouveia, M. *et al.* Unappreciated subcontinental admixture in Europeans and European Americans  
850 and implications for genetic epidemiology studies. *Nature Communications* **14**, 6802 (2023).
- 851 10. Peter, B. M. Admixture, population structure, and F-statistics. *Genetics* **202**, 1485–1501 (2016).
- 852 11. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies,  
853 targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47**, D1005–D1012 (2019).
- 854 12. Barrie, W. *et al.* Elevated genetic risk for multiple sclerosis emerged in steppe pastoralist popula-  
855 tions. *Nature* **625**, 321–328 (2024).
- 856 13. Schick, U. M. *et al.* Genome-wide association study of platelet count identifies ancestry-specific  
857 loci in Hispanic/Latino Americans. *The American Journal of Human Genetics* **98**, 229–242 (2016).
- 858 14. Hodonsky, C. J. *et al.* Ancestry-specific associations identified in genome-wide combined-  
859 phenotype study of red blood cell traits emphasize benefits of diversity in genomics. *BMC Ge-*  
860 *nomics* **21**, 1–14 (2020).
- 861 15. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in  
862 GWAS and to boost power. *Nature Genetics* **53**, 195–204 (2021).

- 863 16. Wu, J., Liu, Y. & Zhao, Y. Systematic review on local ancestor inference from a mathematical and  
864 algorithmic perspective. *Frontiers in Genetics* **12**, 639877 (2021).
- 865 17. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots  
866 using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
- 867 18. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using  
868 dense haplotype data. *PLoS Genetics* **8**, e1002453 (2012).
- 869 19. Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*  
870 **28**, 1359–1367 (2012).
- 871 20. Browning, S. R., Waples, R. K. & Browning, B. L. Fast, accurate local ancestry inference with  
872 FLARE. *The American Journal of Human Genetics* **110**, 326–335 (2023).
- 873 21. Brisbin, A. *et al.* PCAdmix: principal components-based assignment of ancestry along each chro-  
874 mosome in individuals with admixed ancestry from two or more populations. *Human Biology* **84**,  
875 343 (2012).
- 876 22. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling  
877 approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*  
878 **93**, 278–288 (2013).
- 879 23. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler trans-  
880 form (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
- 881 24. Sanauallah, A., Zhi, D. & Zhang, S. d-PBWT: dynamic positional Burrows–Wheeler transform.  
882 *Bioinformatics* **37**, 2390–2397 (2021).
- 883 25. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. *Introduction to algorithms* (MIT press,  
884 2022).
- 885 26. Consortium, . G. P. *et al.* A global reference for human genetic variation. *Nature* **526**, 68 (2015).
- 886 27. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus  
887 genotype data. *Genetics* **155**, 945–959 (2000).
- 888 28. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genetics* **2**,  
889 e190 (2006).
- 890 29. McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genetics* **5**,  
891 e1000686 (2009).
- 892 30. Hu, S. *et al.* Leveraging fine-scale population structure reveals conservation in genetic effect sizes  
893 between human populations across a range of human phenotypes. *bioRxiv* 2023–08 (2023).
- 894 31. Haller, B. C. & Messer, P. W. SLiM 3: forward genetic simulations beyond the Wright–Fisher  
895 model. *Molecular Biology and Evolution* **36**, 632–637 (2019).
- 896 32. Sarmanova, A., Morris, T. T. & Lawson, D. J. Population stratification in GWAS meta-analysis  
897 should be standardized to the best available reference datasets. *bioRxiv* (2020).
- 898 33. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature*  
899 **562**, 203–209 (2018).
- 900 34. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-

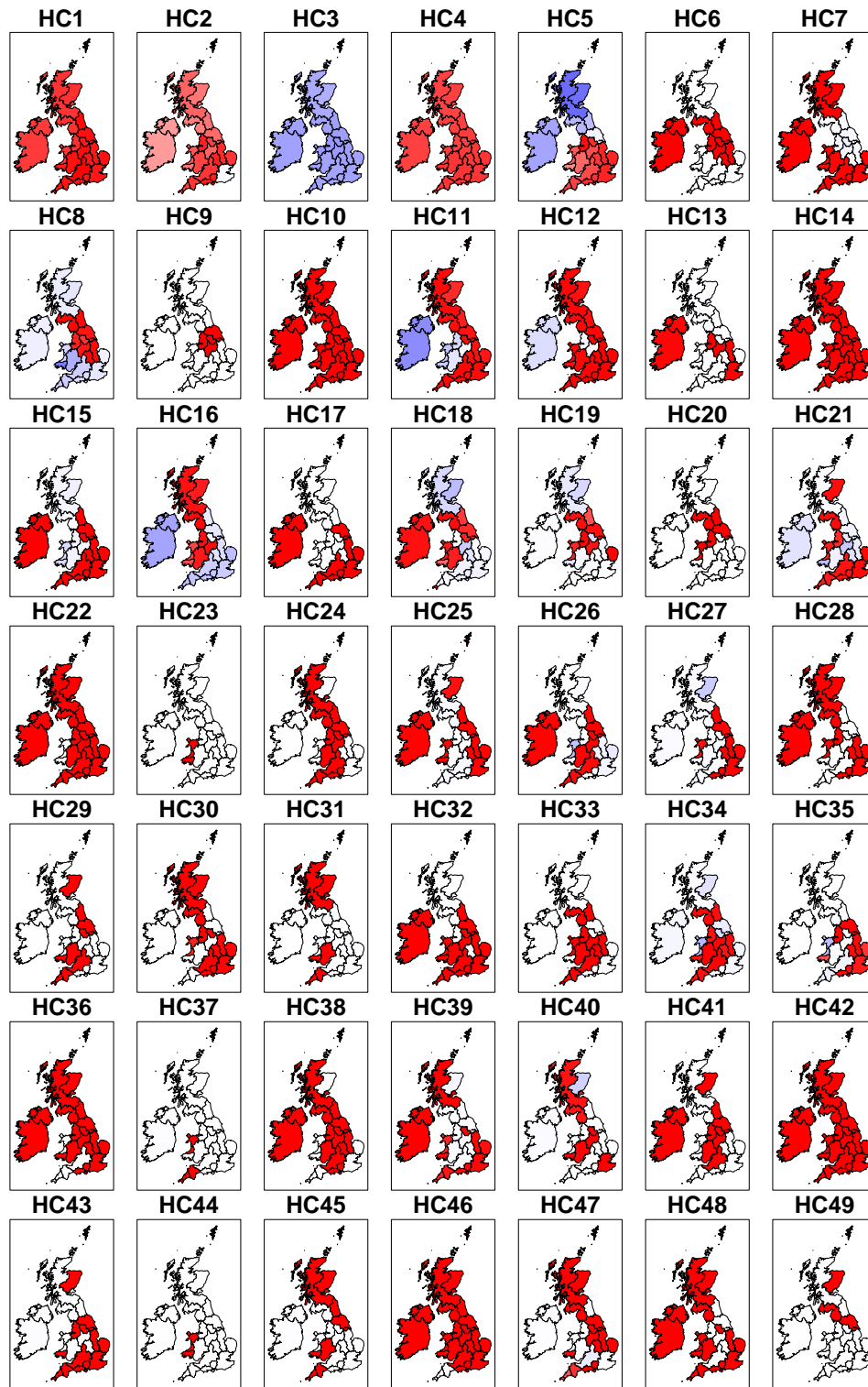


- 901 based scans for positive selection. *Molecular Biology and Evolution* **31**, 2824–2827 (2014).
- 902 35. Wysocki, T., Olesińska, M. & Paradowska-Gorycka, A. Current understanding of an emerging  
903 role of HLA-DRB1 gene in rheumatoid arthritis—from research to clinical practice. *Cells* **9**, 1127  
904 (2020).
- 905 36. Kitaura, K., Shini, T., Matsutani, T. & Suzuki, R. A new high-throughput sequencing method for  
906 determining diversity and similarity of T cell receptor (TCR)  $\alpha$  and  $\beta$  repertoires and identifying  
907 potential new invariant TCR  $\alpha$  chains. *BMC Immunology* **17**, 1–16 (2016).
- 908 37. Howard-McCombe, J. *et al.* Genetic swamping of the critically endangered Scottish wildcat was  
909 recent and accelerated by disease. *Current Biology* **33**, 4761–4769 (2023).
- 910 38. Benton, M. L. *et al.* The influence of evolutionary history on human health and disease. *Nature*  
911 *Reviews Genetics* **22**, 269–283 (2021).
- 912 39. Holmes, E. C. & Twiddy, S. S. The origin, emergence and evolutionary genetics of dengue virus.  
913 *Infection, Genetics and Evolution* **3**, 19–28 (2003).
- 914 40. Messina, J. P. *et al.* Global spread of dengue virus types: mapping the 70 year history. *Trends in*  
915 *Microbiology* **22**, 138–146 (2014).
- 916 41. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale  
917 sequence data. *The American Journal of Human Genetics* **108**, 1880–1890 (2021).
- 918 42. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd*  
919 *acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).
- 920 43. Glémin, S. *et al.* Quantification of GC-biased gene conversion in the human genome. *Genome*  
921 *research* **25**, 1215–1228 (2015).
- 922 44. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature*  
923 **526**, 75–81 (2015).
- 924 45. Mathieson, I. & Terhorst, J. Direct detection of natural selection in Bronze Age Britain. *Genome*  
925 *Research* **32**, 2057–2067 (2022).
- 926 46. Racimo, F., Marnetto, D. & Huerta-Sánchez, E. Signatures of archaic adaptive introgression in  
927 present-day human populations. *Molecular Biology and Evolution* **34**, 296–317 (2017).

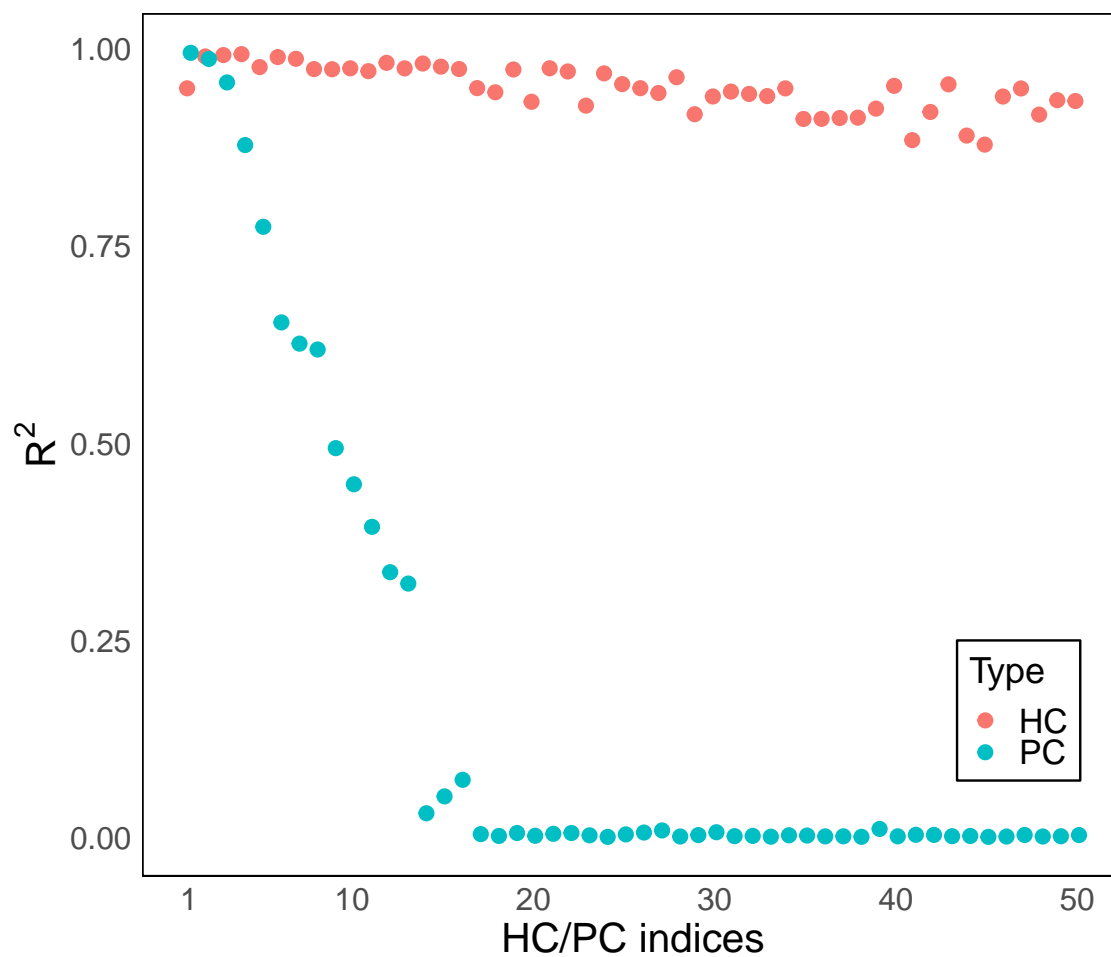
928 **Extended Data**



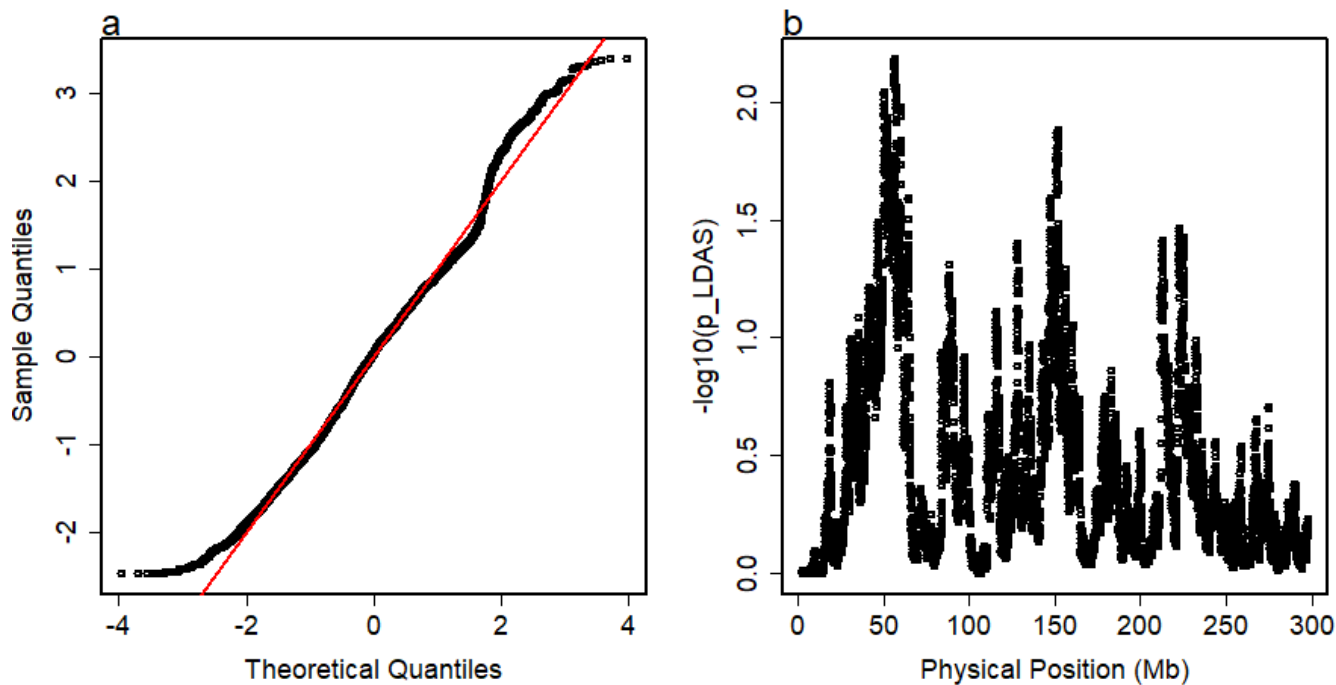
**Extended Data Fig. 1: Two-dimensional plots for the first 18 HCs stratified by UKB self-reported ethnic backgrounds (n=406,773 individuals).**



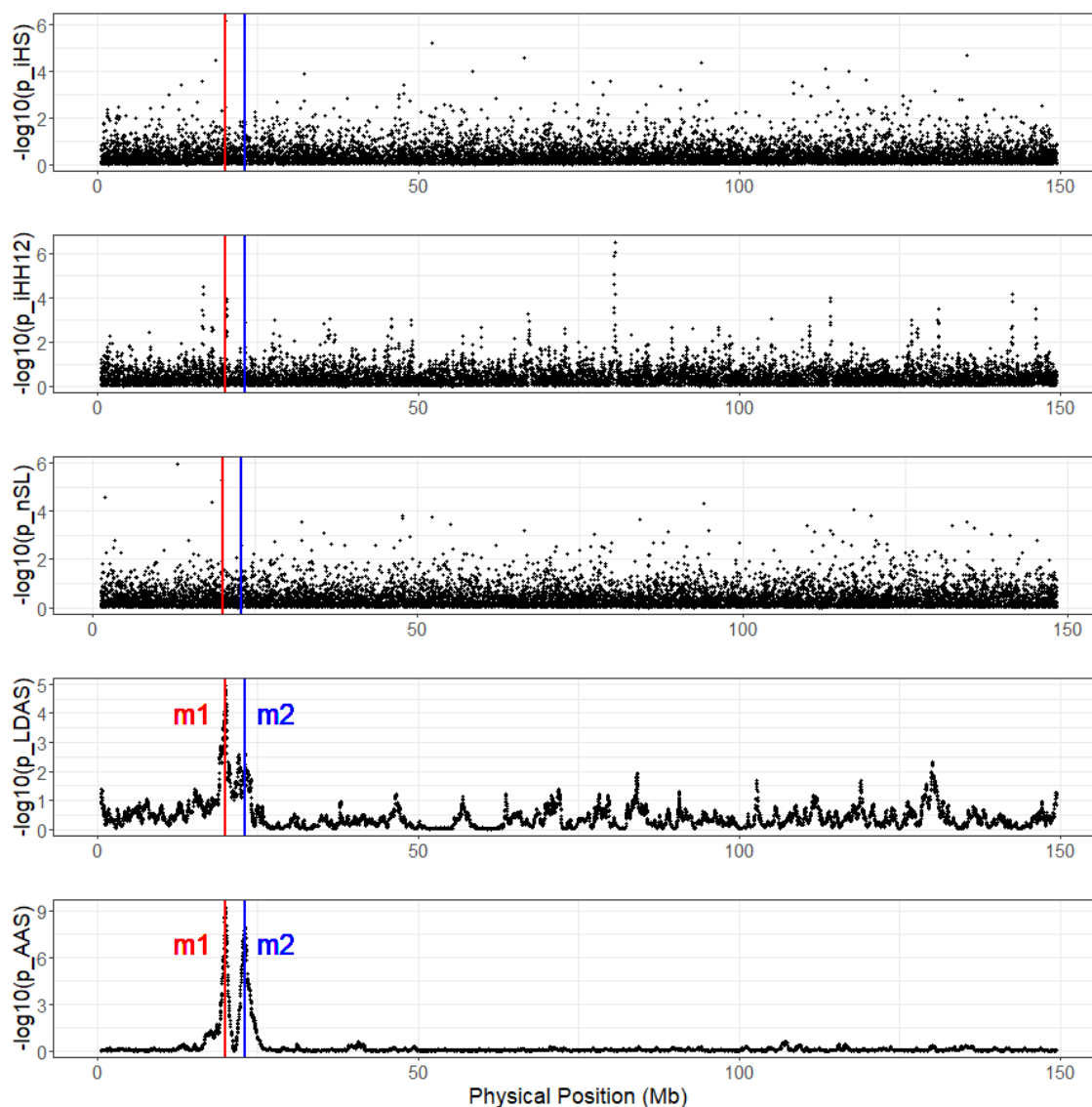
**Extended Data Fig. 2: Visualisation of the average of the first 49 HCs stratified by birthplaces within the UK and Ireland.** This analysis includes n=347,532 individuals. The average HC of each region bigger than, smaller than and equal to the worldwide average is coloured in red, blue and white, respectively.



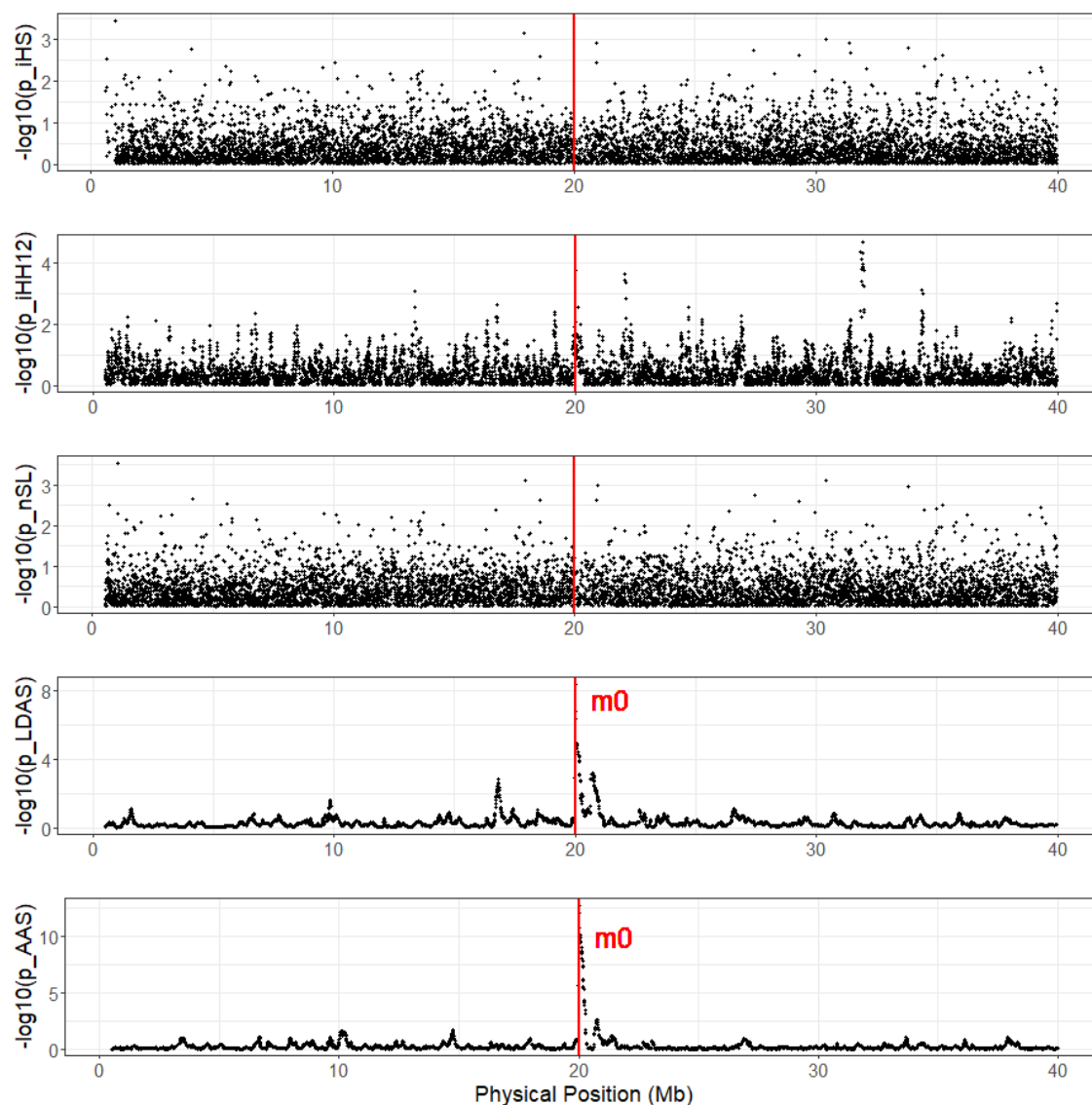
**Extended Data Fig. 3: Average Coefficient of determination for predicting top 50 HCs/PCs computed from odd (even) chromosomes using the first 150 HCs/PCs from even (odd) chromosomes of 406,773 individuals.** The top 50 HCs are well predicted from both plots, while only few top PCs can be predicted with high accuracy.



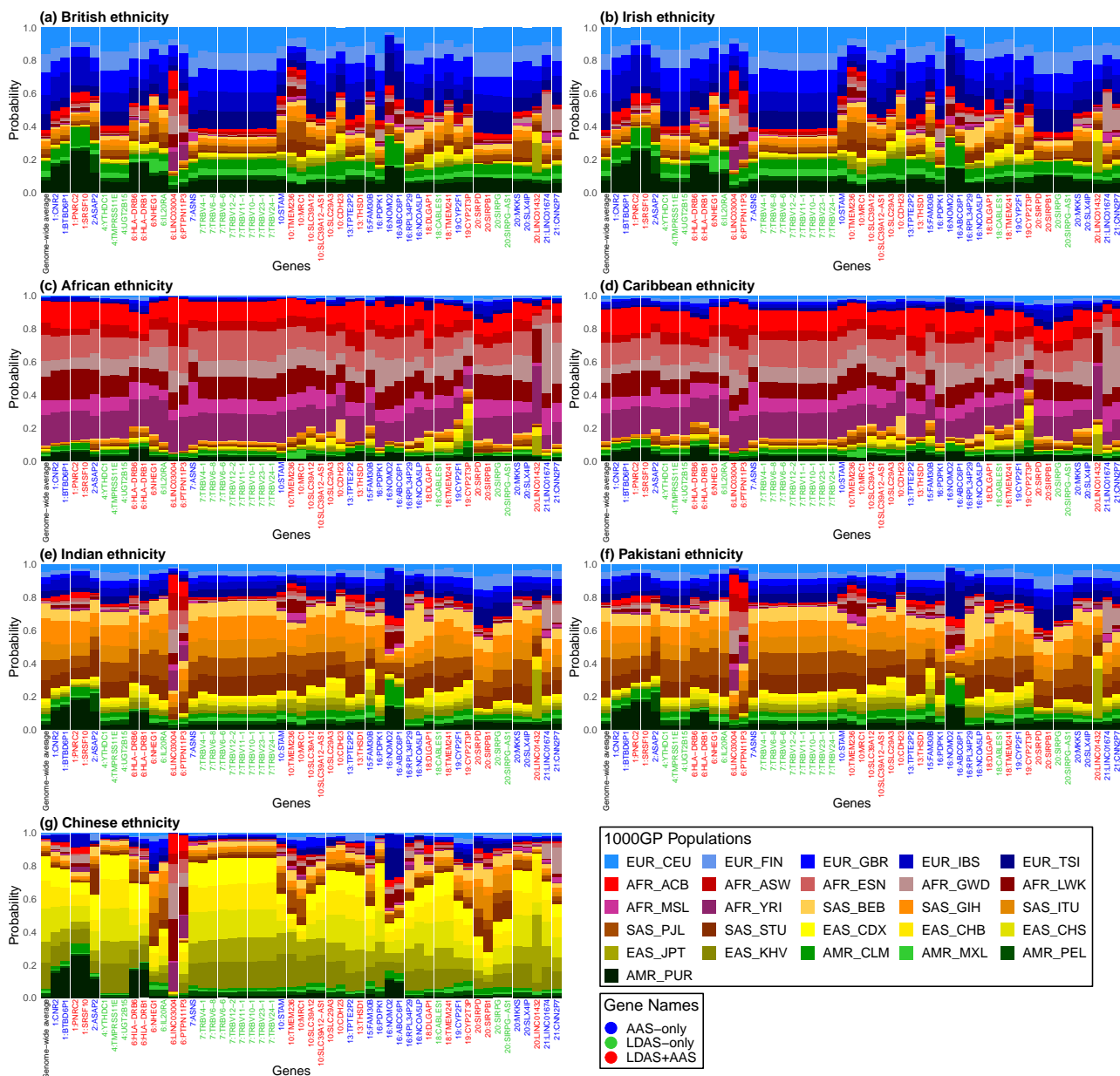
**Extended Data Fig. 4: Distribution of LDAS under the simulation of a 500Mb genome.** a, The quantile-quantile plot of the z-scores of LDAS. b, The P-values (represented in  $-\log_{10}$  scale) under the normality test for detecting low LDAS across the simulated genome.



**Extended Data Fig. 5: iHS, iHH12, nSL, LDAS and AAS under two-loci positive position selection in both ancient and modern populations (reporting the  $-\log_{10}$  of P-values). The red and blue vertical lines indicate the loci under selection in population p1 and p2, respectively.**



**Extended Data Fig. 6: iHS, iHH12, nSL, LDAS and AAS under one-locus positive selection in ancient populations and negative selection in modern population (reporting the  $-\log_{10}$  of P-values). The red vertical line indicates the loci under selection.**



**Extended Data Fig. 7: Average probabilities of 26 1000GP populations at genes with shared LDAS and AAS signals across 7 UK Biobank self-reported ethnicities.** We sampled a representative SNP from each gene with low LDAS or AAS signals (in 26-pop painting) shared between all 7 UKB ethnicities, as visualised in Fig. 6. The genome-wide average probabilities are shown on the left of each plot for comparison.