

Genetic risk for Multiple Sclerosis originated in Pastoralist Steppe populations

Authors

William Barrie¹§, Yaoling Yang^{2,3}§, Kathrine E. Attfield⁴§, Evan Irving-Pease⁵§, Gabriele Scorrano⁵§, Lise Torp Jensen⁶§, Angelos P. Armen⁴, Evangelos Antonios Dimopoulos⁷, Aaron Stern⁸, Alba Refoyo-Martinez⁵, Abigail Ramsøe⁵, Charleen Gaunitz⁵, Fabrice Demeter⁵, Marie Louise S. Jørkov⁹, Stig Bermann Møller¹⁰, Bente Springborg¹⁰, Lutz Klassen¹¹, Inger Marie Hyldgård¹¹, Niels Wickmann¹², Lasse Vinner⁵, Thorfinn Sand Korneliusen⁵, Martin Sikora⁵, Kristian Kristiansen^{5,13}, Santiago Rodriguez³, Rasmus Nielsen^{5,8}, Astrid K. N. Iversen⁴@, Daniel J. Lawson^{2,3}*@, Lars Fugger^{4,14}*@, Eske Willerslev^{1,5}*@

Affiliations

¹Department of Zoology, University of Cambridge, Cambridge, UK,

²Department of Statistical Sciences, School of Mathematics, University of Bristol, Bristol, UK,

³Integrative Epidemiology Unit, Population Health Sciences, University of Bristol, Bristol, UK,

⁴Oxford Centre for Neuroinflammation, Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, University of Oxford, Oxford, UK

⁵Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark,

⁶Department of Clinical Medicine, Aarhus University Hospital, Aarhus, Denmark

⁷Pathogen Genomics and Evolution Group, Department of Veterinary Medicine, University of Cambridge, Cambridge, UK

⁸Department of Integrative Biology, University of California, Berkeley

⁹Laboratory of Biological Anthropology, Department of Forensic Medicine, University of Copenhagen, Denmark

¹⁰Ålborg Historiske Museum, Nordjyske Museer, Vang Mark 25, 9380 Vestbjerg, Denmark

¹¹Museum Østdanmark - Djursland og Randers. Stemannsgade 2, DK-8900 Randers C, Denmark

¹²Museum Vestsjælland, Forten 10, 4300 Holbæk, Denmark

¹³Department of Historical Studies, University of Gothenburg, SE-41255, Gothenburg, Sweden

¹⁴MRC Human Immunology Unit, John Radcliffe Hospital, University of Oxford, Oxford, UK

* Corresponding authors; email: Dan.Lawson@bristol.ac.uk, lars.fugger@ndcn.ox.ac.uk, ew482@cam.ac.uk

§ Joint first authors

@ Joint last authors

SUMMARY

Multiple sclerosis (MS) is a modern neuro-inflammatory and -degenerative disease, which is most prevalent in Northern Europe. Whilst it is known that inherited risk to MS is located within or within close proximity to immune genes it is unknown when, where and how this genetic risk originated. By using the largest ancient genome dataset from the Stone Age, along with new Medieval and post-

44 Medieval genomes, we show that many of the genetic risk variants for MS rose to higher frequency
45 among pastoralists located on the Pontic Steppe, and were brought into Europe by the Yamnaya-
46 related migration approximately 5,000 years ago. We further show that these MS-associated
47 immunogenetic variants underwent positive selection both within the Steppe population, and later in
48 Europe, likely driven by pathogenic challenges coinciding with dietary and lifestyle environmental
49 changes. This study highlights the critical importance of this period as a determinant of modern
50 immune responses and its subsequent impact on the risk of developing MS in a changing
51 environment.

52

53 INTRODUCTION

54 Multiple sclerosis (MS) is an autoimmune disease of the brain and spinal cord that currently affects
55 more than 2.5 million people worldwide. The prevalence varies markedly with ethnicity and
56 geographical location, with the highest prevalence observed in Europe (142.81 per 100.000 people),
57 and Northern Europeans being particularly susceptible to developing the disease¹. The origins and
58 reasons for the geographical variation are poorly understood, yet such biases may hold important
59 clues as to why the prevalence of autoimmune diseases, including MS, has continued to rise during
60 the last 50 years.

61

62 While still elusive, MS etiology is thought to involve gene-gene and gene-environmental interactions.
63 Accumulating evidence suggests that exogenous triggers initiate a cascade of events involving a
64 multitude of cells and immune pathways in genetically vulnerable individuals, which may ultimately
65 lead to MS neuropathology².

66

67 Genome-wide association studies have identified 233 commonly occurring genetic variants that are
68 associated with MS; 32 variants are located in the HLA region and 201 outside the HLA region³. The
69 strongest MS associations are found in the HLA region with the most prominent of these, HLA-
70 DRB1*15:01, conferring an approximately three-fold increase in the risk of MS. Collectively, genetic
71 factors are estimated to explain approximately 30% of the overall disease risk, while environmental
72 and lifestyle factors are considered the major contributors to MS. Such determinants may include
73 geographically varying exposures like infections and low sun exposure/vitamin D deficiency. For
74 instance, while infection with Epstein-Barr virus frequently occurs in childhood and usually is
75 symptomless, delayed infection into early adulthood, as typically observed in countries with high
76 standards of hygiene, is associated with a 32-fold increased risk of MS^{4,5}. Lifestyle factors associated
77 with increased MS risk such as smoking, obesity during adolescence, and nutrition/gut health also
78 vary geographically⁶. Dysregulations including autoimmunity in modern immune systems could also
79 result from the absence of ancient immunological triggers to which humans have evolutionarily
80 adapted, for instance by disturbing the delicate balance of pro- and anti-inflammatory pathways⁷.

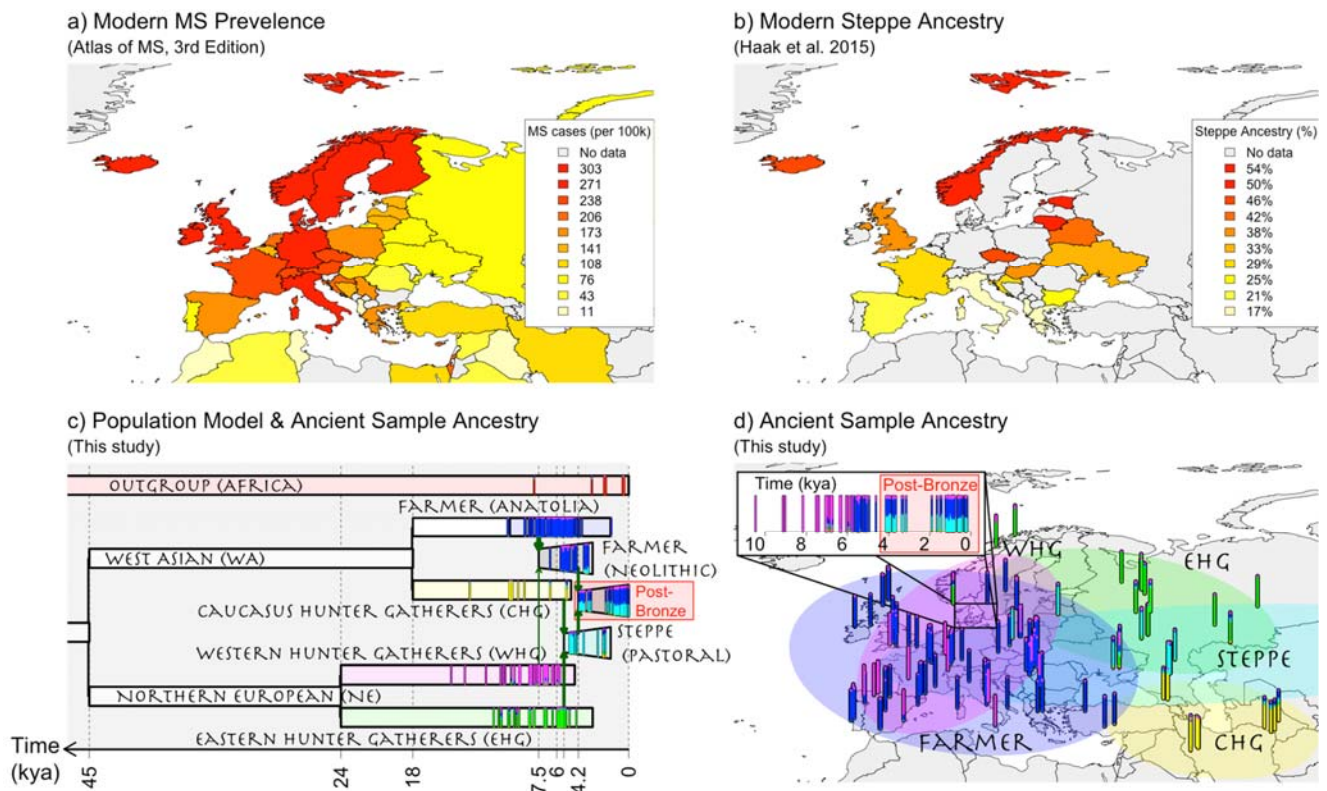
81

82 European ancestry has been postulated to explain part of the global difference in MS prevalence
83 globally in admixed populations⁸. Specifically, cases in African Americans exhibit increased
84 European ancestry in the HLA region compared to controls, with European haplotypes conferring
85 more MS risk for most HLA alleles, including HLA-DRB1*15:01. Conversely, Asian American cases
86 have decreased European ancestry in the HLA region compared to controls. Although Ancient
87 European ancestry and MS risk in Europe are known to be geographically structured (Figure 1a-b),
88 the effect of ancestry variation within Europe on MS prevalence is unknown.

89

90 Modern ancestry is viewed as a mixture of genetic ancestries derived from ancient populations, who
91 can be distinguished by their subsistence lifestyle: Western Hunter-Gatherers (WHG), Eastern

92 Hunter-Gatherers (EHG), Caucasus Hunter-Gatherers (CHG), Anatolian Farmers, and Steppe
93 Pastoralists (Figure 1c-d). Using the largest ancient genome dataset from the Stone Age, presented in
94 the accompanying study ‘Population Genomics of Stone Age Eurasia’⁹, coupled with new Medieval
95 and post-Medieval genomes, we quantified modern European ancestry with respect to these ancient
96 ancestries to identify signals of lifestyle-specific evolution. Then we determined whether the variants
97 associated with an increased risk for MS have undergone positive selection. We asked when selection
98 occurred and whether the targets of selection were specific to diet and lifestyle. Finally, we examined
99 the environmental conditions that may have caused selection for risk variants, including human
100 subsistence practice and exposure to pathogens.
101
102



103
104
105

106 **Figure 1: Population history of Europe is associated with modern-day distribution of MS.**

107 *a) Modern-day geographical distribution of MS in Europe. Prevalence data for MS (cases per*
108 *100,000) was obtained from ¹. b) Steppe ancestry in modern samples as estimated by ¹⁰. c-d) A model*
109 *of European prehistory¹¹ onto which our reference samples have been projected using NNLS (see*
110 *Methods), and the same data represented spatially. Chronologically, Western Hunter-Gatherers*
111 *(WHG) and Eastern Hunter-Gatherers (EHG) were largely replaced by Anatolian Farmers amid*
112 *demographic changes during the “Neolithic transition” around 9,000 years ago. Later migrations*
113 *during the Bronze Age about 5,000 years ago brought a roughly equal Steppe ancestry component*
114 *from the Pontic-Caspian Steppe to Europe, an ancestry descended from the EHG from the Middle*
115 *Don River region and Caucasus Hunter-Gatherers (CHG)⁹. Steppe ancestry has been associated with*
116 *the Yamnaya culture and then with the expansion westwards of the Corded Ware Complex and Bell*
117 *Beaker culture, and the eastwards expansion in the form of the Afanasievo culture^{12,10}. Samples are*
118 *vertical bars representing their “admixture estimate” estimated by NNLS (methods) from six*

119 *ancestries: EHG (green), WHG (pink), CHG (yellow), Farmer (blue), Steppe (cyan) or an Outgroup*
120 *(represented by ancient Africans, red). Important population expansions are shown as growing bars*
121 *and “recent” (post-Bronze age) non-reference admixed populations are shown for the Denmark time-*
122 *transect (see Supplementary Figure 1.1 for details).*

123

124 RESULTS

125 We obtained local ancestry (i.e. ancestry at specific loci) labels for ~410,000 self-identified “white
126 British” individuals in the UK Biobank¹³, using a reference panel of 318 ancient DNA (aDNA)
127 samples⁹ (Figure 1; Supplementary Figure 1.1) from the Mesolithic and Neolithic, including Steppe
128 pastoralists. Comparing the ancestry at each labelled single nucleotide polymorphism (SNP,
129 n=549,323) to genome-wide ancestry in the UK Biobank provided a “local ancestry anomaly score”
130 (Methods), for which two regions stood out as having undergone the most significant ancestry-
131 specific evolution in this period: LCT/MCM6, regulating lactase persistence¹⁴, and the HLA region
132 (Figure 2, top).

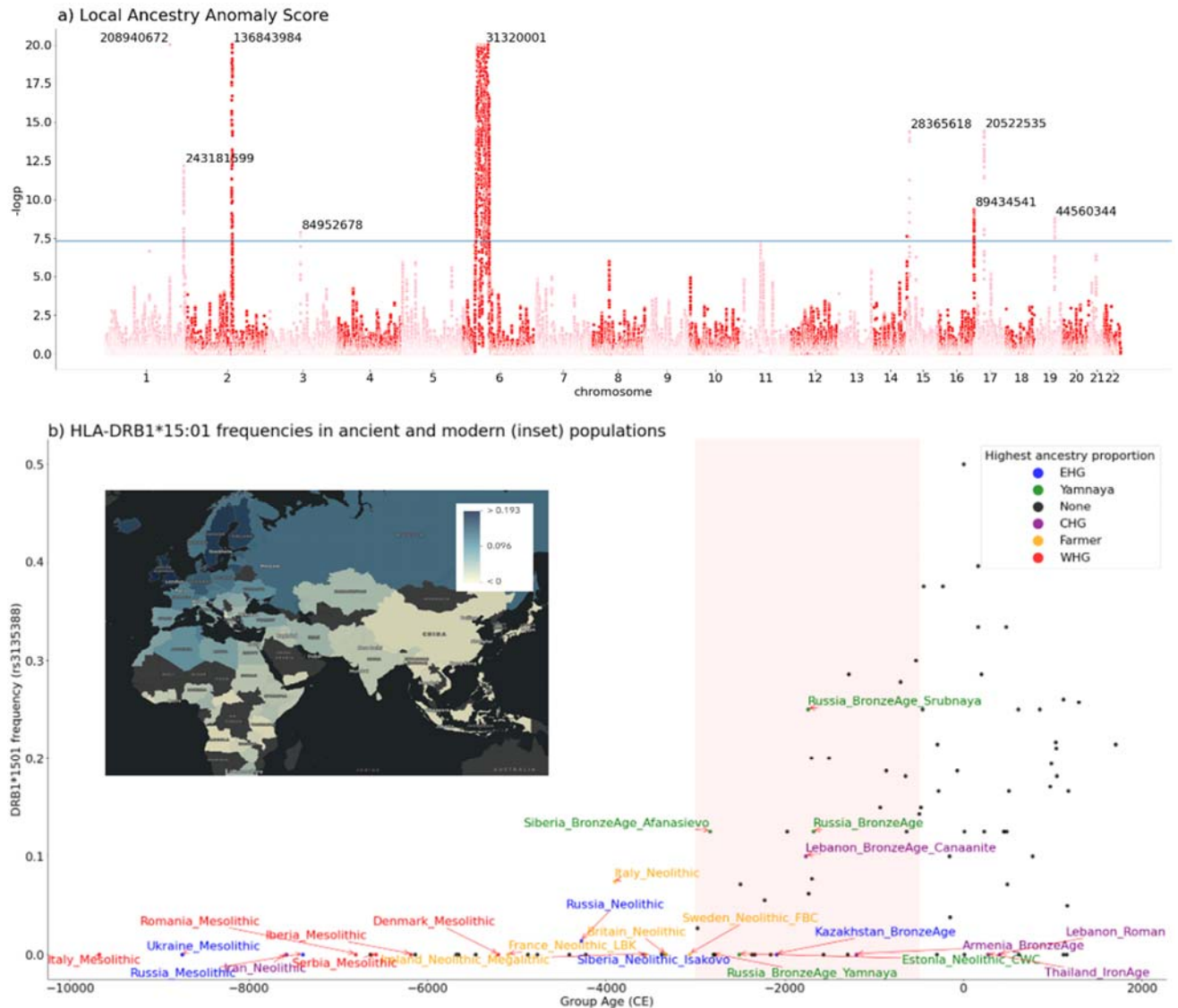
133

134 To determine whether this evolution of the HLA region has subsequently impacted diseases that are
135 strongly associated with risk alleles found within this region, we investigated the history of variants
136 associated with two HLA-associated autoimmune diseases, multiple sclerosis (MS) and rheumatoid
137 arthritis (RA), using the largest ancient genome dataset from the Stone Age coupled with 86 new
138 Medieval and post-Medieval genomes from Denmark (Supplementary Figure 1.1, Supplementary
139 Note 1, ST1). Alongside modern data, with our newly published genomes we have an almost
140 complete transect from approximately 10,000 years ago to the present.

141

142 The allele frequencies of SNPs conferring the highest risk for MS (all in the HLA class II region) in
143 our ancient groups show striking patterns. In particular the tag SNP (rs3135388-T) for HLA-
144 DRB1*15:01, the largest risk factor for MS, first appeared in an Italian Neolithic individual (sampleId
145 R3 from Grotta Continenza, C14 dated to between 5,836-5,723 BCE, coverage 4.05X) and rapidly
146 increased in frequency around the time of the emergence of the Yamnaya culture around 5,300 years
147 ago in Steppe and Steppe-derived populations (Figure 2). From risk allele frequencies of individuals
148 in the UK Biobank born in, and of a ‘typical ancestral background’ for, each country⁹, we found
149 HLA-DRB1*15:01 frequency peaks in modern populations of Finland, Sweden and Iceland, and in
150 ancient populations with high Steppe ancestry (Figure 2, inset).

151



152

153

Figure 2. Areas of unusual local ancestry in the genome, and ancient and modern frequencies of DRB1*15:01.

154

155

*a): Local Ancestry Anomaly Score measuring the difference between the local ancestry and the genome-wide average (capped at $-\log_{10}(p)=20$; see Methods). b) HLA-DRB1*15:01 frequencies in ancient and modern (inset) populations; this is the highest effect variant for MS risk (calculated using rs3135388 tag SNP). For the ancient data, for each ancestry (CHG, EHG, WHG, Farmer, Steppe) the five populations with the highest amount of that ancestry are coloured and labelled. DRB1*15:01 was present before the Steppe expansion, but rose to high frequency during the Yamnaya formation (shaded red). The geographical distribution of DRB1*15:01 frequency in modern populations in the UK Biobank is also shown (inset).*

156

157

158

159

160

161

162

163

164

To investigate the risk of a particular ancestry at all MS-associated fine-mapped loci³ present in the UK Biobank imputed dataset ($n=205/233$, see methods), we used the local ancestry dataset to calculate a risk ratio (see Methods: Weighted Average Prevalence) for each ancestry. For MS, Steppe ancestry has the highest risk ratio in nearly all HLA SNPs, while Farmer and ‘Outgroup’ ancestry (represented by ancient Africans) are often the most protective (Figure 3, top), meaning a Steppe-derived haplotype at these positions confers MS risk.

165

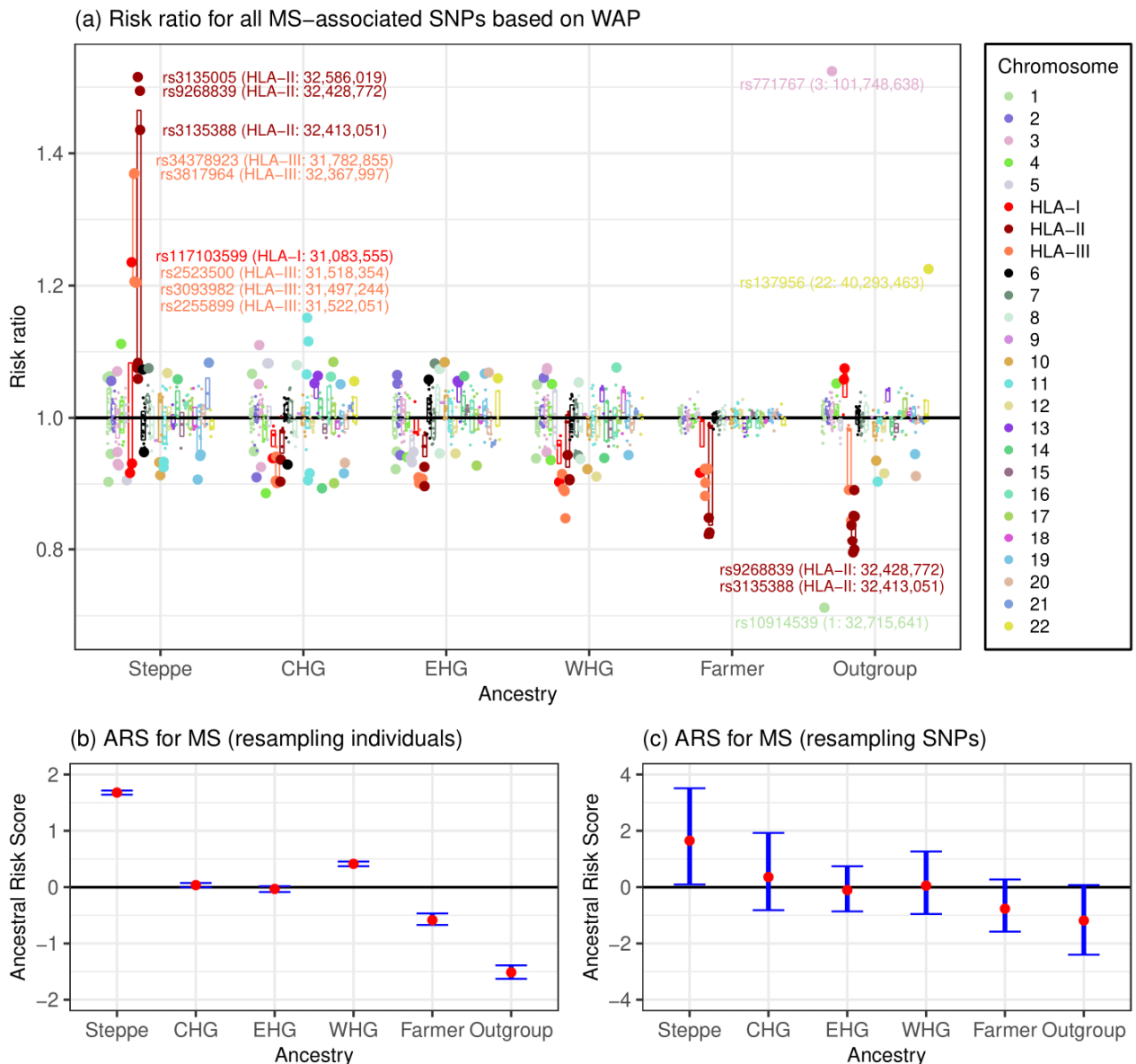
166

167

168

169

170



171

172

173 **Figure 3: Associations between local ancestry and MS in a modern population.**

174 *a) Risk ratio of SNPs for MS based on weighted average prevalence (WAP; see Methods), when*
 175 *decomposed by inferred ancestry. Each ancestry is assigned a mean and confidence interval based on*
 176 *bootstrap resampling, for each chromosome (faded where non-significant). The three HLA regions*
 177 *are split from the rest of chromosome 6, and SNPs with risk ratio >1.2 or <0.8 are annotated. b-c)*
 178 *Genome-wide Ancestral Risk Scores (ARS, see Methods) for MS. Confidence intervals are estimated*
 179 *by either bootstrapping over individuals (b, which can be interpreted as testing power to reject a null*
 180 *of no association between MS and ancestry) and bootstrapping over SNPs (c, which can be*
 181 *interpreted as testing whether ancestry is associated with MS genome-wide).*

182

183 Having shown that some ancestries carry higher risk, we calculated an aggregate risk for each
 184 ancestry across the same SNPs using a new statistic, the Ancestral Risk Score (ARS). ARS is
 185 computed in a large modern sample with local ancestry labels, estimating the relative risk for a

186 modern individual consisting of entirely one ancestry, mitigating the effects of low aDNA sample
187 numbers and bias¹⁵, and being robust to intervening drift and selection. We used effect size estimates
188 from previous association studies, under an additive model, with confidence intervals obtained via an
189 accelerated bootstrap¹⁶ (Supplementary Note 4). In the ARS for MS (Figure 3 bottom), Steppe
190 ancestry had a large and significant risk, followed by WHG, CHG and EHG; Neolithic Farmer and
191 Outgroup ancestry had the lowest ARS (Figure 3). Therefore Steppe ancestry is contributing the most
192 risk for MS overall. We tested for a genome-wide association by resampling loci, and found that
193 Steppe risk is much reduced but still clearly exceeds Farmer, a pattern which holds even when
194 excluding SNPs on the HLA (Supplementary Note 4, Figure S4.1).

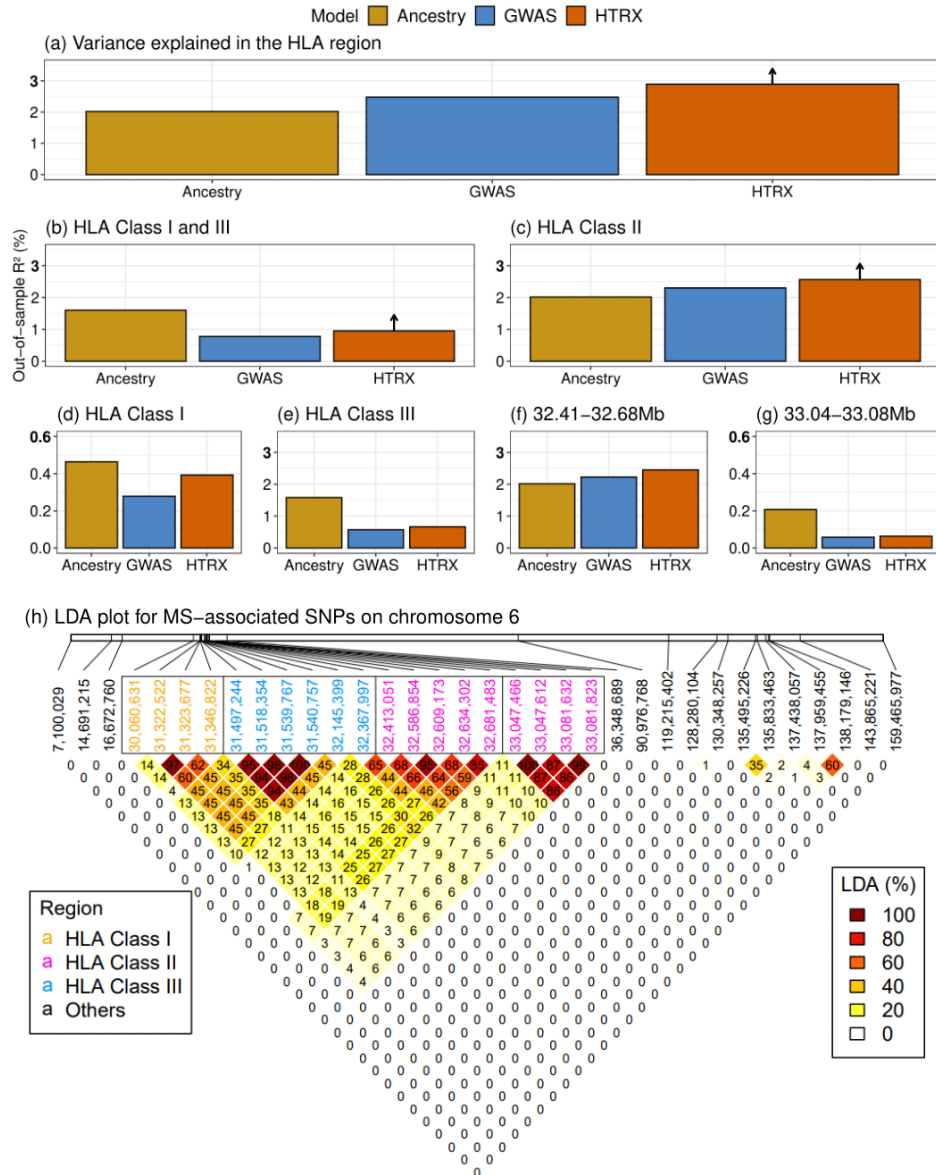
195

196 The fact that all but two MS-associated HLA SNPs confer risk within Steppe ancestry implies that
197 this risk has a common evolutionary history. We therefore investigated whether ancestry was
198 important for prediction using three types of association study in the UK Biobank for disease-
199 associated SNPs, controlling for age, sex and the first 18 PCs. The first of these is a regular SNP-
200 based association as conducted in GWAS. The second uses local ancestry probabilities instead of
201 genotype values (Supplementary Note 3). The third is based on Haplotype trend regression (HTR)
202 which is used to detect interactions between SNPs¹⁷ by treating each haplotype's probability as a
203 feature from which to predict a trait, instead of using SNPs as in a regular GWAS. We developed a
204 new method called Haplotype Trend Regression with eXtra flexibility (HTRX, Supplementary Note
205 5) that searches for haplotype patterns that include single SNPs and non-contiguous haplotypes. To
206 prevent overfitting, we reported out-of-sample variation explained, and showed by simulation (see
207 Supplementary Figure 4.4) that HTRX predicts the same variance as regular GWAS when interactions
208 are absent, but explains more variance when the interaction strength increases.

209

210 Although our cohort of self-identified “white British” individuals is relatively under-powered with
211 respect to MS (cases=1,949; controls=398,049; prevalence=0.487%), MS was associated with Steppe
212 and Farmer ancestry ($p < 1e-10$) in the HLA region (Supplementary Figure 4.1). In 3 out of 4 main LD
213 blocks within the HLA (class I, two subregions of class II determined by LD blocks at 32.41-32.68Mb
214 and 33.04-33.08Mb, and class III), local ancestry explains significantly more variation in total than
215 SNP variation (Figure 4; measured by average out-of-sample McFadden's R^2 for logistic regression,
216 see Methods). While increased ancestry performance over GWAS can be explained by tagging of
217 SNPs outside the region, increased HTRX performance over GWAS quantifies the total effect of a
218 haplotype, including rare SNPs and epistasis. Across the entire HLA region, haplotypes explain at
219 least 17% more out-of-sample variation than GWAS (2.90%, compared to 2.48%). Interaction signals
220 are also observed within class I, within class II, and between class I and III.

221



222

223 **Figure 4: MS association in the HLA.**

224 Comparison of variance explained in MS within the UK Biobank, for all fine-mapped HLA SNPs with
 225 an independent contribution³. The plots compare GWAS (treating SNPs as having independent effect),
 226 local ancestry at those SNPs, and HTRX (haplotypes) after accounting for covariates (Methods). a) is
 227 for fine-mapped MS-associated SNPs in the HLA. b) is HLA class I and -III, c) is HLA class II, d) is
 228 HLA class I, e) is HLA class III, f) and g) are subregions of HLA class II chosen from LD. HTRX has
 229 small “up-arrows” where these are lower bounds (Methods). h) Genetic correlations in the HLA
 230 region at our time-depth from Ancestry-based LD (LDA, see Methods) and Supplementary Figure 6.5
 231 for LD.

232

233 This interaction risk can be attributed to particular ancestries: for example, multiple haplotypes at the
 234 32.41-32.68Mb region are Steppe-associated and have high MS odds ratios. We further tested whether
 235 co-occurring ancestries at each loci were associated with MS (Methods; Supplementary Figure 4.2),
 236 but found no evidence that risk was associated with anything other than Steppe ancestry.

237

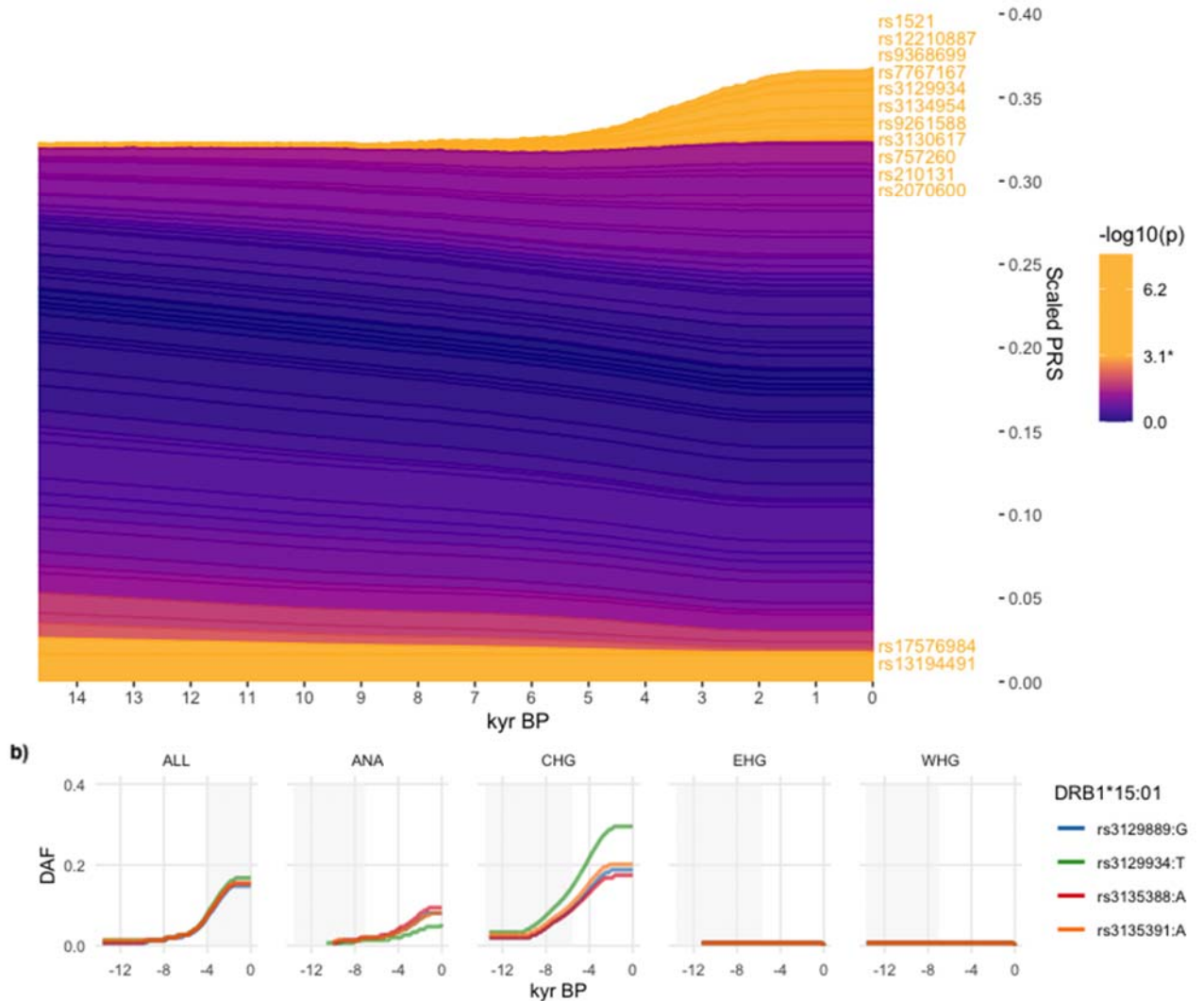
238 Having established that Steppe ancestry contributes most of the HLA-associated risk for MS, we
239 investigated evidence for polygenic selection on the disease-associated variants using two methods.
240 Firstly, we used a novel chromosome painting technique based on inference of a sample's nearest
241 neighbours in the marginal trees of an ARG that contains labelled individuals (Irving-Pease et al.,
242 *submitted*). The resulting ancestral path labels, for haplotypes in both ancient and modern individuals,
243 allowed us to infer allele frequency trajectories for risk associated variants, while controlling for
244 changes in admixture proportions through time. These paths extend backwards from the present day to
245 approximately 15,000 years ago, and are labelled with the unique population that a path travels
246 through. We stress that the path labels are not representative of a continuous population, but represent
247 a path backwards in time that encompasses that ancestry. For example, the CHG path originates in
248 Caucasus hunter-gatherers, before merging with EHG to form the Steppe population, and then merges
249 with other ancestries in later European populations (Figure 1).

250
251 Because not all fine-mapped SNPs had ancestral path labels (missing OR=10.4%) and due to the
252 difficulty in accurately inferring HLA alleles in ancient samples¹⁸, we LD-pruned genome-wide
253 significant summary statistics from the same study³ for which we did have ancestry path labels (n=62,
254 see methods). This allowed us to test for polygenic selection across disease-associated variants using
255 CLUES¹⁹ and PALM²⁰.

256
257 For MS, we found evidence that disease risk was selectively increased when considering all ancestries
258 collectively ($p=5.06e-05$; $\omega=0.0029$), between 5,000-2,000 years ago (Figure 5). Conditioning on each
259 of the four long-term ancestral paths (CHG, EHG, WHG and ANA), we found a statistically
260 significant signal of selection in CHG ($p=6.45e-3$; $\omega=0.009$). None of the other ancestral paths
261 reached nominal significance, although ANA ($p=0.0743$; $\omega=0.011$) and EHG ($p=0.064$; $\omega=0.0045$)
262 paths were close. Again, it is likely that the selection occurred in the pastoralist population of the
263 Steppe, as that population consists of approximately half CHG ancestry¹¹ (Figure 1). The SNP driving
264 the largest change in genetic risk over time was rs3129934, in both the pan-ancestry ($p=9.52e-06$;
265 $s=0.017$) and CHG ($p=0.019$; $s=0.008$) analyses, which tags the HLA-DRB1*15:01 haplotype²¹. We
266 also tested three other alleles that tag the HLA-DRB1*15:01 haplotype (rs3129889, rs3135388 and
267 rs3135391) for evidence of selection, and found that the ancestry stratified signal was consistently
268 strongest in CHG (Figure 5). None of the four tag SNPs were detected on either the EHG or WHG
269 backgrounds, indicating that the HLA-DRB1*15:01 haplotype likely originated in the basal
270 population ancestral to both ANA and CHG.

271

a) Multiple sclerosis ($r^2 < 0.05$; window 250 kb) (n = 62) | All ancestries | $\omega = 0.012$ | se = 0.0029 | z = 4.053 | p = 5.06e-05



272

273

Figure 5: Evidence for selection on MS-associated SNPs.

274

a) Stacked line plot of the pan-ancestry PALM analysis for MS, showing the contribution of alleles to disease risk over time. Individual SNPs are stacked, with their trajectories polarised to show the frequency of the positive risk allele and weighted by their scaled effect size: when a given SNP bar becomes wider over time the risk allele has increased in frequency, and vice versa. SNPs are sorted by their marginal p-value and direction of effect, with selected SNPs that increase risk plotted on top. SNPs are also coloured by their marginal p-values, and significant SNPs are shown in yellow. The y-axis shows the scaled polygenic risk score (PRS), which ranges from 0 to 1, representing the maximum possible additive genetic risk in a population.

275

276

277

278

279

280

281

282

b) Maximum likelihood trajectories for four SNPs tagging DRB1*15:01. The background is shaded for the approximate time period in which the ancestry existed as an actual population. None of the tagging alleles are present on the EHG or WHG ancestral paths.

283

284

285

286

Our second selection measure introduces a new statistic, Linkage Disequilibrium of Ancestry (LDA). LDA is the correlation between ancestries across SNPs, measuring whether recombination events between ancestries are high compared to recombination within ancestries. From this we constructed

287

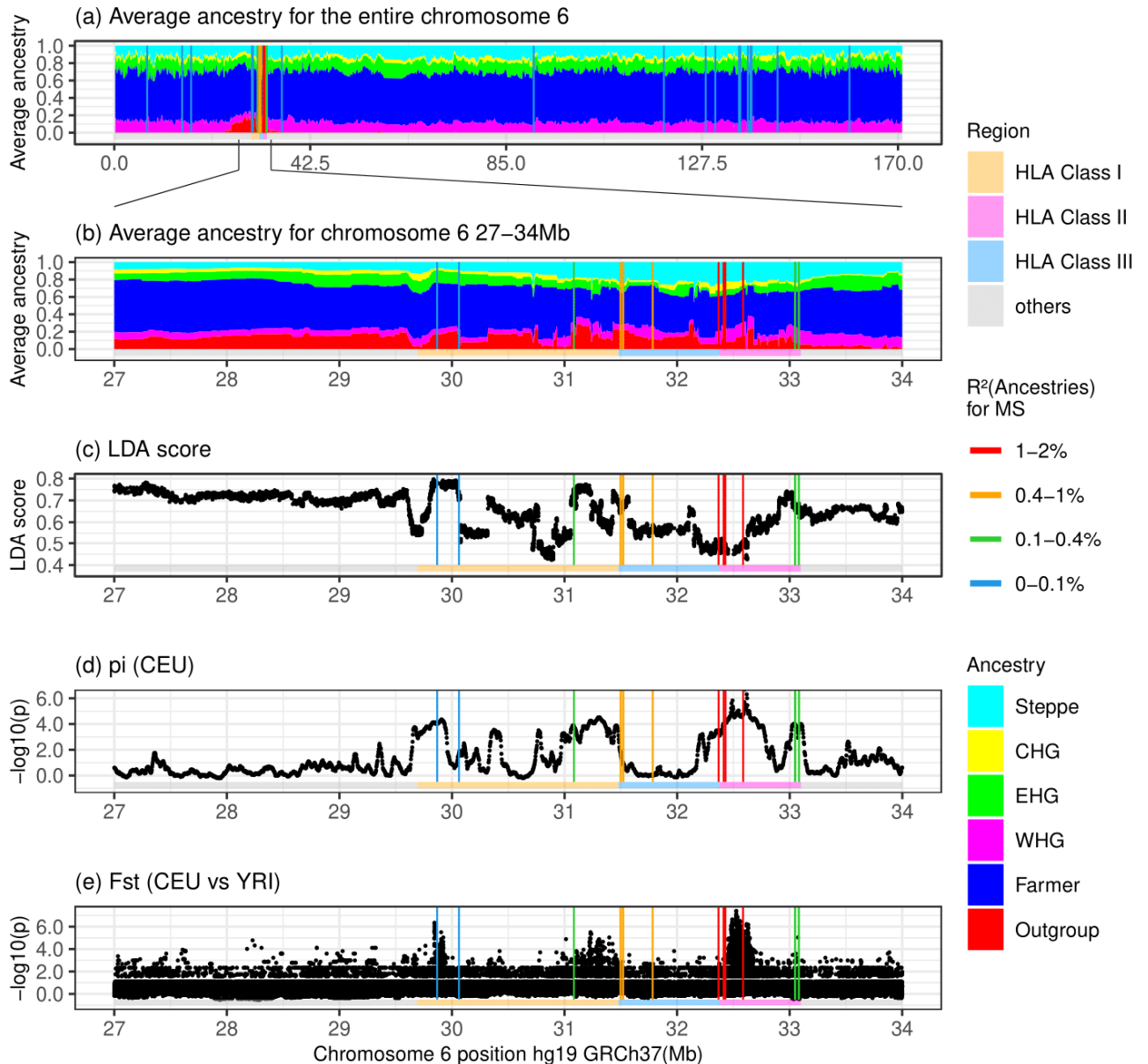
288

289 an “LDA score” using the fine-mapped SNPs, which is the total amount of genome in LDA with a
290 given SNP. A high LDA score indicates that the haplotype inherited from the reference population is
291 longer than expected, while a low score indicates that the haplotype is shorter than expected (i.e.
292 underwent more recombination). For example, the LCT/MCM6 region exhibits a high LDA score
293 (Supplementary Figure 6.4), as expected from a relatively recent selective sweep²².

294
295 The HLA has significantly *lower* LDA scores than the rest of chromosome 6 (Supplementary Figure
296 6.4). We simulated the LDA score under selection (Supplementary Figure 6.1; Methods), which
297 showed that when SNP frequencies are increasing in the most recent population, single locus selection
298 cannot explain this signal (Supplementary Figure 6.2-3). Instead, different loci in LD must have
299 independently reached high frequency in different ancestral populations that admixed, with selection
300 favouring haplotypes of mixed ancestry over single-ancestry haplotypes. Although multi-SNP
301 selection has been modelled²³, the interaction with prior population structure is less explored and is
302 important for the HLA, justifying a new term, “recombinant favouring selection”.

303
304 The HLA region contains the highest “Outgroup” ancestry anywhere on the genome (Figure 6),
305 reflecting high nucleotide diversity. Unlike other measures of balancing selection such as F_{st} (Figure
306 6), LDA describes excess ancestry LD from specific, dated populations and therefore need not be
307 correlated with them. For the HLA class II region, the selection measures all line up (LDA score, F_{st} ,
308 π), but for class I, the LDA score has an additional non-diverse minimum at 30.8Mb, implying that
309 here the genome is ancestrally diverse but genetically strongly constrained. The LDA score is thus
310 informative about the type of selection being detected, and whether it has been subject to change.

311



312

313

Figure 6: Signatures of selection at the HLA locus showing different regions of the HLA

314

(coloured bar) and locations of MS-associated SNPs (vertical lines, coloured by the variance

315

explained by 6 ancestries). a): Whole Chromosome 6 “local ancestry” decomposition by genetic

316

position. b). HLA “local ancestry” decomposition. c): LDA score; low values are indicative of

317

selection for multiple linked loci, while high values indicate positive selection. d): π scores

318

(nucleotide diversity) for CEU (Northern and Western European ancestry). MS-associated SNPs fall

319

in highly diverse regions of the HLA. e): F_{st} scores (divergence between two populations) for CEU vs

320

YRI(Yoruba); locally higher scores indicate regions that have undergone differential selection

321

between the two populations.

322

323

Because MS would not have conferred a fitness advantage on ancient individuals, it is likely that this

324

selection was driven by traits with shared genetic architecture, of which increased risk for MS in the

325

present is a consequence. We therefore looked at LD-pruned MS-associated SNPs that showed

326

statistically significant evidence for selection using CLUES (n=26) and which also had a genome-

327 wide significant trait association ($p < 5e-8$) in any of the 4,359 traits from the UK Biobank¹³ (UK
328 Biobank Neale Lab, Round 2: <http://www.nealelab.is/uk-biobank/>). We found that many selected
329 SNPs are also associated with celiac disease ($n=15$), white blood cell/neutrophil count ($n=15/n=15$),
330 hypothyroidism ($n=14$) and haemoglobin concentration ($n=14$) (Supplementary Figure 7.1). This
331 raised the possibility that the selection had increased risk for both MS and celiac disease, and when
332 we tested celiac disease for polygenic selection, we found significant evidence for positive selection,
333 increasing genetic risk ($p=9.65e-3$; $\omega=0.846$, Supplementary Note 6).

334
335 Because the UK Biobank is underpowered with respect to many traits and diseases, we also undertook
336 a manual literature search (see methods) for all SNPs that reached genome-wide significance for
337 association with MS in the summary stats (i.e., not LD-pruned, as independence is not required) and
338 which showed statistically significant evidence for selection using CLUES ($n=94$). We found that
339 most of the alleles under positive selection are associated with protective effects against specific
340 pathogens (virus, bacteria, fungi and parasites) and/or infectious diseases within one or several
341 ancestral paths (disease or pathogen associated/total selected in ancestry path: pan-ancestry 36/44;
342 ANA 24/31; CHG 25/29; EHG 27/35; WHG 9/10, Supplementary Note 8, ST13, Supplementary
343 Figure 8.1), although we note that GWAS data for many infectious diseases are not available. We
344 observed that the selected alleles had protective associations with several chronic viruses (EBV, VZV,
345 HSV, and CMV) and to viruses or diseases not associated with transmission in small hunter-gatherer
346 groups (e.g., measles, mumps, influenza, whooping cough). Moreover, many selected alleles
347 conferred a reduction of risk of parasites, of skin and subcutaneous tissue, gastrointestinal, respiratory,
348 urinary tract, and sexually transmitted infections, or of pathogens associated with these or other
349 infections (e.g., malaria, toxoplasmosis, entamoeba histolytica, clostridium difficile, tuberculosis,
350 streptococcus pyrogenes, and chlamydia) (Supplementary Note 8, ST13, Supplementary Figure 8.1).

351
352 We contrasted these findings for MS with results for RA, a common inflammatory HLA class II-
353 associated disease that primarily affects the joints causing pain, swelling and stiffness²⁴, which shows
354 a strikingly different ancestry risk profile. HLA-DRB1*04:01 is the largest genetic risk factor for RA;
355 in the CLUES analysis, the tag SNP for this allele (rs660895) displayed evidence of continuous
356 negative selection until approximately 3,000 years ago ($p=4.63e-4$, Supplementary Figure 5.1). We
357 found that WHG and EHG ancestries often confer the most risk at SNPs associated with RA (Relative
358 Risk ratio of RA-associated SNPs based on WAP, see Methods); and these ancestries have
359 contributed the greatest risk for RA on aggregate, reflected in a higher ARS for these ancestries
360 (Supplementary Note 4), while Steppe and Outgroup ancestry have the lowest scores (Supplementary
361 Figure 3.1). These results were recapitulated in the local ancestry GWAS (Supplementary Note 3).

362
363 We found that RA-associated SNPs have undergone negative polygenic selection ($p=7.93e-3$,
364 Supplementary Figure 5.1) over the last approximately 15,000 years; when this is decomposed by
365 ancestry path, we found significant evidence for negative selection in both the CHG ($p=3.09e-5$) and
366 ANA ($p=1.20e-3$) ancestry paths. We found no evidence for negative selection in the EHG and WHG
367 paths, although both show a trend of increasing risk, and EHG nears significance ($p=0.0842$).

368
369 These results demonstrate that genetic risk for RA was higher in the distant past, in contrast to MS,
370 with RA-associated risk variants present at higher frequencies in European hunter-gatherer
371 populations before the arrival of agriculture. In order to understand what caused the high risk in
372 hunter-gatherer populations and subsequent negative selection, we again undertook a manual
373 literature search for pleiotropic effects of SNPs associated with RA. Because the number of SNPs that
374 reached genome-wide significance in the GWAS study and also showed statistically significant

375 evidence for directional selection was large, we only analysed LD-pruned SNPs (n=42). We found
376 that the majority of selected SNPs were associated with protection against distinct pathogens and/or
377 infectious diseases across all paths (disease or pathogen associated/total selected in ancestry path:
378 pan-ancestry 9/13; ANA 10/13; CHG 8/11; EHG 10/16; WHG 10/12). We found that selected RA-
379 risk alleles were often linked to the same pathogens or diseases as in the MS analysis, although the
380 number of protective associations to distinct pathogens were fewer (Supplementary Note 8, ST14,
381 Supplementary Figure 8.1).

382

383 DISCUSSION

384 The last 10,000 years have seen some of the most extreme global experiments in lifestyle with the
385 emergence of farming in some regions and a pastoral lifestyle in others. While 5,000 years ago farmer
386 ancestry predominated across Europe, a relatively diverged ancestry arrived with the Steppe
387 migrations around this time. We have shown that this ancestry contributes the most genetic risk for
388 MS today, and that these variants were the result of positive selection coinciding with the emergence
389 of a pastoralist lifestyle on the Pontic-Caspian Steppe, and continued selection in the subsequent
390 admixed post-Stone Age populations in Europe. This ultimately created a legacy of heterogeneity in
391 MS risk observed across Europe today. These results address the long-standing debate around the
392 north-south gradient in MS prevalence in Europe, and suggest that the Steppe ancestry gradient in
393 modern populations - specifically at the HLA region - across the continent causes this phenomenon in
394 combination with environmental factors. Furthermore, while epistasis between MS-associated variants
395 in the HLA region has been demonstrated before^{25, 26, 27, 28}, we have shown that accounting for this
396 explains 17% more variance than independent SNPs effects alone. Many of the haplotypes carrying
397 these risk alleles have ancestry-specific origins, which could be exploited for individual risk
398 prediction and may offer a pathway from ancestry associations into a mechanistic understanding of
399 MS risk. We have contrasted these findings with results for rheumatoid arthritis (RA), another HLA
400 class II associated chronic inflammatory disease, and found that the genetic risk for RA exhibits a
401 contrasting pattern: genetic risk was highest in Stone Age hunter-gatherer ancestry and decreased over
402 time.

403

404 Our interpretation of this history is that co-evolution between pathogens and their human hosts has
405 resulted in massive and divergent ancestry-specific selection on immune response genes according to
406 lifestyle and environment, driven by a range of pathogenic drivers, and “recombinant favouring
407 selection” after these populations merged. The Late Neolithic and Early Bronze Age was a time of
408 massively increased infectious diseases in human populations, due to increased population density as
409 well as contact with, and consumption of, domesticated animals. Many diseases trace their origins to
410 this period, such as tuberculosis (TB) caused by the intracellular bacteria *Mycobacterium tuberculosis*
411 or *Mycobacterium bovis*^{29, 30}, bubonic plague caused by *Yersinia pestis*^{31, 32, 33}, herpes simplex virus³⁴,
412 and chickenpox caused by varicella-zoster virus³⁵, and we have shown that many of the MS- and RA-
413 associated variants under selection confer resistance to a range of infectious diseases and pathogens
414 (Supplementary Note 8). For example, HLA-DRB1*15:01 is associated with protection against TB³⁶
415 and increased risk for lepromatous leprosy³⁷. However, we are underpowered to detect specific
416 associations beyond this hypothesis due to poor knowledge of the distribution and diversity of past
417 diseases, poor preservation of endogenous pathogens in the archaeological record, and a lack of well-
418 powered GWAS studies for many infectious diseases.

419

420 A pattern that repeatedly appears is that of lifestyle change driving changes in risk and phenotypic
421 outcomes. We have shown that in the past environmental changes driven by lifestyle innovation
422 inadvertently drove an increase in genetic risk for MS. Today, with increasing prevalence of MS cases

423 observed over the last five decades^{38, 39}), we again observe a striking correlation with changes in our
424 environment, including lifestyle choices and improved hygiene, which no longer favours this previous
425 genetic architecture. Instead, the fine balance of genetically-driven cells within the immune system,
426 which are needed to combat a broad repertoire of pathogens without harming self-tissue, has been met
427 with new challenges, including a potential absence of requirement. For example, while a population of
428 immune cells, T helper 1 (Th1), direct strong cellular immune responses against intracellular
429 pathogens, T helper 2 (Th2) cells mediate humoral immune responses against extracellular bacteria
430 and parasites and further have the capacity to guide the restoring of homeostasis, thus preventing
431 damage of the infected tissue via immune-regulatory cytokines. We have shown that the majority of
432 selected MS-associated SNPs are associated with protection against a wide range of pathogens,
433 consistent with strong but balanced Th1/Th2 immunity in the Bronze age, where a diversification of
434 pathogens likely took place. In contrast, although MS pathogenesis is complex and multicellular of
435 nature, CD4⁺ Th cells, in particular IFN- γ producing Th1 cells and IL-17-producing Th17 cells play a
436 key role in disease development². The skewed Th1/Th2 balance observed in MS may partly result
437 from the developed world's increased sanitation, which has led to drastically reduced burden of
438 parasites, which the immune system had evolved to efficiently combat⁴⁰. In the case of RA, the
439 exposure of Hunter Gatherer populations to the respiratory or gastrointestinal pathogens linked to
440 triggering RA⁴¹ was likely low. The new pathogenic challenges associated with agriculture, animal
441 domestication, pastoralism, and higher population densities might have substantially increased the risk
442 of developing RA in genetically predisposed individuals, resulting in negative selection. If true, this
443 would present a parallel between RA in the Bronze Age and MS today, in which lifestyle changes
444 have exposed previously favourable genetic variants as autoimmune disease risks.

445
446 More broadly, it is clear that this was a critical period in human history during which highly
447 genetically and culturally divergent populations evolved and eventually mixed. These separate
448 histories dictate the genetic risk and prevalence of several autoimmune diseases today. Surprisingly,
449 the emergence of the pastoralist Steppe lifestyle may have had an impact on immune response as great
450 as or greater than the emergence of farming during the Neolithic transition, commonly held to be the
451 greatest lifestyle change in human history.

452

453 DATA AVAILABILITY

454 All collapsed and paired-end sequence data for novel samples sequenced in this study will be made
455 publicly available on the European Nucleotide Archive, together with trimmed sequence alignment
456 map files, aligned using human build GRCh37. Previously published ancient genomic data used in
457 this study are detailed in ST15, and are all already publicly available.

458

459 CODE AVAILABILITY

460 The modified version of CLUES used in this study is available from [https://github.com/standard-](https://github.com/standard-aaron/clues)
461 [aaron/clues](https://github.com/standard-aaron/clues). The pipeline and conda environment necessary to replicate the analysis of allele
462 frequency trajectories and polygenic selection in Supplementary Note 6 are available on Github at
463 https://github.com/ekirving/ms_paper. The code to create Ancestry Anomaly scores based on
464 Chromosome painting is on Github at https://github.com/danjlawson/ms_paper. The codes to
465 compute LDA and LDA score are available on Github at
466 <https://github.com/YaolingYang/LDAandLDAscore>. The codes to implement HTRX and its
467 simulation are on Github at <https://github.com/YaolingYang/HTRX>. The codes to implement ARS
468 calculation are on Github at https://github.com/will-camb/ms_paper.

469

470 ACKNOWLEDGEMENTS

471 We extend our thanks to all the former and current staff at the Lundbeck Foundation GeoGenetics
472 Centre and the GeoGenetics Sequencing Core, and to colleagues across the many institutions detailed
473 below. We are particularly grateful to Maria Madrona, Lærke Hansen and Julie Bitz-Thorsen for
474 laboratory assistance; and to Julie Hansen, Sandra Mularczyk, Katja Thorø Michler, Emilie Neerup
475 Nielsen for their help with sampling, and to Line Olsen as project manager for the Lundbeck
476 Foundation GeoGenetics Centre project. We thank UK Biobank Ltd. for access to the UK Biobank
477 genomic resource. We are thankful to Illumina Inc. for collaboration. E.W. thanks St. John's College,
478 Cambridge, for providing a stimulating environment of discussion and learning.

479

480 AUTHOR CONTRIBUTIONS

481 W.B., Y.Y., K.E.A., E.K.I-P, G.S., and L.T.J. contributed equally to this work.

482 A.I., D.J.L., L.F., and E.W. led the study.

483 W.B., A.R-M., L.F., R.N., and E.W. conceptualised the study.

484 R.N., K.K., L.F., and E.W. acquired funding for research.

485 A.R., C.G., F.D., M.L.S.J., S.B.M., B.S., L.K., I.M.H., N.W., L.V., and T.S.K., were involved in
486 sample collection and processing

487 W.B., Y.Y., E.K.I-P, A.S., S.R., and D.J.L. were involved in developing and applying methodology.

488 W.B., Y.Y., E.K.I-P, G.S., A.A., A.R., E.A.D., M.S., S.R., A.I., and D.J.L. undertook formal analyses
489 of data.

490 W.B., Y.Y., K.E.A., E.K.I-P, and L.T.J., A.I., L.F., and E.W. drafted the main text (W.B. led this).

491 W.B., Y.Y., E.K.I-P, G.S., L.T.J., E.A.D., A.S., F.D., M.L.S.J., S.B.M., B.S., L.K., I.M.H., N.W.,

492 L.V., A.I., and D.J.L. drafted supplementary notes and materials.

493 W.B., Y.Y., K.E.A., E.K.I-P, L.T.J., A.A., K.K., R.N., A.I., D.J.L., L.F., and E.W. were involved in
494 reviewing drafts and editing.

495 All co-authors read, commented on, and agreed upon the submitted manuscript.

496

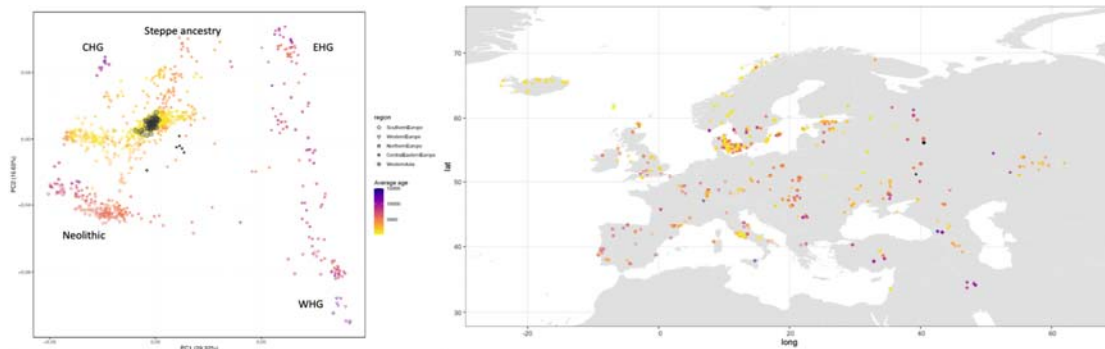
497

498 COMPETING INTERESTS

499 The authors declare no competing interests

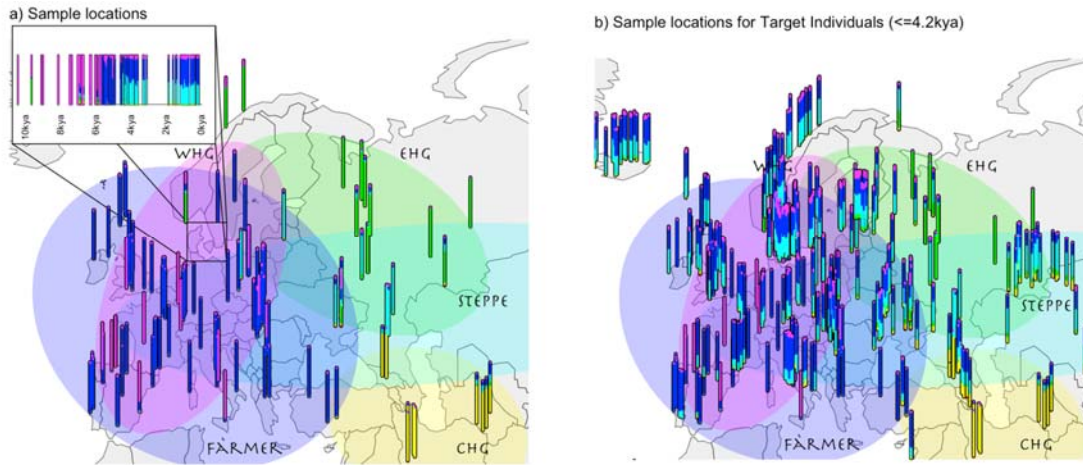
500

501 Supplementary Figures



502

503



504

505 **Supplementary Figure 1.1. Ancient sample PCA, map, ancestry proportions through time for**

506 **samples in Denmark.** (1) PC1 vs PC2 of the filtered Western Eurasian ancient samples included in

507 this study. Black circled points are Danish Medieval and post-Medieval samples published here for

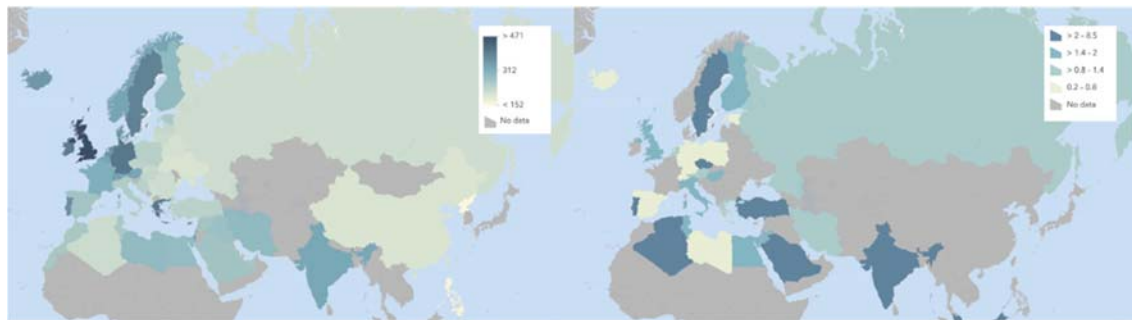
508 the first time. Major component ancestry locations are labelled. (2) Map of ancient filtered Western

509 Eurasian ancient samples included in this study (3a) Map of reference data and time transect of

510 Denmark as in Figure 1. (3b) More recent ancient data (samples <4,200 years ago) not used as

511 reference, showing the clines of the main ancestry components from (3a).

512



513

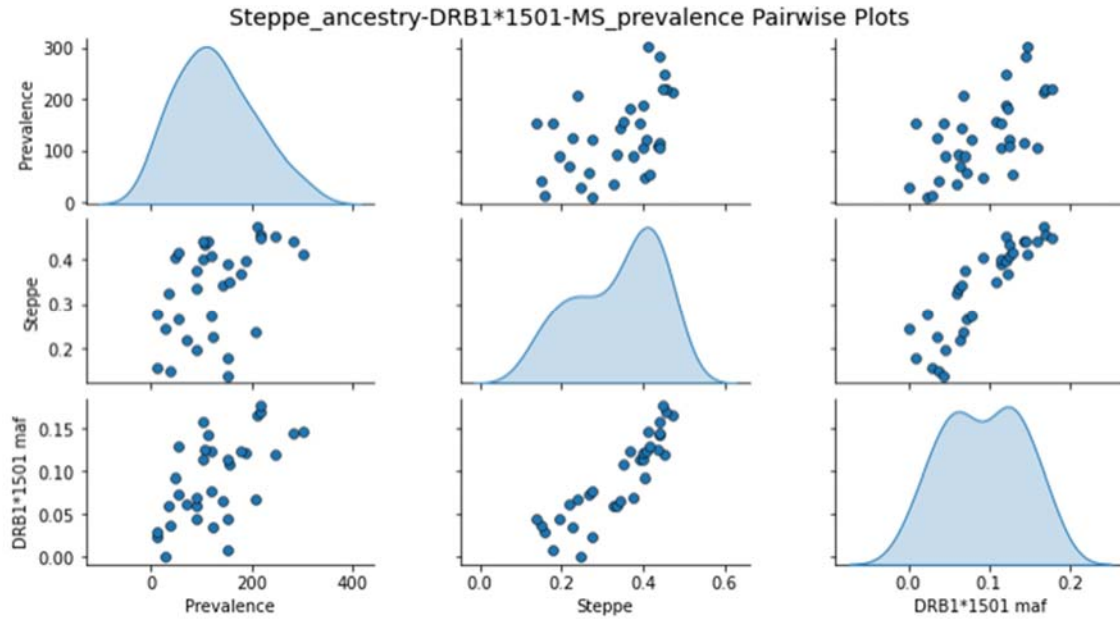
514 **Supplementary Figure 1.2. Modern prevalences of RA (left) and CD (right).**

515 Modern-day geographical distribution of RA and CD prevalence in Europe. Prevalence data for RA

516 (cases per 100,000) was obtained from ⁴². For CD, the seroprevalence (%) is based on the presence of

517 transglutaminase and/or endomysial autoantibodies; data were obtained from ⁴³.

518

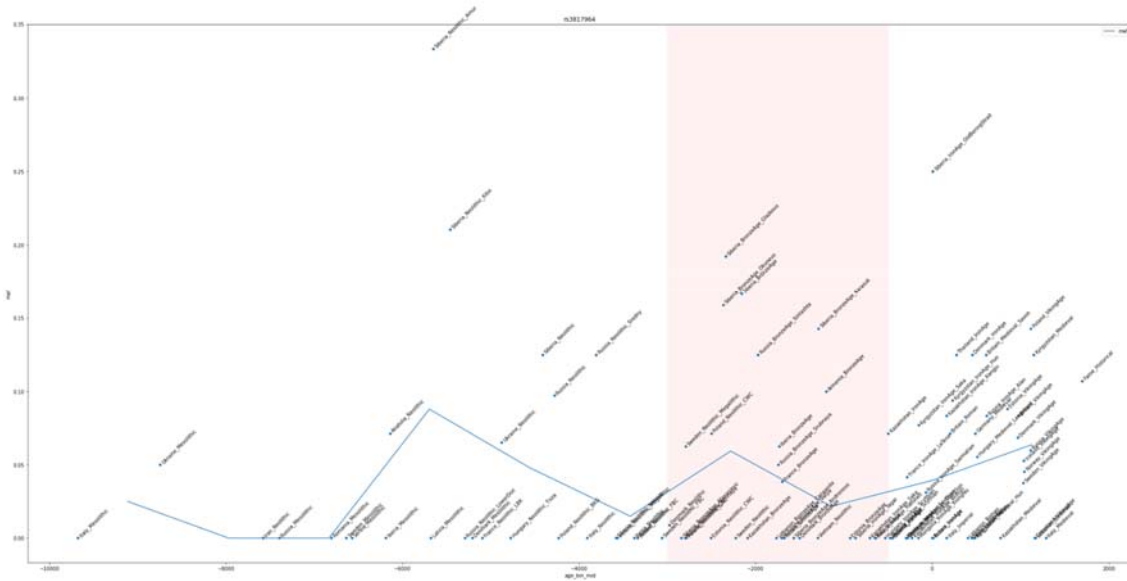


519

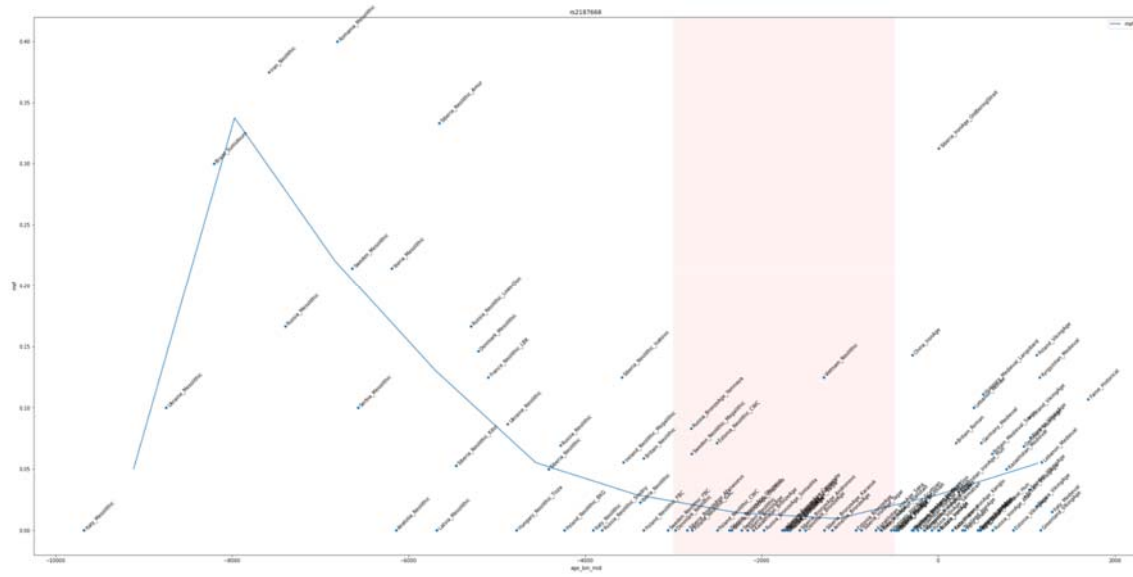
520

521

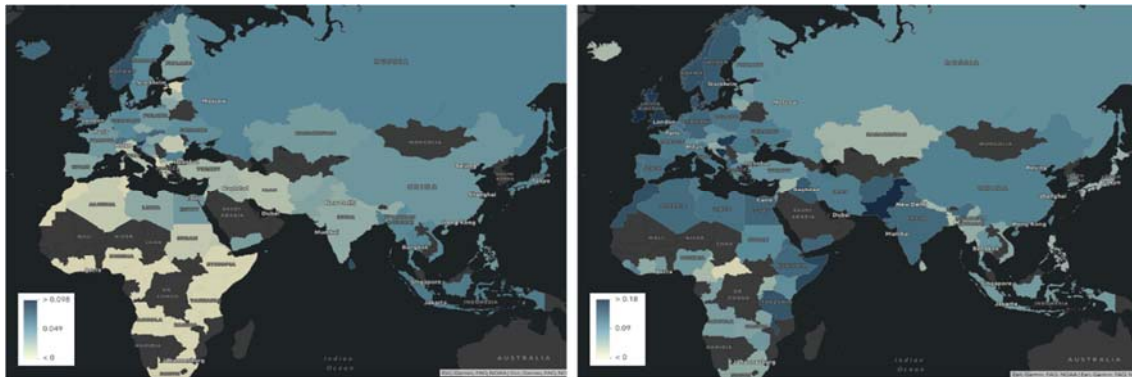
Supplementary Figure 1.3 Association between genome-wide Steppe ancestry, MS prevalence and DRB1*15:01 frequency in modern populations in the UK Biobank.



522



523



524

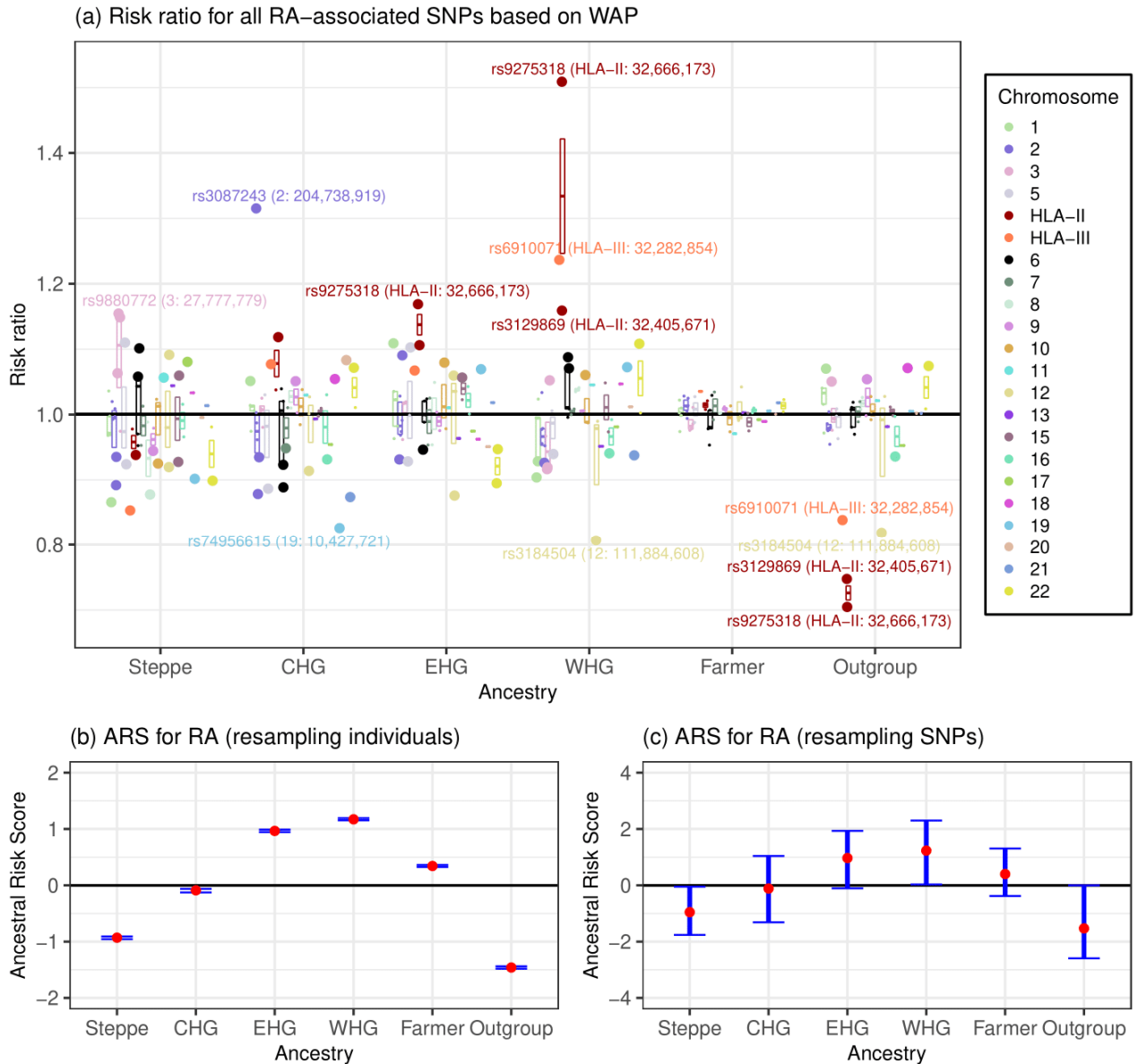
525 **Supplementary Figure 2.1. Ancient and modern prevalences of HLA-DRB1*04:01 (rs3817964)**
526 **and HLA-DQ2.5 (rs2187668).**

527 Top and middle: Ancient distributions of HLA-DRB1*04:01, the largest genetic risk factor in RA,
528 and HLA-DQ2.5, the largest genetic risk factor in CD. Average frequency across all populations is
529 shown (blue line, 10 time bins) as well as the Bronze Age (red shading).

530 Bottom: Modern distribution of HLA-DRB1*04:01 (left) and HLA-DQ2.5 (right) in populations in
531 the UK Biobank. NB the tag SNPs may be less effective at tagging these types in non-European
532 populations, so we urge caution in interpretation - especially in African populations.

533

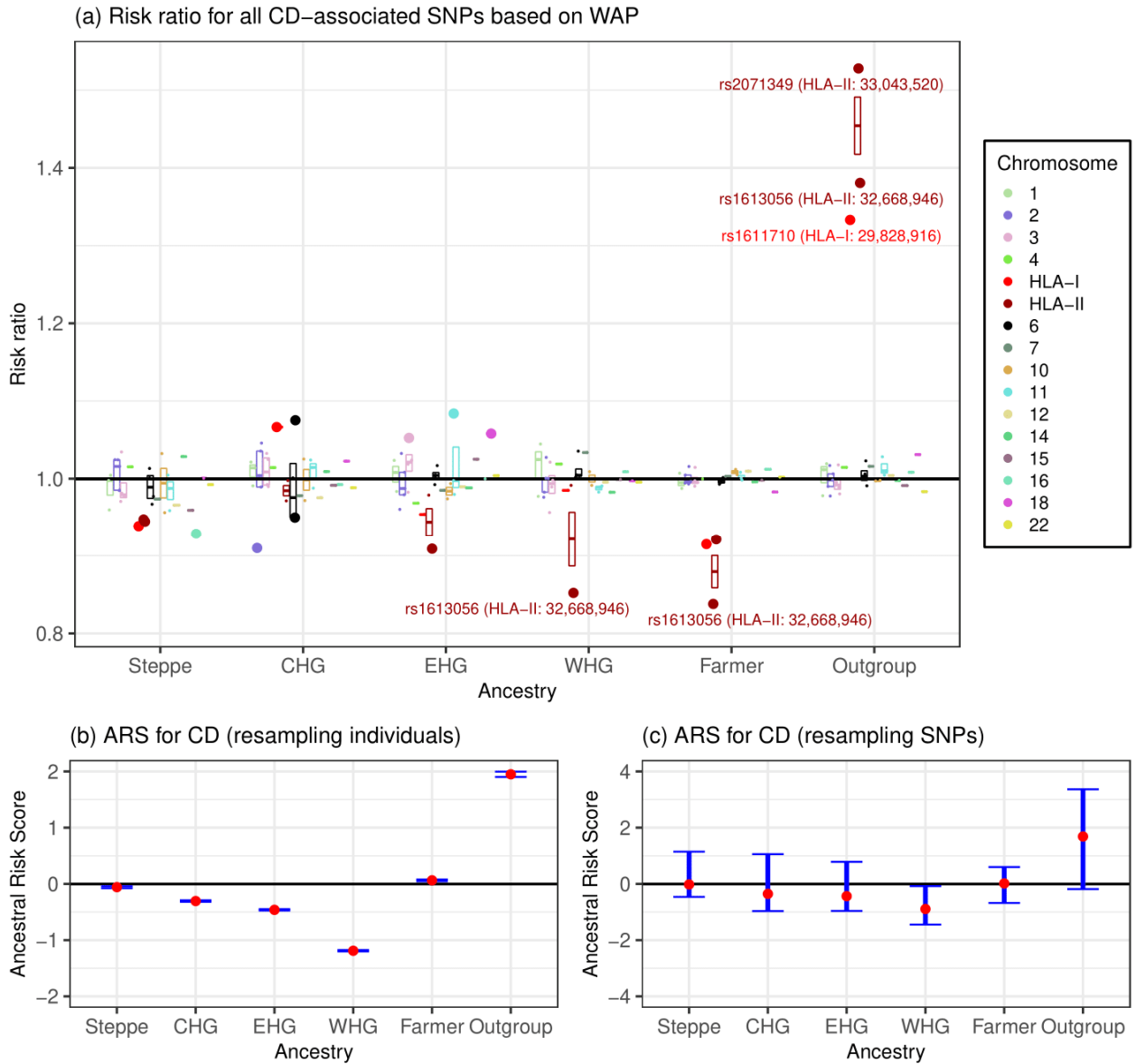
534



535
536
537
538
539
540
541
542
543
544
545

Supplementary Figure 3.1: Associations between local ancestry and RA in a modern population.

a) Risk ratio of SNPs for RA based on weighted average prevalence (WAP; see Methods), when decomposed by inferred ancestry. Each ancestry is assigned a mean and confidence interval based on bootstrap resampling, for each chromosome (faded where non-significant). SNPs with risk ratio >1.15 or <0.85 are annotated. b-c) Genome-wide Ancestral Risk Scores (ARS, see Methods) for RA. Confidence intervals are estimated by either bootstrapping over individuals (b, which can be interpreted as testing power to reject a null of no association between RA and ancestry) and bootstrapping over SNPs (c, which can be interpreted as testing whether ancestry is associated with RA genome-wide).



546

547

Supplementary Figure 3.2: Associations between local ancestry and CD in a modern population.

548

a) Risk ratio of SNPs for CD based on weighted average prevalence (WAP; see Methods), when decomposed by inferred ancestry. Each ancestry is assigned a mean and confidence interval based on bootstrap resampling, for each chromosome (faded where non-significant). SNPs with risk ratio >1.15 or <0.85 are annotated.

550

b-c) Genome-wide Ancestral Risk Scores (ARS, see Methods) for CD.

552

Confidence intervals are estimated by either bootstrapping over individuals (b, which can be

553

interpreted as testing power to reject a null of no association between CD and ancestry) and

554

bootstrapping over SNPs (c, which can be interpreted as testing whether ancestry is associated with

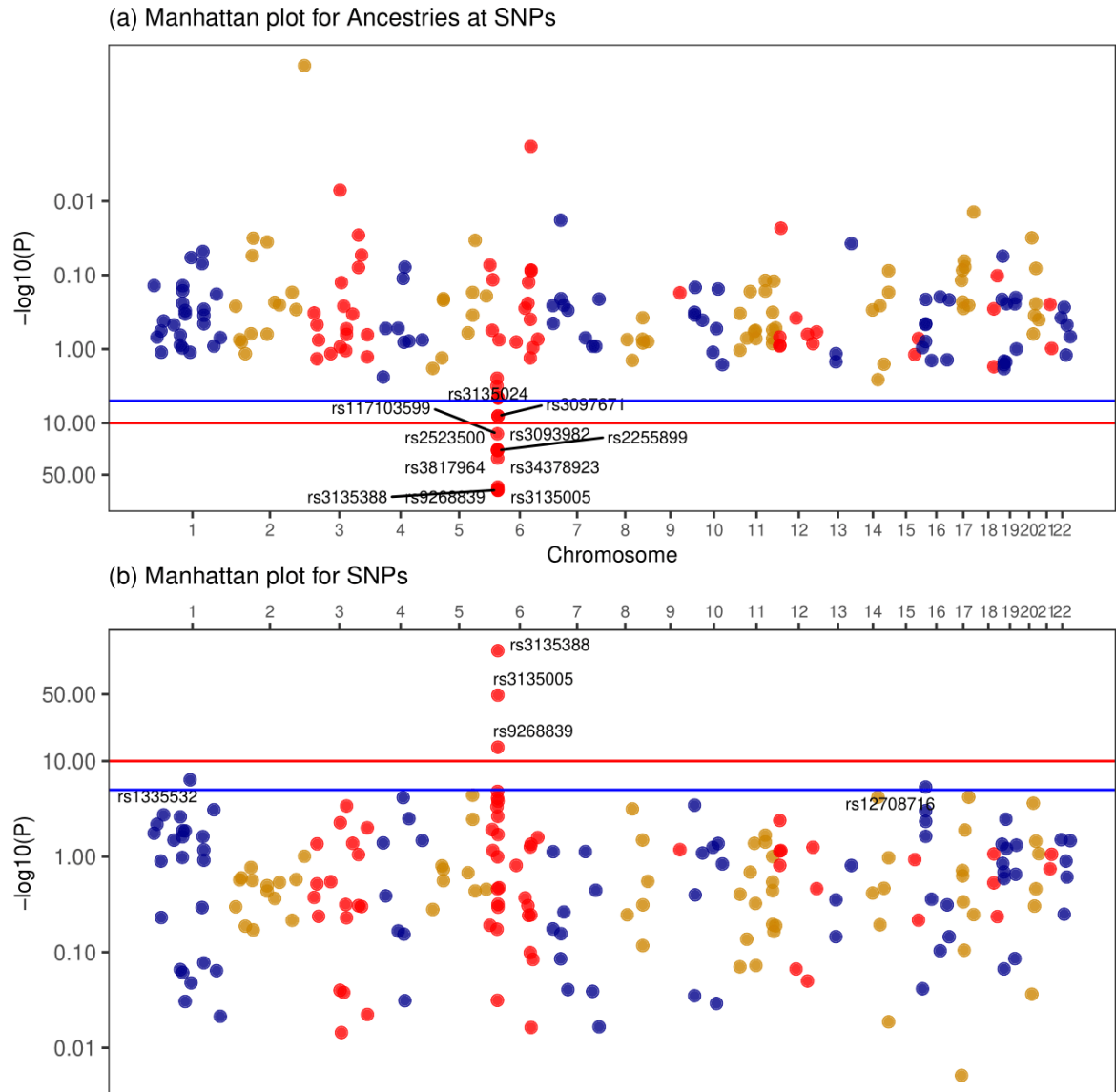
555

CD genome-wide).

556

557

558



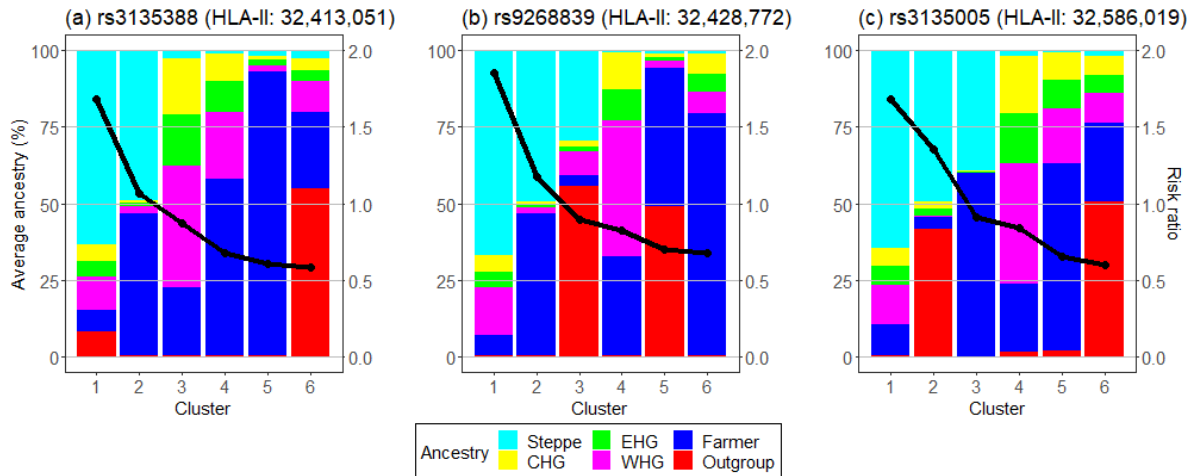
559

560 **Supplementary Figure 4.1: Association with MS risk at externally ascertained SNPs, for (top)**

561 **ancestry, and (bottom) SNPs.**

562 Due to the UK Biobank being less powered (having fewer cases) than the Case-Control study from
563 which these SNPs were found, the only statistically significant association is in the HLA.

564



565

566

567

Supplementary Figure 4.2: Comparison between MS-risk and local ancestry for 3 example SNPs.

568

In the HLA class II region, all SNPs share a pattern in which high Steppe ancestry is associated with

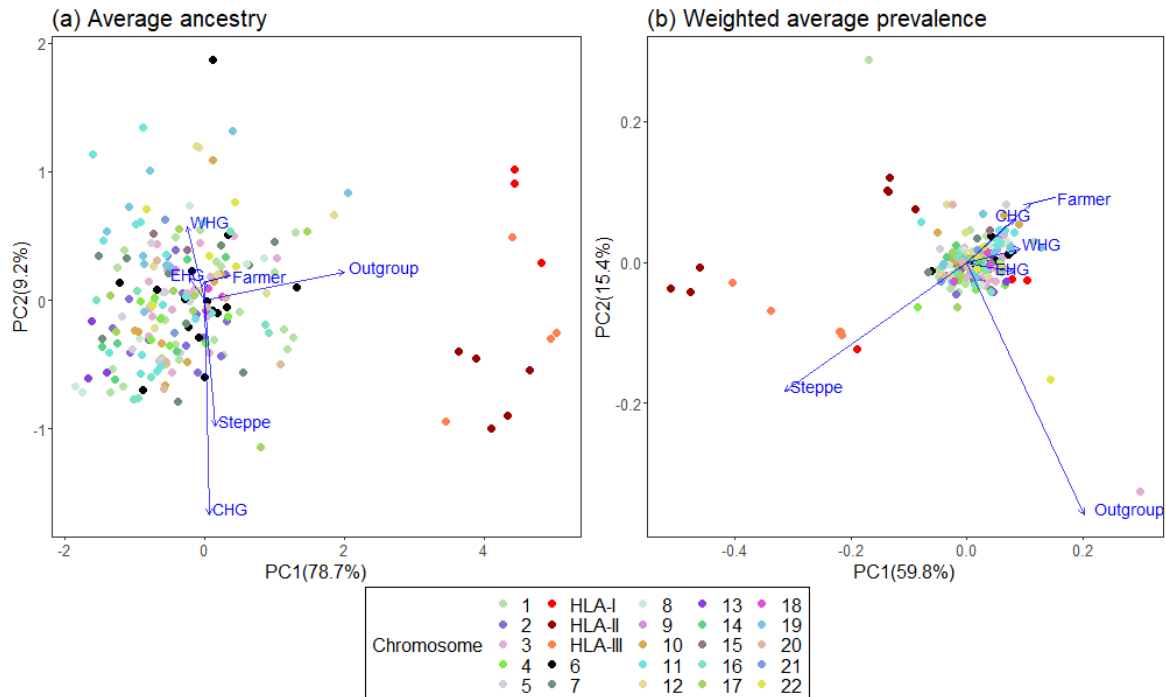
569

high MS-risk. The risk decreases monotonically and is not present in the Steppe precursor populations

570

(Hunter Gatherers), but is with the admixed Bronze-age European populations (Steppe + Farmer).

571



572

573

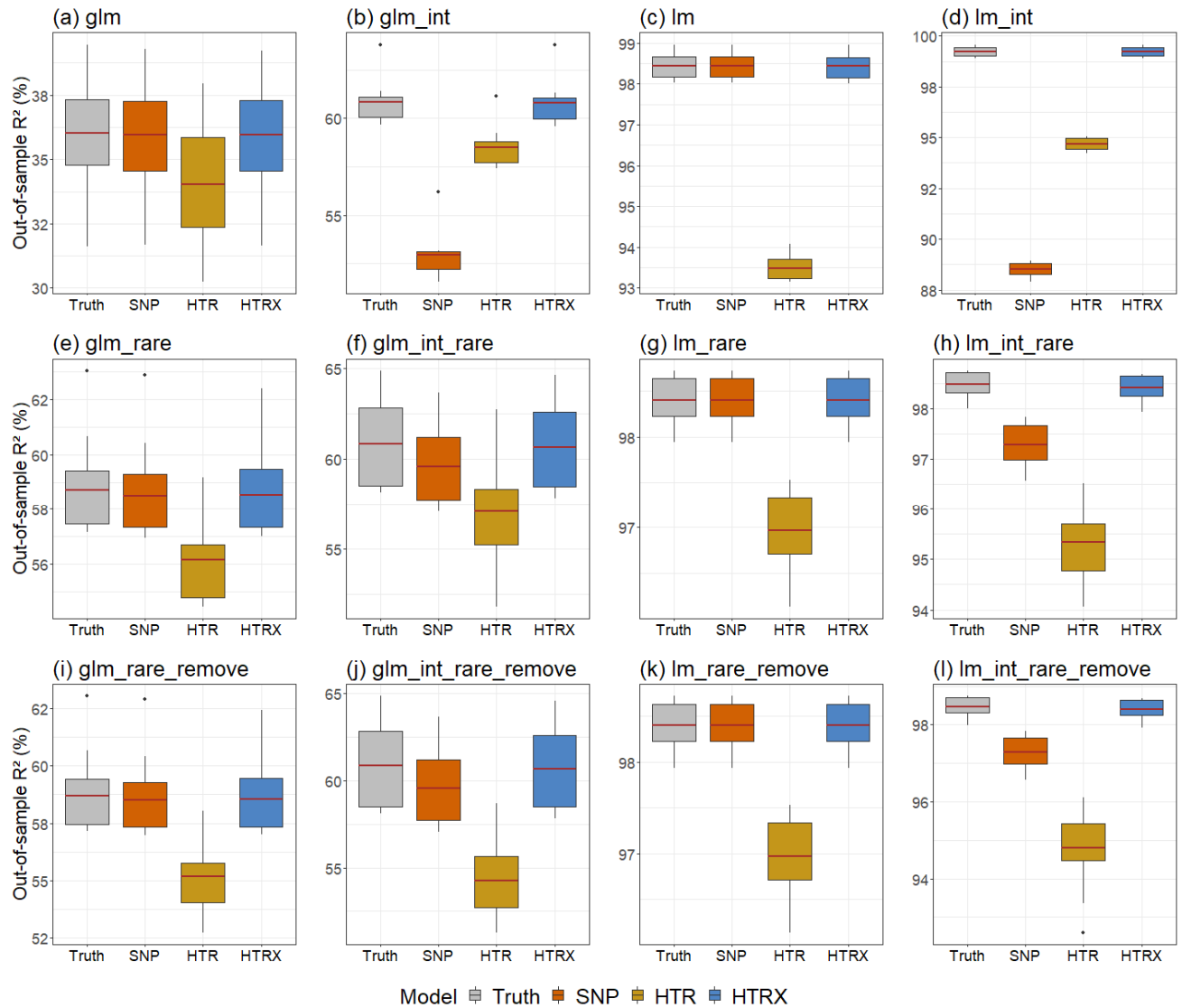
574

575

576

577

Supplementary Figure 4.3: Decomposition of individuals ancestry at MS risk SNPs in terms of (left) the ancestry of those SNPs alone, or (right) the Weighted average prevalence of MS in each ancestry after “logit” transformation.



578

579

580 **Supplementary Figure 4.4: Simulation study with four SNPs showing the boxplots of out-of-**
 581 **sample variance (with the red line representing the average) explained by HTRX compared to**
 582 **GWAS, HTR and the true model.**

582

583

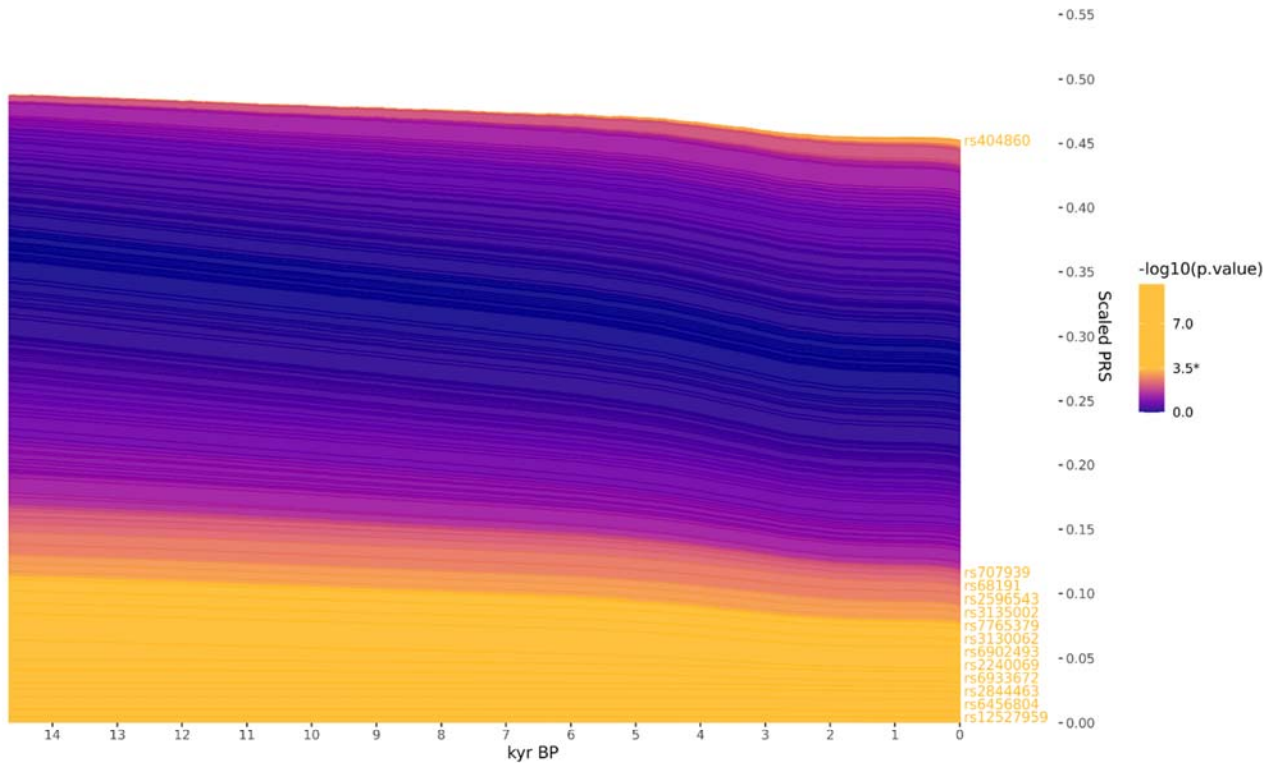
584

585

586

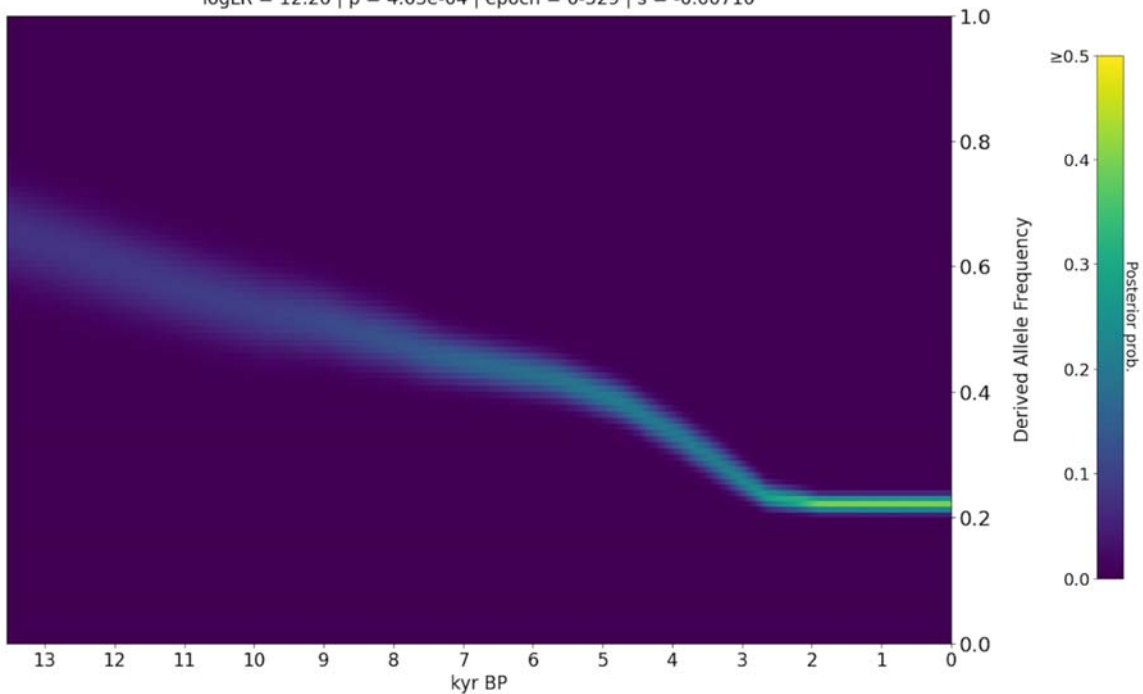
The total variance explained by HTRX is the same as SNP and bigger than HTR when there are no interactions. When interaction (with subtitle "int") exists, HTRX significantly outperforms GWAS and HTR. In all situations, HTRX works similarly to the truth.

a) Rheumatoid arthritis ($r^2 < 0.05$; window 250 kb) ($n = 153$) | All ancestries | $\omega = -0.005$ | $se = 0.0021$ | $z = -2.655$ | $p = 0.00793$



b) rs660895 | chr6:32577380 | Gene(s): HLA-DRB1 - HLA-DQA1 | A/G

$\log LR = 12.26$ | $p = 4.63e-04$ | epoch = 0.529 | $s = -0.00710$



587

588

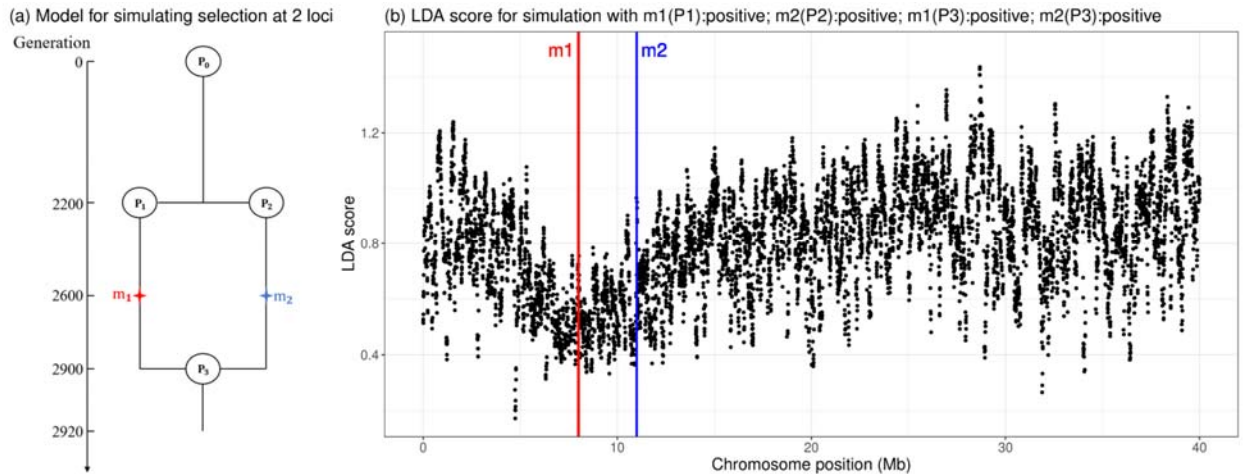
589

590

Supplementary Figure 5.1: Evidence for selection on RA-associated SNPs.

a) Stacked line plot of the pan-ancestry PALM analysis for RA, showing the contribution of alleles to disease risk over time. Individual SNPs are stacked, with their trajectories polarised to show the

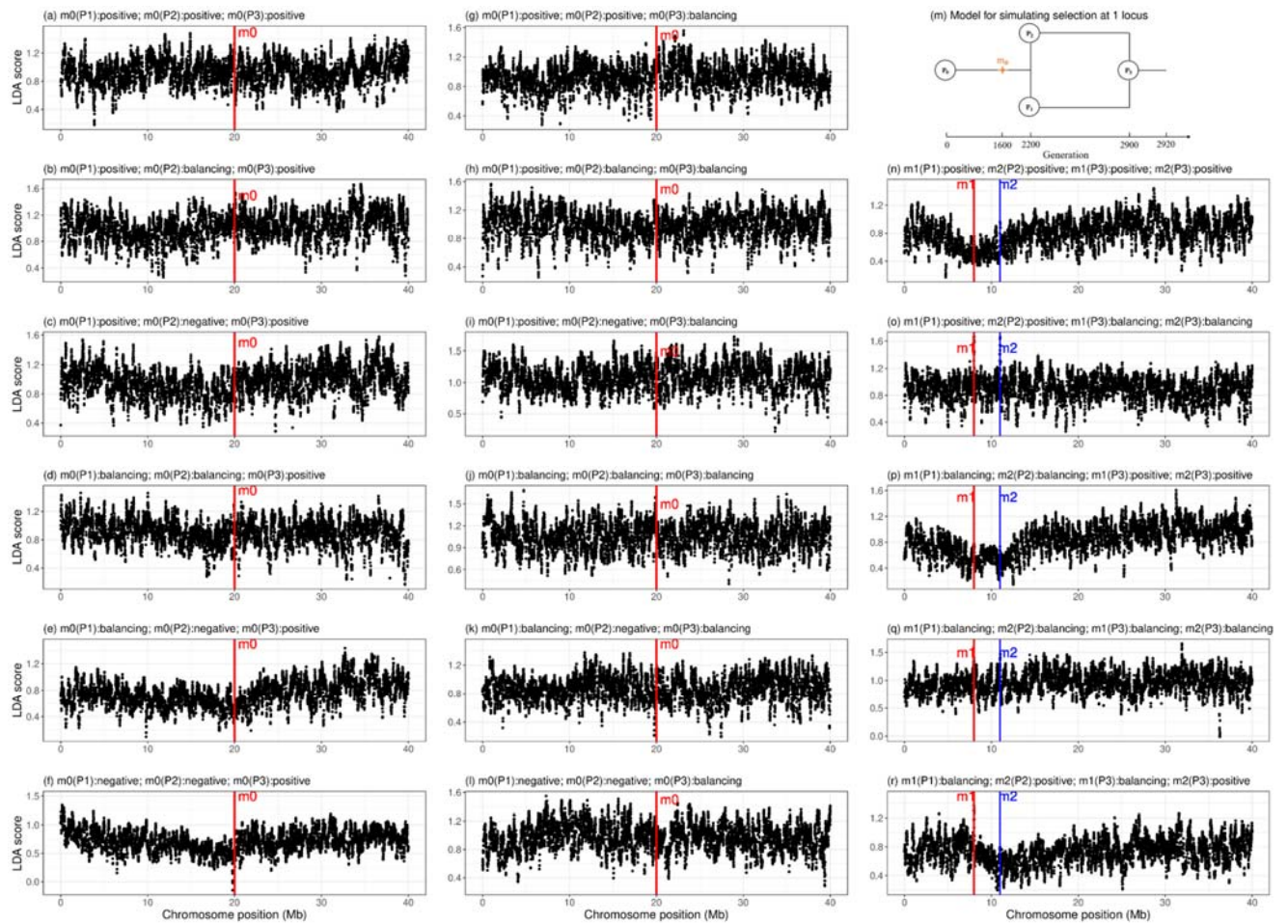
591 frequency of the positive risk allele and weighted by their scaled effect size: when a given SNP bar
592 becomes wider over time the risk allele has increased in frequency, and vice versa. SNPs are sorted by
593 their marginal p-value and direction of effect, with selected SNPs that increase risk plotted on top.
594 SNPs are also coloured by their marginal p-values, and significant SNPs are shown in yellow. The y-
595 axis shows the scaled polygenic risk score (PRS), which ranges from 0 to 1, representing the
596 maximum possible additive genetic risk in a population.
597 b) Posterior likelihood trajectory for rs660895, tagging HLA-DRB1*04:01, inferred by CLUES.
598



599
600
601
602
603
604

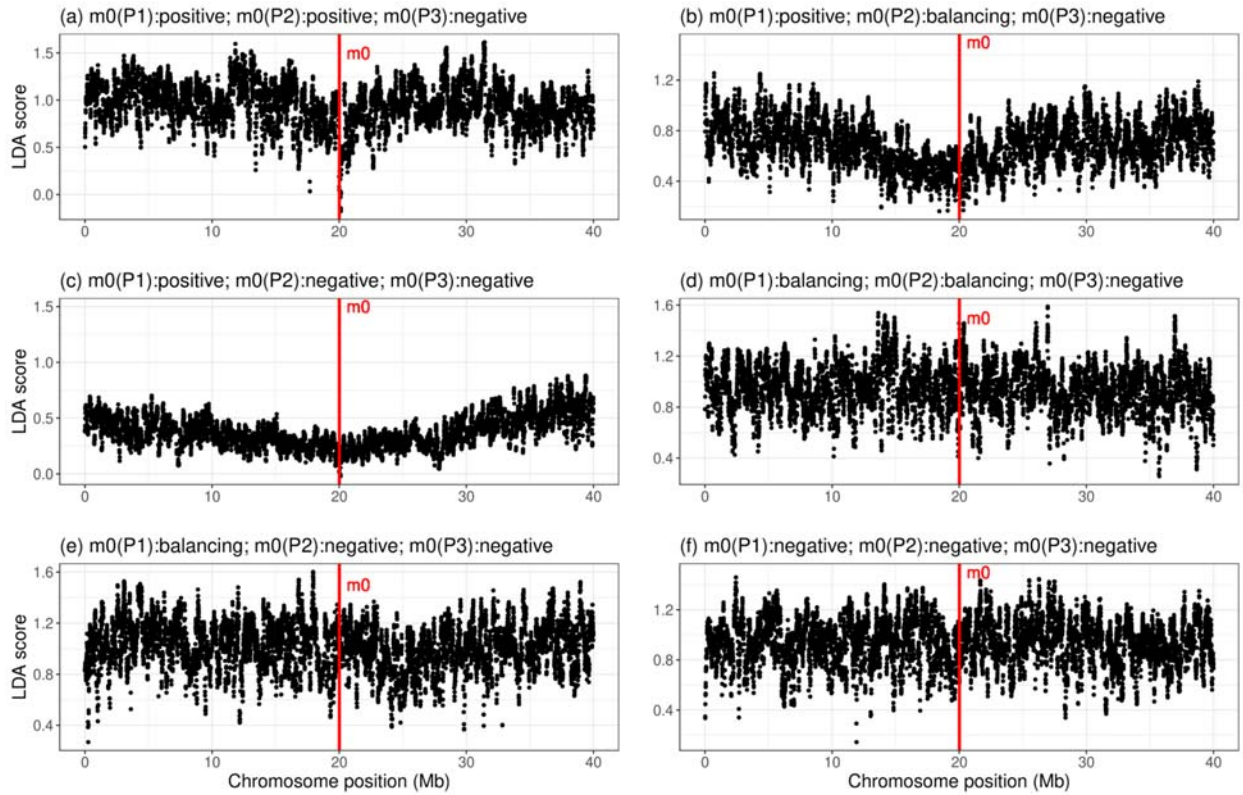
Supplementary Figure 6.1: Simulating Low LDA score.

Left: A simulated history in which a single population splits into two (“Steppe” and “Farmer”) after 2200 generations and experiences positive selection on different loci (m_1 in P_1 and m_2 in P_2). After 2900 generations the populations merge (“Europeans”) but selection continues on *both* loci.



605
606
607
608
609
610
611
612

Supplementary Figure 6.2: LDAS simulation with positive or balancing selection in the modern population. The left two columns show simulations with a single variant satisfying the observed constraint that modern-day frequencies are not decreasing (i.e. not negative selection). The right column shows simulations with two variants, also obeying this constraint. The model for simulating 2 loci is the same as in Supplementary Figure 6.1, and that for 1 locus is in the top right of this plot (which differs only in the location of the selected variant in the separated populations).



613

614

615

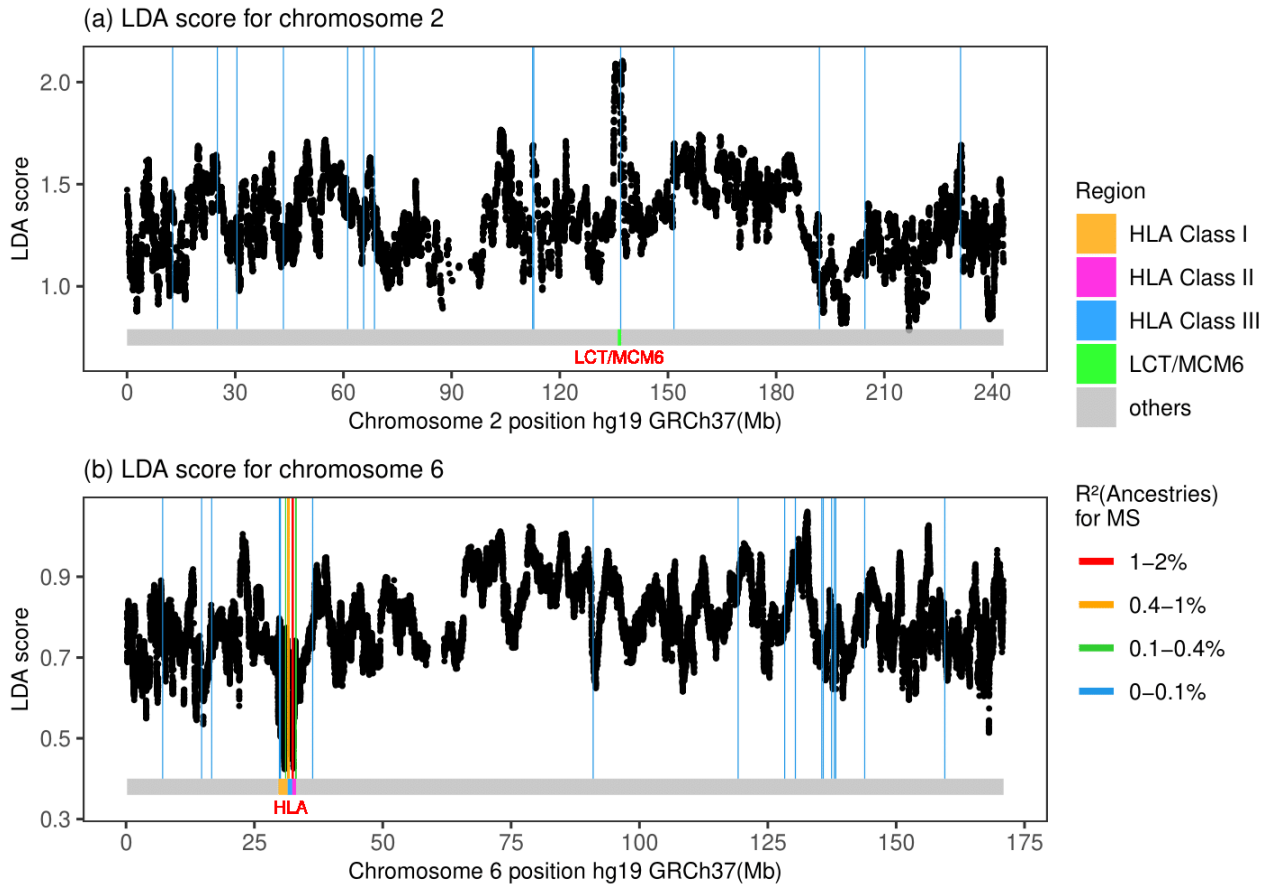
616

617

618

619

Supplementary Figure 6.3: LDAS simulation with single locus negatively selected in the modern population. In two cases this generates a low LDAS score, which requires recent negative selection (which is ruled out for HLA by the observed frequency trend). In this case, one ancestry dominates the region and recombination to the other conveys risk. The model used is in the top right of Supplementary Figure 6.2.

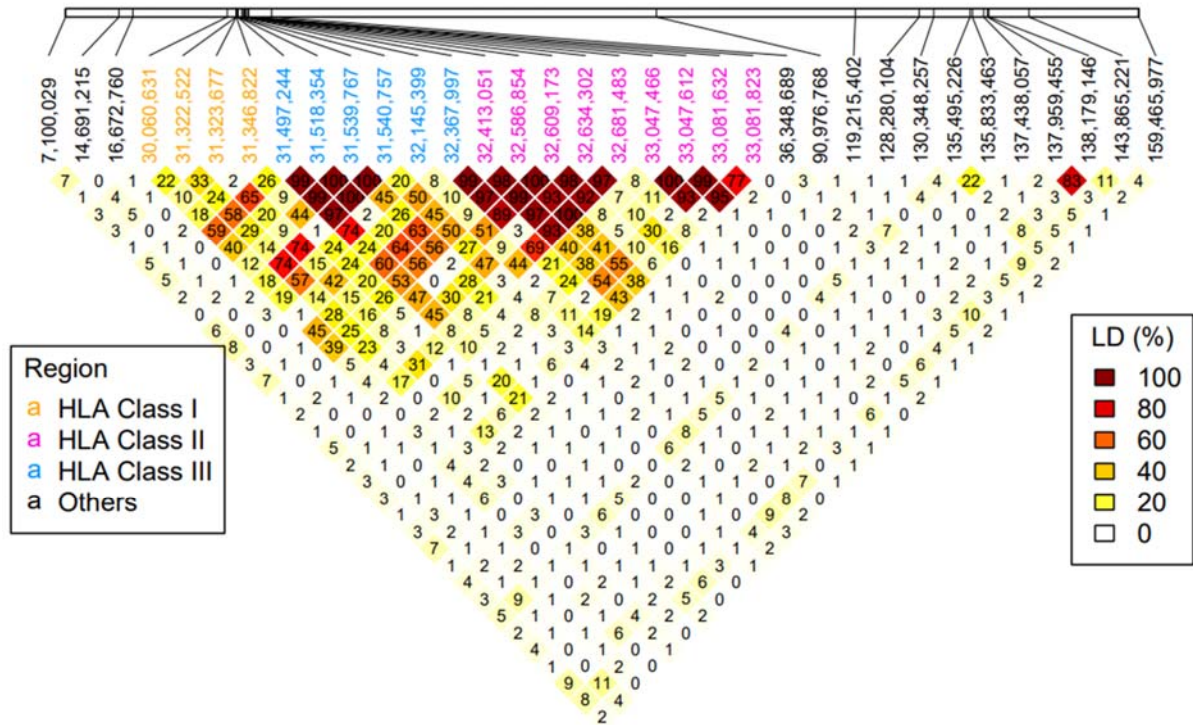


620

621

622 **Supplementary Figure 6.4: LDAS on chromosome 6 and 2.** LDA score is a) high in the

623 LCT/MCM6 region while is b) low in the HLA region.



624

625

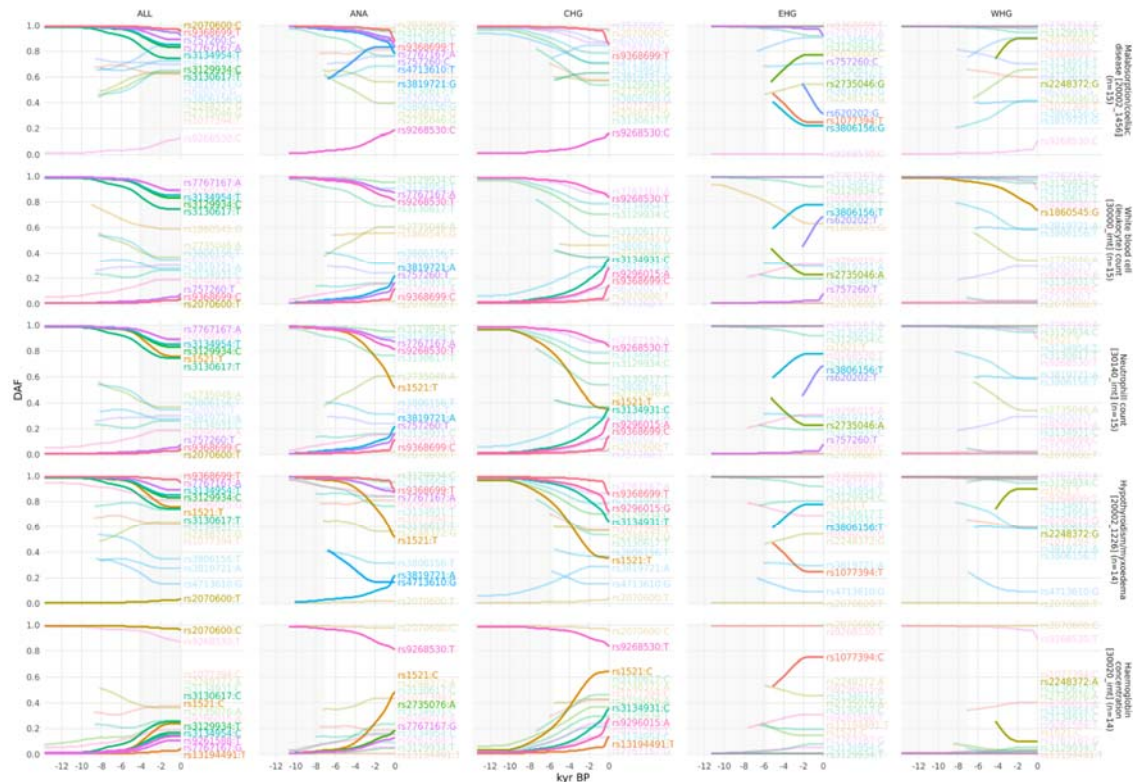
Supplementary Figure 6.5: Pairwise Linkage Disequilibrium (LD) plot (measured by D') for all the MS-associated SNPs on chromosome 6.

626

627

628

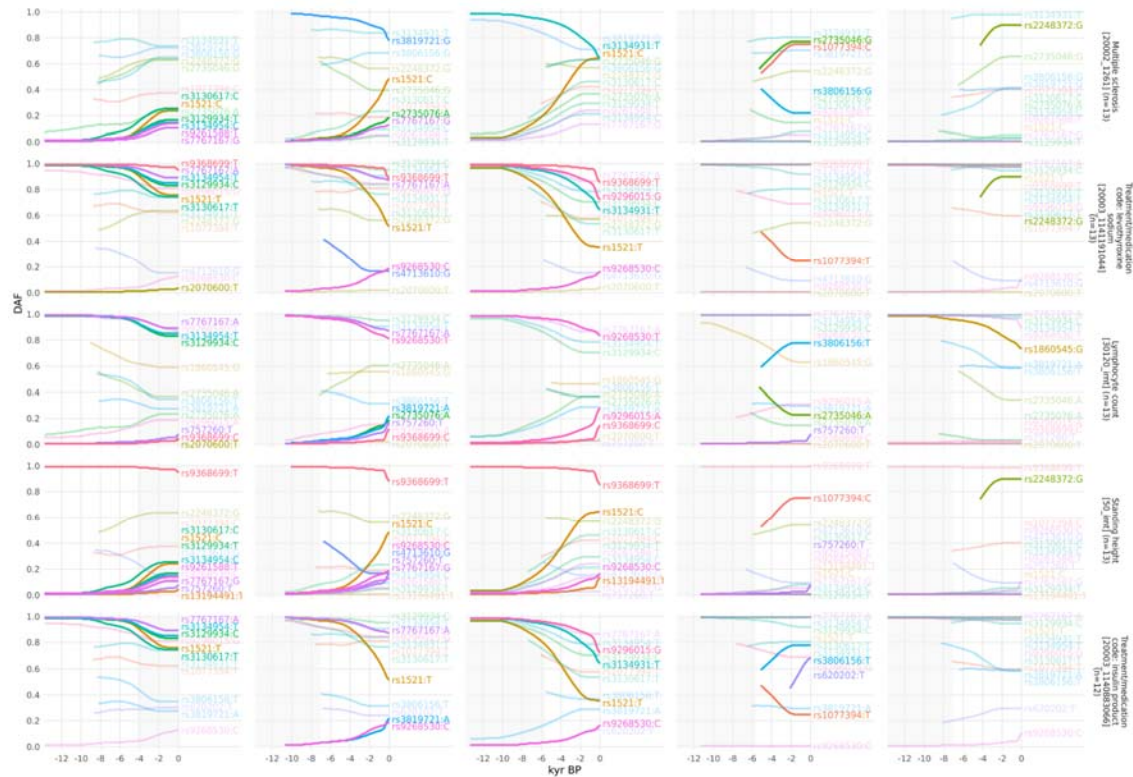
629



630

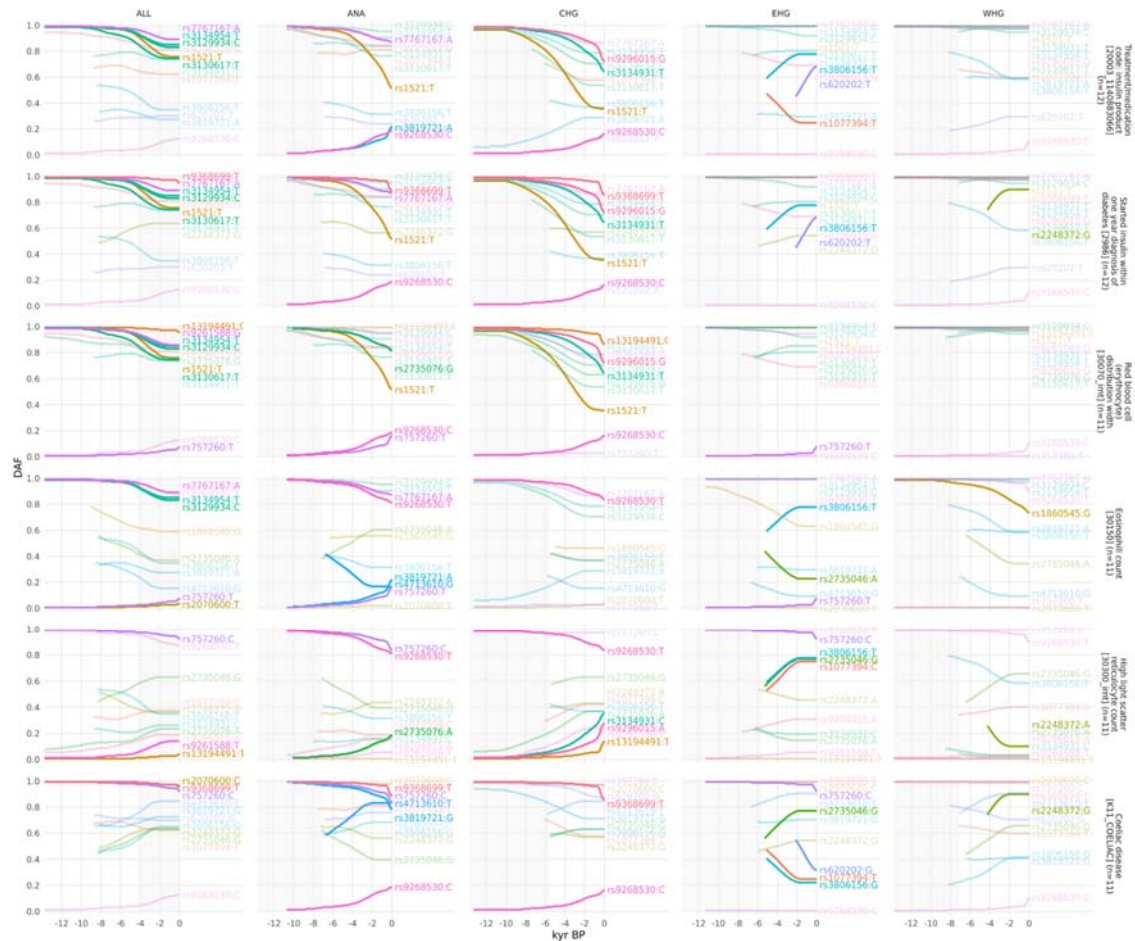
631 **Supplementary Figure 7.1 Allele frequency plots for positively selected MS-associated SNPs that**
632 **are also associated with other phenotypes in the UK Biobank. Traits 1-5.**

633 SNPs are shown with their maximum likelihood trajectories, and polarised by the direction of their
634 effect on the marginal UK Biobank trait (i.e. showing the ‘risk’ allele). Phenotypes are ordered
635 according to the number of common SNPs, non-significant SNPs are shown with partial transparency,
636 portions of the trajectory with low posterior density are cropped off, and the background is shaded for
637 the approximate time period in which the ancestry existed as an actual population.
638



639 **Supplementary Figure 7.2 Allele frequency plots for positively selected MS-associated SNPs that**
640 **are also associated with other phenotypes in the UK Biobank. Traits 6-10.**

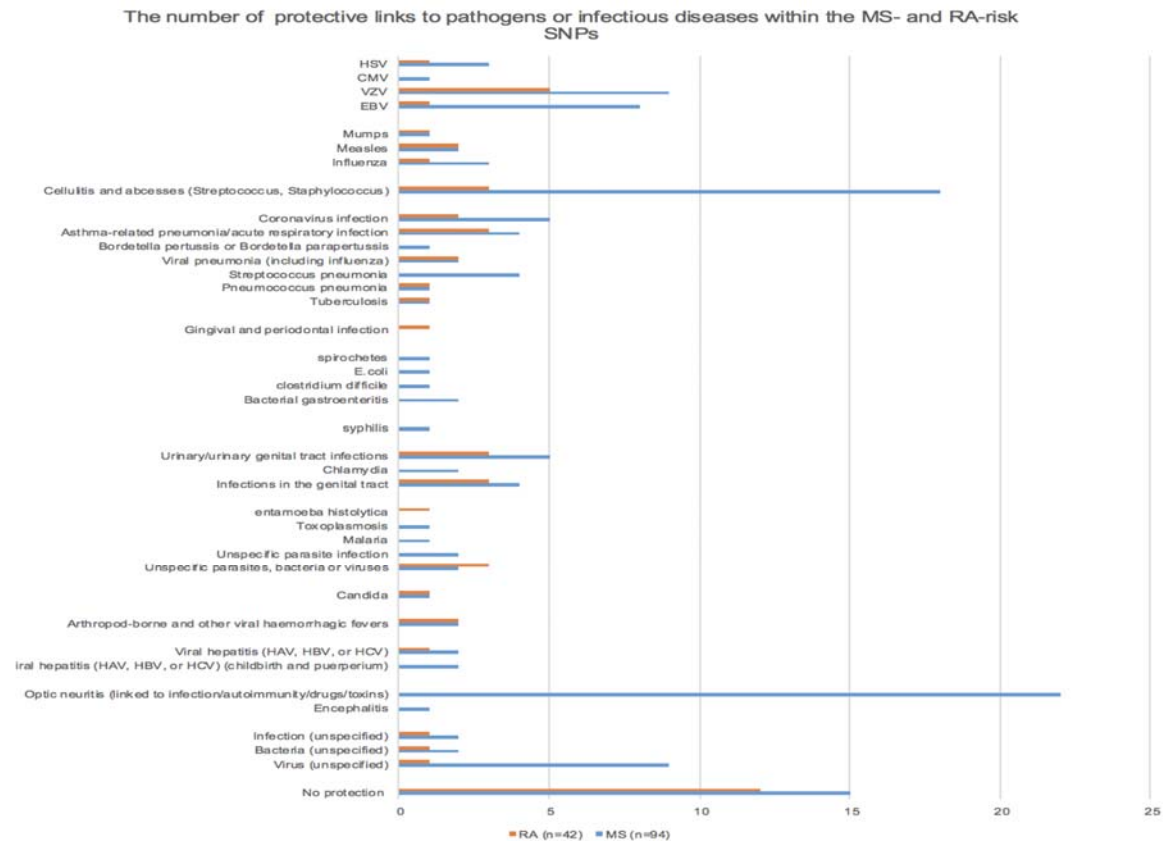
641 SNPs are shown with their maximum likelihood trajectories, and polarised by the direction of their
642 effect on the marginal UK Biobank trait (i.e. showing the ‘risk’ allele). Phenotypes are ordered
643 according to the number of common SNPs, non-significant SNPs are shown with partial transparency,
644 portions of the trajectory with low posterior density are cropped off, and the background is shaded for
645 the approximate time period in which the ancestry existed as an actual population. Note that many
646 phenotypes are underpowered in the UKBiobank GWAS, hence why MS appears as just the joint 7th
647 in this list.
648



649
650
651
652
653
654
655
656
657

Supplementary Figure 7.3 Allele frequency plots for positively selected MS-associated SNPs that are also associated with other phenotypes in the UK Biobank. Traits 11-15.

SNPs are shown with their maximum likelihood trajectories, and polarised by the direction of their effect on the marginal UK Biobank trait (i.e. showing the ‘risk’ allele). Phenotypes are ordered according to the number of common SNPs, non-significant SNPs are shown with partial transparency, portions of the trajectory with low posterior density are cropped off, and the background is shaded for the approximate time period in which the ancestry existed as an actual population.



658

659 **Supplementary Figure 8.1 The number of protective associations with pathogens or infectious**
 660 **diseases for the MA- and RA-associated selected SNPs**

661 The number of protective associations to specific pathogens and/or diseases associated with the MS-
 662 and RA-SNPs that showed statistically significant evidence for selection using CLUES. One SNP can
 663 have a link to more than one pathogen and/or disease (see ST13 and ST14 for details on each SNP).
 664 Fifteen and twelve SNPs had no detectable links to any pathogen or infectious disease in the MS and
 665 RA SNP sets, respectively.

666

667

668 **METHODS**

669 **Data Generation**

670 **Overview**

671 In order to examine variants associated with phenotypes backwards in time, we assembled a large
 672 ancient DNA dataset. Here we present new genomic data from 86 ancient individuals from Medieval
 673 and post-Medieval periods from Denmark (Supplementary Figure 1, Supplementary Note 1, ST1).
 674 The samples range in age from around the XIth to the XVIIIth century. We extracted ancient DNA
 675 from tooth cementum or petrous bone and shotgun sequenced the 86 genomes to a depth of genomic
 676 coverage ranging from 0.02 X to 1.6 X (mean = 0.39 X and median = 0.27 X). The genomes of the
 677 new 86 individuals were imputed using the 1,000 Genomes phased data as a reference panel by an
 678 imputation method designed for low coverage genomes (GLIMPSE⁴⁴), and we also imputed 1,664
 679 ancient genomes presented in the accompanying study ‘Population Genomics of Stone Age Eurasia’⁹.
 680 Depending on the specific data quality requirements for the downstream analyses, we filtered out
 681 samples with poor coverage, variant sites with low MAF and with low imputation quality (average

682 genotype probability < 0.98). Our dataset of ancient individuals span approximately 15,000 years
683 across Eurasia (Supplementary Figure 1).

684

685 **Ancient data DNA extraction and library preparation**

686 Laboratory work was conducted in the dedicated ancient DNA clean-room facilities at the Lundbeck
687 Foundation GeoGenetics Centre (Globe Institute, University of Copenhagen). A total of 86 Medieval
688 and post-Medieval human samples from Denmark (ST2) were processed using semi-automated
689 procedures. Each sample was processed in parallel. For each extract non USER-treated and USER-
690 treated (NEB) libraries were built⁴⁵. All libraries were sequenced on the NovaSeq6000 instrument at
691 the GeoGenetics Sequencing Core, Copenhagen, using S4 200 cycles kits version 1.5. A more
692 detailed description of DNA extraction and library preparation can be found in Supplementary Note 1.

693

694 **Basic bioinformatics**

695 The sequencing data was demultiplexed using the Illumina software BCL Convert
696 (https://emea.support.illumina.com/sequencing/sequencing_software/bcl-convert.html, Illumina Inc.) .
697 Adapter sequences were trimmed and overlapping reads were collapsed using AdapterRemoval
698 (2.2.4⁴⁶). Single-end collapsed reads of at least 30bp and paired-end reads were mapped to the human
699 reference genome build 37 using bwa (0.7.17⁴⁷) with seeding disabled to allow for higher sensitivity.
700 Paired- and single-end reads for each library and lane were merged, and duplicates were marked using
701 Picard MarkDuplicates (2.18.26, <http://picard.sourceforge.net>) with a pixel distance of 12000. Read
702 depth and coverage were determined using samtools (1.10⁴⁸) with the all sites used in the calculation
703 (-a). Data was then merged to sample level and duplicates were marked again.

704

705 **DNA authentication**

706 In order to determine the authenticity of the ancient reads, post-mortem DNA damage patterns were
707 quantified using mapDamage2.0⁴⁹. Next, two different methods were used to estimate the levels of
708 contamination. Firstly, we applied ContamMix in order to quantify the fraction of exogenous reads in
709 the mitochondrial reads by comparing the mtDNA consensus genome to possible contaminant
710 genomes⁵⁰. The consensus was constructed using an in-house perl script that used sites with at least 5x
711 coverage, and bases were only called if observed in at least 70% of reads covering the site. Lastly, we
712 applied ANGSD (0.931⁵¹) to estimate nuclear contamination by quantifying heterozygosity
713 on the X chromosome in males. Both contamination estimates only used filtered reads
714 with a base quality of ≥ 20 and mapping quality of ≥ 30 .

715

716 **Imputation**

717 We combined the 86 newly sequenced Medieval and post-Medieval Danish individuals with 1,664
718 previously published ancient genomes⁹. We then excluded individuals showing:
719 contamination (more than 5%); low autosomal coverage (less than 0.1 X); low
720 genome-wide average imputation genotype probability (less than 0.98), and we chose
721 the best quality sample in a close relative pair (first or second degree relative). A total
722 of 1,557 individuals passed all filters, and were used in downstream analyses. We
723 restricted the analysis to SNPs with imputation INFO score ≥ 0.5 and MAF ≥ 0.05 .

724

725 **Kinship analysis and uniparental haplogroup inferences**

726 READ⁵² was used to detect the degree of relatedness between pairs of individuals.

727 The mtDNA haplogroups of the Medieval and post-Medieval individuals were assigned using
728 HaploGrep2⁵³. Y chromosome haplogroup assignment was inferred following the workflow already
729 published⁵⁴. More details can be found in Supplementary Note 2.

730

731 **Population genetic analyses**

732 We used principal component analysis (PCA) (Supplementary Figure 1.1) to investigate the overall
733 population structure of the dataset. We used plink⁵⁵, excluding SNPs with minor allele frequency
734 (MAF) < 0.05 in the imputed panel. Based on 1,210 ancient western Eurasia imputed genomes, the
735 Medieval and post-Medieval samples cluster very close to each other, displaying a relatively low
736 genetic variability and situated within the genetic variability observed in the post-Bronze Age western
737 Eurasian populations.

738

739 We used three methods to estimate ancestry components in our ancient samples: model-based
740 clustering (ADMIXTURE⁵⁶) (Supplementary Note 1, Figure S1.1) on a subset of 826,248 SNPs;
741 qpAdm⁵⁷ (Supplementary Note 1 Figure S1.2 and Table S1.1) with a reference panel of three genetic
742 ancestries (WHG, Neolithic Farmer, and Steppe) on the same 826,248 SNPs. We performed qpAdm
743 applying the option “allsnps: YES” and a set of 7 outgroups was used as “right populations”:
744 Siberia_UpperPaleolithic_UstIshim, Siberia_UpperPaleolithic_Yana,
745 Russia_UpperPaleolithic_Sunghir, Switzerland_Mesolithic, Iran_Neolithic, Siberia_Neolithic,
746 USA_Beringia. We set a minimum threshold of 100,000 SNPs and only results with $p > 0.05$ only
747 have been considered. Finally we ran chromosome painting⁵⁸ using a panel of 7 ancestries (as on the
748 UK Biobank). We ran chromosome painting on all ancient individuals not in the reference panel,
749 using a reference panel of ancient donors grouped into populations to represent specific ancestries:
750 western hunter-gatherer (WHG), eastern hunter-gatherer (EHG), Caucasus hunter-gatherer (CHG),
751 Neolithic Farmer, Yamnaya, African and EastAsian (method described in ⁹ Supplementary Note 3h).
752 Painting followed the pipeline of ⁵⁹ based on GLOBETROTTER⁶⁰, with admixture proportions
753 estimated using Non-Negative Least squares. We also painted individuals born in Denmark of a
754 typical ancestry based on density-based clustering of the first 18 PCs⁹. This generated both local
755 ancestry probabilities and genome-wide ancestry fractions for each painted individual. The reference
756 panel used for chromosome painting was designed to capture the various components of European
757 ancestry only, and so we urge caution in interpreting these results for non-European samples.

758

759 This dataset provides the opportunity to study the population history of Denmark from the Mesolithic
760 to the post-Medieval period, covering around 10,000 years, which can be considered a typical
761 Northern European population. Our results clearly demonstrate the impact of previously described
762 demographic events, including the influx of Neolithic Farmer ancestry ~9,000 years ago and Steppe
763 ancestry ~5,000 years ago^{12, 10}. We highlight genetic continuity from the Bronze Age to the post-
764 Medieval period (Supplementary Note 1 Figure S1.1), although qpAdm detected a small increase in
765 the Neolithic Farmer component during the Viking Age (Supplementary Note 1 Figure S1.2 and
766 Table S1.1), while the Medieval period marked a time of increased genetic diversity, likely reflecting
767 increased mobility across Europe. This genetic continuity is further confirmed by the haplogroups
768 identified in the uniparental genetic markers (Supplementary Note 2). Together, these results suggest
769 that after the Bronze Age Steppe migration there was no other major gene flow into Denmark from
770 populations with significantly different Neolithic and Bronze Age ancestry compositions, and
771 therefore no changes in these ancestry components in the Danish population.

772

773 **Local ancestry**

774 We used two estimates of local ancestry from ⁹: (1) first coalescent labels generated by running
775 Chromopainter⁵⁸ on all “White British” individuals in the UK Biobank, using the same reference
776 panel described above. Henceforth ‘local ancestry’. (2) Ancestry path labels in GBR, FIN and TSI
777 1000G populations⁶¹) and 1015 ancient genomes generated using a neural network to assign ancestry
778 paths based on a sample’s nearest neighbours at the first five informative nodes of a marginal tree
779 sequence, where an informative node is defined as one which has at least one leaf from the reference
780 set of ancient samples described above (⁹ Supplementary Note S3i). Henceforth ‘ancestry path labels’.
781

782 **SNP associations**

783 We aimed to generate SNP associations from previous studies for each phenotype in a consistent
784 approach. To generate a list of SNPs associated with multiple sclerosis (MS), rheumatoid arthritis
785 (RA) and celiac disease (CD), we used two approaches: in the first, we downloaded fine-mapped
786 SNPs from previous association studies. For each fine-mapped SNP, if the SNP did not have an
787 ancestry path label, we found the SNP in highest LD that did, with a minimum threshold of $r^2 \geq 0.7$
788 in the GBR, FIN and TSI 1000G populations using LDLinkR⁶². The final SNPs used for each
789 phenotype can be found in ST4 (MS), ST5 (RA), and ST6 (CD).
790

791 For MS, we used data from ³. For non-MHC SNPs, we used the ‘discovery’ SNPs with P(joined) and
792 OR(joined) generated in the replication phase. For MHC variants, we searched the literature for the
793 reported HLA alleles and amino-acid polymorphisms (ST3). In total, we generated 205 SNPs which
794 were either fine-mapped or in high LD with a fine-mapped SNP (15 MHC, 190 non-MHC).
795

796 For RA, we downloaded 57 genome-wide significant non-MHC SNPs for seropositive RA in
797 Europeans⁶³. We retrieved MHC associations separately (⁶⁴, with associated ORs and p-values from
798 ⁶⁵). In total, we generated 51 SNPs which were either fine-mapped or in high LD with a fine-mapped
799 SNP (3 MHC, 48 non-MHC).
800

801 For CD, we retrieved non-MHC SNPs from ⁶⁶. We used MHC SNPs from ⁶⁷, with associated ORs and
802 p-values from ⁶⁸. In total, this generated 32 SNPs which were either fine-mapped or in high LD with a
803 fine-mapped SNP (3 MHC, 29 non-MHC).
804

805 Secondly, because we could not always find tag SNPs for fine-mapped SNPs that were present in our
806 ancestry path labels dataset, we found that we were losing significant signal from the HLA, therefore
807 we generated a second set of SNP associations. We downloaded full summary statistics for each
808 disease (MS: ³; RA: ⁶⁹, CD: <http://www.nealelab.is/uk-biobank/>), restricted to sites present in the
809 ancestry path labels dataset, and ran Plink’s (PLINK v1.90b4.4⁷⁰) clump method (parameters: --
810 clump-p1 5e-8 --clump-r2 0.05 --clump-kb 250 as in ⁷¹ using LD in the GBR, FIN and TSI 1000G
811 populations⁶¹ to extract genome-wide significant independent SNPs.
812

813 In the main text we report results for the first set of SNPs (‘fine-mapped’) for analyses involving local
814 ancestry in modern data, and the second set of SNPs (‘pruned’) for analyses involving polygenic
815 measures of selection (CLUES/PALM).
816

817 **REGIONS OF UNUSUAL ANCESTRY AND GENE ENRICHMENT**

818 To assess which regions of ancestry were unusual, we converted the ancestry estimates to a Z-score.
819 Specifically, we let $A(i, j, k)$ denote the probability of the k th ancestry ($k = 1, \dots, K$) at the j th SNP
820 ($j = 1, \dots, J$) of a chromosome for the i th individual ($i = 1, \dots, N$). We then computed the mean

821 painting for each SNP, $A(j, k) = \frac{1}{N} \sum_{i=1}^N A(i, j, k)$. From this we estimated a scale parameter μ_k
822 and deviation parameter σ_k using a block-median approach. Specifically we partitioned the genome
823 into 0.5Mb regions, and within each, computed the mean and standard deviation of the ancestry. The
824 parameter estimates are then the median values over the whole genome. We then computed an
825 anomaly score for each SNP for each ancestry $Z(j, k) = (A(j, k) - \mu_k) / \sigma_k$.

826

827 To arrive at an anomaly score for each SNP aggregated over all ancestries, we also had to account for
828 correlations in the ancestry paintings. Instead of scaling each ancestry deviation $A^*(j, k) = A(j, k) -$
829 μ_k by its standard deviation, we instead “whitened” them, i.e. rotated the data to have an independent
830 signal. Let $C = A^{*T} A^*$ by a $K \times K$ covariance matrix, and let $C^{-1} = UDV^T$ be the Singular Value
831 Decomposition. Then $W = UD^{1/2}$ is the whitening matrix from which $Z = A^*W$ are normally
832 distributed with covariance matrix $\text{diag}(1)$ under the null that A^* is normally distributed with mean 0
833 and unknown covariance Σ . The “ancestry anomaly score” test statistic for each SNP is $t(j) =$
834 $\sum_{k=1}^K Z(j, k)^2$, which is Chi-squared distributed with K degrees of freedom under the null, and we
835 reported p-values from this.

836

837 To test for gene enrichment we formed a list of all SNPs reaching genome-wide significance ($p <$
838 5^{-8}) and using the R package *gprofiler2*⁷² converted these to a unique list of genes. We then used *gost*
839 to perform an enrichment test for each GO term, for which we used default p-value correction via the
840 *g:Profiler SCS* method. This is an empirical correction based on performing random lookups of the
841 same number of genes under the null, to control the error rate and ensure that 95% of reported
842 categories (at $p=0.05$) are correct.

843

844 **ALLELE FREQUENCY PLOTS OVER TIME**

845 To investigate how effect allele frequencies have changed over time, we extracted high effect alleles
846 for each phenotype from the ancient data. We excluded all non-Eurasian samples, grouped them by
847 ‘groupLabel’, excluded any group with fewer than 4 samples, and coloured points by ancestry
848 proportion according to genome-wide NNLS based on chromosome painting (above).

849

850 **CLUSTER ANALYSIS**

851 In order to understand whether risk-conferring haplotypes evolved in the Steppe population, or in a
852 pre- or post-dating population in which Steppe ancestry is high, we used k-means clustering on the
853 dosage of each ancestry for each selected significant SNP and investigated the dosage distribution of
854 clusters with significantly higher MS prevalence. For the target SNPs, the Elbow method⁷³ suggested
855 selecting around 5-7 clusters, of which we chose 6. After performing the k-means cluster analysis, we
856 calculated the average probability for each ancestry for case individuals. Furthermore, we calculated
857 the prevalence of MS in each cluster, and performed a one-sample t-test to investigate whether it
858 differs from the overall MS prevalence (0.487%). This tests whether particular combinations of
859 ancestry are associated with the phenotype at a SNP. Clusters with high MS risk-ratios have high
860 Steppe components (Supplementary Figure 4.2), leading to the conclusion that Steppe ancestry alone
861 is driving this signal.

862

863 **WEIGHTED AVERAGE PREVALENCE**

864 In order to quantify the risk of each ancestry for each SNP, we calculated the weighted average
865 prevalence (WAP) for each ancestry based on the result of k-means clustering (above).

866

867 For the j th SNP, let $P_{jkm} = n_{jm}P_{jkm}$ denote the sum of the k th ancestry probabilities of all the
868 individuals in the m th cluster ($k, m = 1, \dots, 6$), where n_{jm} is the cluster size of the m th cluster. Let
869 π_{jm} denote the prevalence of MS in the m th cluster, the weighted average prevalence for the k th
870 ancestry is defined as:

$$871 \quad \pi_{jk} = \frac{P_{jkm}\pi_{jm}}{\sum_{m=1}^6 P_{jkm}},$$

872 where P_{jkm} is defined as the weight for each cluster.

873

874 For each ancestry, WAP measures the association of that ancestry with MS risk across all clusters. To
875 make a clear comparison, we calculated the risk ratio (compared to the overall MS prevalence) for
876 each ancestry at each SNP, and assigned a mean and confidence interval for the risk ratios of each
877 ancestry at each chromosome (Figure 3, Supplementary Figure 3.1 and 3.2).

878

879 **PCA/UMAP OF WAP/AVERAGE DOSAGE**

880 We performed principal component analysis (PCA) on the average ancestry probability and WAP at
881 each MS-associated SNP (Supplementary Figure 4.3). The former shows that all of the HLA SNPs
882 except three from HLA class II and III have much larger Outgroup components compared with the
883 others. The latter analysis indicates a strong association between Steppe and MS risk. Also, Outgroup
884 ancestry at rs10914539 from chromosome 1 exceptionally reduces the incidence of MS, while
885 Outgroup ancestry at rs771767 (chromosome 3) and rs137956 (chromosome 22) significantly boosts
886 MS risk.

887

888 **ANCESTRAL RISK SCORES**

889 Following methods developed in Irving-Pease et al. (*submitted*), we calculated the effect allele
890 painting frequency for a given ancestry $f_{\{anc,i\}}$ for SNP i using the formula:

$$891 \quad f_{\{anc,i\}} = \frac{\sum_j^{M_{effect}} \text{Painting certainty}_{\{j,i,anc\}}}{\sum_j^{M_{alt}} \text{Painting certainty}_{\{j,i,anc\}} + \sum_j^{M_{effect}} \text{Painting certainty}_{\{j,i,anc\}}},$$

892 where there are M_{effect} individuals homozygous for the effect allele, M_{alt} individuals homozygous
893 for the other allele, and $\sum_j^{M_{effect}} \text{Painting certainty}_{\{j,i,anc\}}$ is the sum of the painting
894 probabilities for that ancestry anc in individuals homozygous for the effect allele at SNP i . This can
895 be interpreted as an estimate of an ancestral contribution to effect allele frequency in a modern
896 population. Per-SNP painting frequencies can be found in ST4, ST5, and ST6.

897

898 To calculate the ancestral risk score (ARS) we summed over all I pruned SNPs in an additive model:

$$899 \quad ARS_{anc} = \sum_i^I f_{\{anc,i\}} * \beta_{i.}$$

900

901 We then ran a transformation step as in ⁷⁴. To obtain 95% confidence intervals, we ran an accelerated
902 bootstrap over loci, which accounts for the skew of data to better estimate confidence intervals ⁷⁵.

903

904 **LOCAL ANCESTRY AND GENOTYPE GWAS**

905 We used the UK Biobank to fit GWAS models for local ancestry values and genotype values
906 separately, using only SNPs known to be associated with the phenotype ('fine-mapped' SNPs). We

907 used the following phenotype codes for each phenotype: MS: Data-Field 131043; RA: Data-Field
908 131849 (seropositive); CD: Data-Field 21068.
909

910 Let Y_i denote the phenotype status for the i th individual ($i = 1, \dots, 399998$), which takes value 1 for a
911 case and 0 for control, and let $\pi_i = Pr(Y_i = 1)$ denote the probability that this individual has the
912 event. Let X_{ijk} denote the k th ancestry probability ($k = 1, \dots, K$) for the j th SNP ($j = 1, \dots, 205$) of
913 the i th individual. C_{ic} is the c th predictor ($c = 1, \dots, N_c$) for the i th individual. We used the following
914 logistic regression model for GWAS, which assumes the effects of alleles are additive:

$$915 \quad Y_i \sim Bin(1, \pi_i); \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{k=1}^K \beta_{jk} X_{ijk} + \sum_{c=1}^{N_c} \gamma_c C_{ic}.$$

916
917 We used $N_c=20$ predictors in the GWAS models, including sex, age and the first 18 PCs, which are
918 sufficient to capture most of the population structure in the UK Biobank⁷⁶.
919

920 First, we built the model with $K = 1$. By using only one ancestry probability in each model, we aimed
921 to find the statistical significance of each SNP under each ancestry. Then, we built the model with
922 $K = 5$, i.e. using all 6 local ancestry probabilities which sum to 1. We calculated the variance
923 explained by each SNP by summing up the variance explained by X_{ijk} ($k=1, \dots, 5$).
924

925 We considered fitting the multivariate models by using all the SNPs as covariates. However, the
926 dataset only contains 1,982 cases. Even though only one ancestry is included, the multivariate model
927 contains 191 predictors, which could result in overfitting problems. Therefore, the GWAS models are
928 preferred over multivariate models.
929

930 We also fitted a logistic regression model for GWAS using the genotype data as follows:

$$931 \quad Y_i \sim Bin(1, \pi_i); \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_j X_{ij} + \sum_{c=1}^{N_c} \gamma_c C_{ic},$$

932 where $X_{ij} \in \{0, 1, 2\}$ denotes the number of copies of the reference allele of the j th SNP ($j =$
933 $1, \dots, 205$) that the i th individual has, and C_{ic} ($c = 1, \dots, N_c$) denotes the covariates including age, sex
934 and first 18 PCs for the i th individual, where $N_c=20$. Due to the UK Biobank being underpowered
935 compared to the Case-Control study from which these SNPs were found, the only statistically
936 significant (at $p < 10^{-5}$) association is in the HLA class II tagging HLA-DRB1*15:01.
937

938 **COMPARISON OF GWAS MODELS USING PAINTING AND GENOTYPE DATA**

939 We compared the variance explained by SNPs from the GWAS model using the painting data (all 6
940 local ancestry probabilities) with that from GWAS model using the genotype data. McFadden's
941 pseudo R squared measure⁷⁷ is widely used for estimating the variance explained by the logistic
942 regression models. McFadden's pseudo R squared is defined as

$$943 \quad R^2 = 1 - \frac{\ln(L_M)}{\ln(L_0)},$$

944 where L_M and L_0 are the likelihoods for the fitted and the null model, respectively. Taking overfitting
945 into account, we propose the adjusted McFadden's pseudo R squared by penalizing the number of
946 predictors:

$$947 \quad Adjusted R^2 = 1 - \frac{\ln(L_M)/(N - k)}{\ln(L_0)/(N - 1)},$$

948 where N is the sample size and k is the number of predictors.

949

950 Specifically, $R^2(SNPs)$ is calculated as the extra variance in addition to sex, age and 18 PCs that can
951 be explained by SNPs:

$$952 \quad R^2(SNPs) = R^2(\text{sex} + \text{age} + 18 \text{ PCs} + SNPs) - R^2(\text{sex} + \text{age} + 18 \text{ PCs}).$$

953

954 Notably, two SNPs stand out for explaining much larger variance than others when fitting the GWAS
955 model using the genotype data, but overall more SNPs from GWAS painting explain more than 0.1%
956 variance, which indicates the painting data are probably more efficient for estimating the effect sizes
957 of SNPs and detecting significant SNPs. Also, some SNPs from GWAS models using painting data
958 explain almost the same amount of variance, suggesting that these SNPs consist of very similar
959 ancestries.

960

961 **HAPLOTYPE TREND REGRESSION WITH eXtra FLEXIBILITY (HTRX)**

962 We propose Haplotype Trend Regression with eXtra flexibility (HTRX) which searches for haplotype
963 patterns that include single SNPs and non-contiguous haplotypes. HTRX is an association between a
964 template of n SNPs and a phenotype. A template gives a value for each SNP taking values of '0' or
965 '1', reflecting whether the reference allele of each SNP is present or absent, or an 'X' meaning either
966 value is allowed. For example, haplotype '1X0' corresponds to a 3-SNP haplotype where the first
967 SNP is the alternative allele and the third SNP is the reference allele, while the second SNP can be
968 either the reference or the alternative allele. Therefore, haplotype '1X0' is essentially only a 2-SNP
969 haplotype.

970

971 To examine the association between a haplotype and a binary phenotype, we replace the genotype
972 term with a haplotype from the standard GWAS model:

$$973 \quad Y_i \sim \text{Bin}(1, \pi_i); \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_j H_{ij} + \sum_{c=1}^{N_c} \gamma_c C_{ic},$$

974 where H_{ij} denotes the j th haplotype probability for the i th individual:

$$H_{ij} = \begin{cases} 1 & \text{if } i\text{th individual has haplotype } j \text{ in both genomes,} \\ 1/2 & \text{if } i\text{th individual has haplotype } j \text{ in one of the two genomes,} \\ 0 & \text{otherwise.} \end{cases}$$

975

976

977 HTRX can identify gene-gene interactions, and is superior to HTR not only because it can extract
978 combinations of significant SNPs within a region, leading to improved predictive performance, but
979 the haplotypes are more interpretable as multi-SNP haplotypes are only reported when they lead to
980 increased predictive performance.

981

982 **HTRX Model selection procedure for shorter haplotypes**

983 Fitting HTRX models directly on the whole dataset can lead to significant overfitting, especially when
984 the number of SNPs increases. When overfitting occurs, the models experience poorer predictive
985 accuracy against unseen data. Further, HTRX introduces an enormous model space which much be
986 searched.

987

988 To address these problems, we implement a two-step procedure.

989

990 Step 1: select **candidate** models. This is to address the model search problem, and is chosen to obtain
991 a set of models more diverse than traditional bootstrap resampling (Efron, 1979⁷⁸).

992

993 (1) Randomly sample a subset (50%) of data. Specifically, when the outcome is binary, stratified
994 sampling is used to ensure the subset has approximately the same proportion of cases and controls as
995 the whole data;

996

997 (2) Start from a model with fixed covariates (18 PCs, sex and age), and perform forward regression on
998 the subset, i.e. iteratively choose a feature (in addition to the fixed covariates) to add whose inclusion
999 enables the model to explain the largest variance, and select s models with the lowest Bayesian
1000 Information Criteria (BIC)⁷⁹ to enter the candidate model pool;

1001

1002 (3) repeat (1)-(2) B times, and select all the different models in the candidate model pool as the
1003 candidate models.

1004

1005 Step 2: select the best model using 10-fold cross-validation.

1006

1007 (1) Randomly split the whole data into 10 groups with approximately equal sizes, using stratified
1008 sampling when the outcome is binary;

1009

1010 (2) In each of the 10 folds, use a different group as the test dataset, and take the remaining groups as
1011 the training dataset. Then, fit all the candidate models on the training dataset, and use these fitted
1012 models to compute the additional variance explained by features (out-of-sample R^2) in the test
1013 dataset. Finally, select the candidate model with the biggest average out-of-sample R^2 as the best
1014 model.

1015

1016 **HTRX Model selection procedure for longer haplotypes (Cumulative HTRX)**

1017 Longer haplotypes are important for discovering interactions. However, there are $3^k - 1$ haplotypes
1018 in HTRX if the region contains k SNPs, making it unrealistic for regions with large numbers of SNPs.
1019 To address this issue, we proposed cumulative HTRX to control the number of haplotypes, which is
1020 also a two-step procedure.

1021

1022 Step 1: extend haplotypes and select candidate models.

1023

1024 (1) Randomly sample a subset (50%) of data, use stratified sampling when the outcome is binary. This
1025 subset is used for all the analysis in (2) and (3);

1026

1027 (2) Start with L randomly chosen SNPs from the entire k SNPs, and keep the top M haplotypes that
1028 are chosen from the forward regression. Then add another SNP to the M haplotypes to create $3M + 2$
1029 haplotypes. There are $3M$ haplotypes obtained by adding '0', '1' or 'X' to the previous M haplotypes,
1030 as well as 2 bases of the added SNP, i.e. 'XX...X0' and 'XX...X1' (as 'X' was implicitly used in the
1031 previous step). The top M haplotypes from them are then selected using forward regression. Repeat
1032 this process until obtaining M haplotypes which include $k - 1$ SNPs;

1033

1034 (3) Add the last SNP to create $3M + 2$ haplotypes. Afterwards, start from a model with fixed
1035 covariates (18 PCs, sex and age), perform forward regression on the training set, and select s models
1036 with the lowest BIC to enter the candidate model pool;

1037

1038 (4) repeat (1)-(3) B times, and select all the different models in the candidate model pool as the
1039 candidate models.

1040

1041 Step 2: select the best model using 10-fold cross-validation, as described in “**HTRX Model selection**
1042 **procedure for shorter haplotypes**”.

1043

1044 We note that because the search procedure in Step 1(2) may miss some highly predictive haplotypes,
1045 cumulative HTRX acts as a lower bound on the variance explainable by HTRX.

1046

1047 As a model criticism, only common and highly predictive haplotypes (i.e. those with the greatest
1048 adjusted R^2) are correctly identified, but the increased complexity of the search space of HTRX leads
1049 to haplotype subsets that are not significant on their own but are significant when interacting with
1050 other haplotype subsets being missed. This issue would be eased if we increase all the parameters s , l ,
1051 M and B but with higher computational cost, or improve the search by optimizing the order of adding
1052 SNPs. This leads to a decreased certainty that the exact haplotypes proposed are ‘correct’, but together
1053 reinforces the inference that interaction is extremely important.

1054

1055 **Simulation for HTRX**

1056 To investigate how the total variance explained by HTRX compare to GWAS and HTR, we used a
1057 simulation study comparing:

1058 (1) linear models (denoted by "lm") and generalized linear models with a logit link-function (denoted
1059 by "glm");

1060 (2) models with or without actual interaction effects;

1061 (3) models with or without rare SNPs (frequency smaller than 5%);

1062 (4) remove or retain rare haplotypes when rare SNPs exist.

1063

1064 We started from creating the genotypes for 4 different SNPs G_{ijq} ($i = 1, \dots, 100,000$ denotes the
1065 index of individuals, $j = 1$ ("1XXX"), 2 ("X1XX"), 3 ("XX1X") and 4 ("XXX1") represents the index
1066 of SNPs, and $q = 1, 2$ for two genomes as individuals are diploid). If no rare SNPs were included, we
1067 sampled the frequency F_j of these 4 SNPs from 5% to 95%; otherwise, we sampled the frequency of
1068 the first 2 SNPs from 2% to 5% (in practice, we obtained $F_1 = 2.8\%$ and $F_2 = 3.1\%$ under our seed)
1069 while the last 2 SNPs from 5% to 95%. For the i th individual, we sampled $G_{ijq} \sim \text{Bin}(1, F_j)$ for the q th
1070 genome of the j th SNP, and took the average value of two genomes as the genotype for the j th SNP of
1071 the i th individual: $G_{ij} = \frac{G_{ij1} + G_{ij2}}{2}$. Based on the genotype data, we obtained the haplotype data for
1072 each individual, and we considered removing haplotypes rarer than 0.1% or not when rare SNPs were
1073 generated. In addition, we sampled 20 fixed covariates (including sex, age and 18 PCs) C_{ic} where $c =$
1074 $1, \dots, 20$ from UK Biobank for 100000 individuals.

1075

1076 Next, we sampled the effect sizes of SNPs β_{G_j} and covariates β_{C_c} , and standardize them by their

1077 standard deviations: $\beta_{G_j} \sim \frac{U(-1,1)}{sd(G_j)}$ and $\beta_{C_c} \sim \frac{U(-1,1)}{sd(C_c)}$ for each fixed j and c , respectively. When

1078 interaction exists, we created a fixed effect size for haplotype "11XX" as twice the average absolute

1079 SNP effects: $\beta_{H_1} = \frac{1}{2} \sum_{j=1}^4 |\beta_{G_j}|$ where H_1 refers to "11XX"; otherwise, $H_1 = 0$. Note that $F_{H_1} =$

1080 0.09% when rare SNPs are included.

1081

1082 Finally, we sampled the outcome based on the outcome score (for the i th individual)

$$1083 \quad O_i = \sum_{c=1}^{20} \beta_c C_{ic} + \gamma (\sum_{j=1}^4 \beta_{G_j} G_{ij} + \beta_{H_1} H_1) + e_i + w,$$

1084 where γ is the effect scale of SNPs and haplotype "11XX", $e_i \sim N(0,0.1)$ is the random error and w is
1085 a fixed intercept term. For linear models, the outcome $Y_i = O_i$; while for generalized linear models,

1086 we sampled the outcome from binomial distribution: $Y_i \sim \text{Bin}(1, \pi_i)$, where $\pi_i = \frac{e^{O_i}}{1+e^{O_i}}$ is the

1087 probability that the i th individual has the case.

1088

1089 As the simulation is intended to compare the variance explained by HTRX, HTR and SNPs (GWAS)
1090 in addition to fixed covariates, we tripled the effect sizes of SNPs and haplotype "11XX" (if
1091 interaction exists) by setting $\gamma = 3$. In "glm", to ensure a reasonable case prevalence (e.g. below 5%),
1092 we set $w = -7$, which was also applied in "lm".

1093

1094 We applied the procedure described in “**HTRX Model selection procedure for shorter haplotypes**”
1095 for HTRX, HTR and GWAS, and visualized the distribution of the out-of-sample R^2 for each of the
1096 best models selected by each method in Supplementary Figure 4.4. In both "lm" and "glm", HTRX
1097 has equal predictive performance as the true model. It performs as well as GWAS when the
1098 interaction effects is absent, explains more variance when an interaction is present, and is significantly
1099 more explanatory than HTR. When rare SNPs are included, the only effective interaction term is rare.
1100 In this case the difference between GWAS and HTRX becomes smaller as expected, and removing the
1101 rare haplotypes hardly reduces the performance of HTRX.

1102

1103 In conclusion, we demonstrate through simulation that our HTRX implementation a) searches
1104 haplotype space effectively, and b) protects against overfitting. This makes it a superior approach
1105 compared to HTR and GWAS to integration SNP effects with gene-gene interaction. Its robustness
1106 also retains when there are rare effective SNPs and haplotypes.

1107

1108 **POLYGENIC SELECTION TEST**

1109 We inferred allele frequency trajectories and selection coefficients for a set of LD-puned genome-
1110 wide significant trait associated variants using a modified version of the software CLUES¹⁹. To
1111 account for population structure in our samples, we applied a novel chromosome painting technique
1112 based on inference of a sample's nearest neighbours in the marginal trees of an ARG that contains
1113 labelled individuals (Irving-Pease et al., *submitted*). We ran CLUES using a time-series of imputed
1114 aDNA genotype probabilities obtained from 1,015 ancient West Eurasian samples that passed all
1115 quality control filters. We produced four additional models for each trait associated variant, by
1116 conditioning the analysis on one of the four ancestral path labels from our chromosome painting
1117 model: either Western hunter-gatherers (WHG), Eastern hunter-gatherers (EHG), Caucasus hunter-
1118 gatherers (CHG), or Anatolian farmers (ANA). We then inferred polygenic selection gradients (ω)
1119 and p-values for each of MS, CD and RA, in all ancestral paths, using the software PALM²⁰. Full
1120 methods and results can be found in Supplementary Note 6.

1121

1122 **ANCESTRY LINKAGE DISEQUILIBRIUM (LDA) AND ANCESTRY LINKAGE 1123 DISEQUILIBRIUM SCORE (LDAS)**

1124 In population genetics, linkage disequilibrium (LD) is defined as the non-random association of
1125 alleles at different loci in a given population⁸⁰. We propose an ancestry linkage disequilibrium (LDA)
1126 approach to measure the association of ancestries between SNPs.

1127

1128 Let $A(i, j, k)$ denote the probability of the k th ancestry ($k = 1, \dots, K$) at the j th SNP ($j = 1, \dots, J$) of a
 1129 chromosome for the i th individual ($i = 1, \dots, N$).

1130

1131 We define the distance between SNP l and m as the average L_2 norm between ancestries at those
 1132 SNPs. Specifically we compute the L_2 norm for the i th genome as

$$1133 \quad D_i(l, m) = \|A(i, l, \cdot) - A(i, m, \cdot)\|_2 = \sqrt{\frac{1}{K} \sum_{k=1}^K (A(i, l, k) - A(i, m, k))^2}.$$

1134

1135 Then we compute the distance between SNP l and m by averaging $D_i(l, m)$:

$$1136 \quad D(l, m) = \frac{1}{N} \sum_{i=1}^N D_i(l, m).$$

1137

1138 We define $D^*(l, m)$ as the theoretical distance between SNP l and m if there were no linkage
 1139 disequilibrium of ancestry (LDA) between them. $D^*(l, m)$ is estimated by

$$1140 \quad D^*(l, m) \approx \frac{1}{N} \sum_{i=1}^N \|A(i^*, l, \cdot) - A(i, m, \cdot)\|_2,$$

1141 where $i^* \in \{1, \dots, N\}$ are re-sampled without replacement at SNP l . Using the empirical distribution
 1142 of ancestry probabilities accounts for variability in both the average ancestry and its distribution
 1143 across SNPs. Ancestry assignment can be very precise in regions of the genome where our reference
 1144 panel matches our data, and uncertain in others where we only have distant relatives of the underlying
 1145 populations.

1146

1147 The LDA between SNP l and m is a similarity, defined in terms of the negative distance $-D(l, m)$
 1148 normalized by the expected value $D^*(l, m)$ under no LD, as:

$$1149 \quad LDA(l, m) = \frac{D^*(l, m) - D(l, m)}{D^*(l, m)}.$$

1150

1151 LDA therefore takes an expected value 0 when haplotypes are randomly assigned at different SNPs,
 1152 and positive values when the ancestries of haplotypes are correlated.

1153

1154 LDA is a pairwise quantity. To arrive at a per-SNP property, we define the LDA score (LDAS) of
 1155 SNP j as the total LDA of this SNP with the rest of the genome, i.e. the integral of the LDA for that
 1156 SNP. Because this quantity decreases to zero as we move away from the target SNP, this is in practice
 1157 computed within an X cM-window (we use $X = 5$ as LDA is approximately zero outside this region in
 1158 our data) on both sides of the SNP. Note that we measure this quantity in terms of the genetic
 1159 distance, and therefore LDAS is measuring the length of ancestry-specific haplotypes compared to
 1160 individual-level recombination rates.

1161

1162 As a technical note, when the SNPs approach either end of the chromosome, they no longer have a
 1163 complete window, which results in a smaller LDAS. This would be appropriate for measuring total
 1164 ancestry correlations, but to make LDAS useful for detecting anomalous SNPs, we use the LDAS of
 1165 the symmetric side of the SNP to estimate the LDAS within the non-existent window.

1166

$$1167 \quad LDAS(j; X) = \begin{cases} \int_{gd(j)-X}^{gd(j)+X} LDA(j, l) dgd & \text{if } X \leq gd(j) \leq tg - X, \\ \int_0^{gd(j)+X} LDA(j, l) dgd + \int_{2gd(j)}^{gd(j)+X} LDA(j, l) dgd & \text{if } gd(j) < X, \\ \int_{gd(j)-X}^{tg} LDA(j, l) dgd + \int_{gd(j)-X}^{2gd(j)-tg} LDA(j, l) dgd & \text{if } gd(j) > tg - X. \end{cases}$$

1167

1168

1169 where $gd(l)$ is the genetic distance (i.e. position in cM) of SNP l , and tg is the total genetic distance
1170 of a chromosome. We also assume the LDA on either end of the chromosome equals the LDA of the
1171 SNP closest to the end: $LDA(j, gd = 0) = LDA(j, l_{min(gd)})$ and $LDA(j, gd = td) =$
1172 $LDA(j, l_{max(gd)})$, where gd is the genetic distance, $l_{min(gd)}$ and $l_{max(gd)}$ are the indexes of the SNP
1173 with the smallest and largest genetic distance, respectively.

1174

1175 The integral $\int_{gd(j)-x}^{gd(j)+x} LDA(j, l) dgd$ is computed assuming linear interpolation of the LDA score
1176 between adjacent SNPs.

1177

1178 LDA thus quantifies the correlations between the ancestry of two SNPs, measuring the proportion of
1179 individuals who have experienced a recombination leading to a change in ancestry, relative to the
1180 genome-wide baseline. The LDA score is the total amount of genome in LDA with each SNP
1181 (measured in recombination map distance).

1182

1183 **SIMULATION FOR SELECTION: LDA**

1184 An ancient population P_0 evolved for 2200 generations before splitting into two sub-populations P_1
1185 (Steppe) and P_2 (Farmer). After evolving 400 generations, we added mutation “ m_1 ” and “ m_2 ” at the
1186 different locus in P_1 and P_2 . Both added mutations were then positively selected in the following 300
1187 generations, after which they merged to P_3 , where both added mutations experienced strong positive
1188 selection for 20 generations. Finally, we sampled 1000 individuals from P_3 to compute their ancestry
1189 proportions of P_1 and P_2 using the "chromosome painting" technique, and calculated the LDA score
1190 of the simulated chromosome positions.

1191

1192 The above describes the simulation in Supplementary Figure 6.1.

1193

1194 We investigated balancing selection at 2 loci as well. The balancing selection in P_1 and P_2 ensured the
1195 mutated allele reaches around 50% frequency, while positive selection made the mutated allele
1196 become almost the only allele. In P_3 , if m_1 or m_2 was positively selected, its frequency reached more
1197 than 80% regardless of whether the allele experienced balancing or positive selection in P_1 or P_2 ,
1198 because we set a strong positive selection. If m_1 or m_2 was balancing selected in P_3 , its frequency
1199 slightly increased, e.g. if m_1 underwent balancing selection in P_1 , it had 25% frequency when P_3 was
1200 created, and the frequency reached around 37.5% after 20 generations of balancing selection in P_3 .

1201

1202 The results (Supplementary Figure 6.2) show that positive selection in P_3 resulted in low LDA scores
1203 around the selected locus, if this allele was not uncommon (i.e. had 50% (balancing selection) or
1204 100% frequency (positive selection) in subpopulation P_1 or P_2). Note that the balancing selection in
1205 P_1 or P_2 worked the same as “weak positive selection”, because m_1 and m_2 were rare when they first
1206 occurred, which were positively selected until 50% frequency.

1207

1208 We also performed simulations for selection at a single locus (Supplementary Figure 6.2&6.3).

1209

1210 Stage 1: We added a mutation m_1 in the 1600 generation in P_0 , which then underwent balancing
1211 selection until generation 2200, when P_0 split into P_1 and P_2 , where the frequency of m_1 was around
1212 50%.

1213

1214 Stage 2: Then we explored different combinations of positive, balancing and negative selection of m_1
1215 in P_1 and P_2 . the frequency of m_1 reached 80%, 50% and 20% when it was positively, balancing or
1216 negatively selected, respectively, until generation 2899. Here we sampled 20 individuals each in P_1
1217 and P_2 as the ancient samples.

1218

1219 Stage 3: Then P_1 and P_2 merged into P_3 in generation 2900. In P_3 , for each combination of selection
1220 in Stage 2, we simulated positive, balancing and negative selection for m_1 . The selection lasted for 20
1221 generations, and then we sampled 4000 individuals from P_3 as the modern population.

1222

1223 Results: when m_1 was positively selected in only one of P_1 and P_2 , and it experienced negative
1224 selection in P_3 , the LDA scores around the locus of m_1 were low. Otherwise, no abnormal LDA
1225 scores were found at m_1 .

1226

1227

1228

1229 REFERENCES

1230

1231 1. Walton, C. *et al.* Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS,

1232 third edition. *Mult. Scler. J.* **26**, 1816–1821 (2020).

1233 2. Attfield, K. E., Jensen, L. T., Kaufmann, M., Friese, M. A. & Fugger, L. The immunology of

1234 multiple sclerosis. *Nat. Rev. Immunol.* (2022) doi:10.1038/s41577-022-00718-z.

1235 3. International Multiple Sclerosis Genetics Consortium *et al.* Multiple sclerosis genomic map

1236 implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).

1237 4. Bjornevik, K. *et al.* Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated

1238 with multiple sclerosis. *Science* **375**, 296–301 (2022).

1239 5. Lanz, T. V. *et al.* Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and

1240 GlialCAM. *Nature* **603**, 321–327 (2022).

1241 6. Olsson, T., Barcellos, L. F. & Alfredsson, L. Interactions between genetic, lifestyle and

1242 environmental risk factors for multiple sclerosis. *Nat. Rev. Neurol.* **13**, 25–36 (2017).

1243 7. Benton, M. L. *et al.* The influence of evolutionary history on human health and disease. *Nat. Rev.*

1244 *Genet.* **22**, 269–283 (2021).

1245 8. Chi, C. *et al.* Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic

1246 ancestry. *PLOS Genet.* **15**, e1007808 (2019).

1247 9. Allentoft, M. E. *et al.* *Population Genomics of Stone Age Eurasia.*

1248 <http://biorxiv.org/lookup/doi/10.1101/2022.05.04.490594> (2022) doi:10.1101/2022.05.04.490594.

1249 10. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages

1250 in Europe. *Nature* **522**, 207–211 (2015).

1251 11. Jones, E. R. *et al.* Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat.*

1252 *Commun.* **6**, 8912 (2015).

1253 12. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172

1254 (2015).

1255 13. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature*

1256 **562**, 203–209 (2018).

1257 14. Itan, Y., Powell, A., Beaumont, M. A., Burger, J. & Thomas, M. G. The Origins of Lactase

- 1258 Persistence in Europe. *PLoS Comput. Biol.* **5**, e1000491 (2009).
- 1259 15. Dehasque, M. *et al.* Inference of natural selection from ancient DNA. *Evol. Lett.* **4**, 94–108
1260 (2020).
- 1261 16. Efron, B. Better Bootstrap Confidence Intervals. *J. Am. Stat. Assoc.* **82**, 171–185 (1987).
- 1262 17. Zaykin, D. V. *et al.* Testing Association of Statistically Inferred Haplotypes with Discrete and
1263 Continuous Traits in Samples of Unrelated Individuals. *Hum. Hered.* **53**, 79–91 (2002).
- 1264 18. Thuesen, N. H., Klausen, M. S., Gopalakrishnan, S., Trolle, T. & Renaud, G. *Benchmarking*
1265 *freely available human leukocyte antigen typing algorithms across varying genes, coverages and*
1266 *typing resolutions*. <http://biorxiv.org/lookup/doi/10.1101/2022.06.28.497888> (2022)
1267 doi:10.1101/2022.06.28.497888.
- 1268 19. Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for inferring
1269 selection and allele frequency trajectories from DNA sequence data. *PLOS Genet.* **15**, e1008384
1270 (2019).
- 1271 20. Stern, A. J., Speidel, L., Zaitlen, N. A. & Nielsen, R. Disentangling selection on genetically
1272 correlated polygenic traits via whole-genome genealogies. *Am. J. Hum. Genet.* **108**, 219–239
1273 (2021).
- 1274 21. Comabella, M. *et al.* Identification of a Novel Risk Locus for Multiple Sclerosis at 13q31.3
1275 by a Pooled Genome-Wide Scan of 500,000 Single Nucleotide Polymorphisms. *PLoS ONE* **3**,
1276 e3490 (2008).
- 1277 22. Bersaglieri, T. *et al.* Genetic Signatures of Strong Recent Positive Selection at the Lactase
1278 Gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
- 1279 23. He, Z., Dai, X., Beaumont, M. & Yu, F. Detecting and Quantifying Natural Selection at Two
1280 Linked Loci from Time Series Data of Allele Frequencies with Forward-in-Time Simulations.
1281 *Genetics* **216**, 521–541 (2020).
- 1282 24. Fugger, L., Jensen, L. T. & Rossjohn, J. Challenges, Progress, and Prospects of Developing
1283 Therapies to Treat Autoimmune Diseases. *Cell* **181**, 63–80 (2020).
- 1284 25. Gregersen, J. W. *et al.* Functional epistasis on a common MHC haplotype associated with
1285 multiple sclerosis. *Nature* **443**, 574–577 (2006).

- 1286 26. Wang, J. H. *et al.* Modeling the cumulative genetic risk for multiple sclerosis from genome-
1287 wide association data. *Genome Med.* **3**, 3 (2011).
- 1288 27. Cotsapas, C. & Mitrovic, M. Genome-wide association studies of multiple sclerosis. *Clin.*
1289 *Transl. Immunol.* **7**, e1018 (2018).
- 1290 28. Slim, L., Chatelain, C., Foucauld, H. de & Azencott, C.-A. A systematic analysis of gene-
1291 gene interaction in multiple sclerosis. *BMC Med. Genomics* **15**, 100 (2022).
- 1292 29. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World
1293 human tuberculosis. *Nature* **514**, 494–497 (2014).
- 1294 30. Sabin, S. *et al.* A seventeenth-century Mycobacterium tuberculosis genome supports a
1295 Neolithic emergence of the Mycobacterium tuberculosis complex. *Genome Biol.* **21**, 201 (2020).
- 1296 31. Rasmussen, S. *et al.* Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago.
1297 *Cell* **163**, 571–582 (2015).
- 1298 32. Spyrou, M. A. *et al.* Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age
1299 origin for bubonic plague. *Nat. Commun.* **9**, 2234 (2018).
- 1300 33. Rascovan, N. *et al.* Emergence and Spread of Basal Lineages of Yersinia pestis during the
1301 Neolithic Decline. *Cell* **176**, 295-305.e10 (2019).
- 1302 34. Guellil, M. *et al.* Ancient herpes simplex 1 genomes reveal recent viral structure in Eurasia.
1303 *Sci. Adv.* **8**, eabo4435.
- 1304 35. Pontremoli, C., Forni, D., Clerici, M., Cagliani, R. & Sironi, M. Possible European Origin of
1305 Circulating Varicella Zoster Virus Strains. *J. Infect. Dis.* jiz227 (2019) doi:10.1093/infdis/jiz227.
- 1306 36. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify
1307 susceptibility loci for multiple common infections. *Nat. Commun.* **8**, 599 (2017).
- 1308 37. Krause-Kyora, B. *et al.* Ancient DNA study reveals HLA susceptibility locus for leprosy in
1309 medieval Europeans. *Nat. Commun.* **9**, 1569 (2018).
- 1310 38. Wallin, M. T. *et al.* The prevalence of MS in the United States: A population-based estimate
1311 using health claims data. *Neurology* **92**, e1029–e1040 (2019).
- 1312 39. Feigin, V. L. *et al.* Global, regional, and national burden of neurological disorders, 1990–
1313 2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **18**, 459–

- 1314 480 (2019).
- 1315 40. Fleming, J. & Fabry, Z. The hygiene hypothesis and multiple sclerosis. *Ann. Neurol.* **61**, 85–
1316 89 (2007).
- 1317 41. Joo, Y. B., Lim, Y.-H., Kim, K.-J., Park, K.-S. & Park, Y.-J. Respiratory viral infections and
1318 the risk of rheumatoid arthritis. *Arthritis Res. Ther.* **21**, 199 (2019).
- 1319 42. Safiri, S. *et al.* Global, regional and national burden of rheumatoid arthritis 1990–2017: a
1320 systematic analysis of the Global Burden of Disease study 2017. *Ann. Rheum. Dis.* **78**, 1463–1471
1321 (2019).
- 1322 43. Lindfors, K. *et al.* Coeliac disease. *Nat. Rev. Dis. Primer* **5**, 3 (2019).
- 1323 44. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and
1324 imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126
1325 (2021).
- 1326 45. Meyer, M. & Kircher, M. Illumina Sequencing Library Preparation for Highly Multiplexed
1327 Target Capture and Sequencing. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5448 (2010).
- 1328 46. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming,
1329 identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
- 1330 47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
1331 *Bioinformatics* **25**, 1754–1760 (2009).
- 1332 48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
1333 2079 (2009).
- 1334 49. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast
1335 approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–
1336 1684 (2013).
- 1337 50. Fu, Q. *et al.* A Revised Timescale for Human Evolution Based on Ancient Mitochondrial
1338 Genomes. *Curr. Biol.* **23**, 553–559 (2013).
- 1339 51. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation
1340 Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).
- 1341 52. Monroy Kuhn, J. M., Jakobsson, M. & Günther, T. Estimating genetic kin relationships in

- 1342 prehistoric populations. *PLOS ONE* **13**, e0195491 (2018).
- 1343 53. Weissensteiner, H. *et al.* HaploGrep 2: mitochondrial haplogroup classification in the era of
1344 high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).
- 1345 54. Scorrano, G., Yediay, F. E., Pinotti, T., Feizabadifarahani, M. & Kristiansen, K. The genetic
1346 and cultural impact of the Steppe migration into Europe. *Ann. Hum. Biol.* **48**, 223–233 (2021).
- 1347 55. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based
1348 Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- 1349 56. Shringarpure, S. S., Bustamante, C. D., Lange, K. & Alexander, D. H. Efficient analysis of
1350 large datasets and sex bias with ADMIXTURE. *BMC Bioinformatics* **17**, 218 (2016).
- 1351 57. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192**, 1065–1093 (2012).
- 1352 58. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of Population Structure using
1353 Dense Haplotype Data. *PLoS Genet.* **8**, e1002453 (2012).
- 1354 59. Margaryan, A. *et al.* Population genomics of the Viking world. *Nature* **585**, 390–396 (2020).
- 1355 60. Hellenthal, G. *et al.* A Genetic Atlas of Human Admixture History. *Science* **343**, 747–751
1356 (2014).
- 1357 61. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
1358 *Nature* **526**, 68–74 (2015).
- 1359 62. Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: An R Package for Rapidly
1360 Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Front. Genet.* **11**, 157
1361 (2020).
- 1362 63. Ishigaki, K. *et al.* *Trans-ancestry genome-wide association study identifies novel genetic*
1363 *mechanisms in rheumatoid arthritis*. <http://medrxiv.org/lookup/doi/10.1101/2021.12.01.21267132>
1364 (2021) doi:10.1101/2021.12.01.21267132.
- 1365 64. Alekseyenko, A. V. *et al.* Causal graph-based analysis of genome-wide association data in
1366 rheumatoid arthritis. *Biol. Direct* **6**, 25 (2011).
- 1367 65. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the
1368 association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).
- 1369 66. Spanish Consortium on the Genetics of Coeliac Disease (CEGEC) *et al.* Dense genotyping

- 1370 identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat.*
1371 *Genet.* **43**, 1193–1201 (2011).
- 1372 67. Monsuur, A. J. *et al.* Effective Detection of Human Leukocyte Antigen Risk Alleles in Celiac
1373 Disease Using Tag Single Nucleotide Polymorphisms. *PLoS ONE* **3**, e2270 (2008).
- 1374 68. Gutierrez-Achury, J. *et al.* Fine mapping in the MHC region accounts for 18% additional
1375 genetic risk for celiac disease. *Nat. Genet.* **47**, 577–578 (2015).
- 1376 69. the RACI consortium *et al.* Genetics of rheumatoid arthritis contributes to biology and drug
1377 discovery. *Nature* **506**, 376–381 (2014).
- 1378 70. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
1379 datasets. *GigaScience* **4**, 7 (2015).
- 1380 71. Ju, D. & Mathieson, I. The evolution of skin pigmentation-associated variation in West
1381 Eurasia. *Proc. Natl. Acad. Sci.* **118**, e2009227118 (2021).
- 1382 72. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package for
1383 gene list functional enrichment analysis and namespace conversion toolset g:Profiler.
1384 *F1000Research* **9**, ELIXIR-709 (2020).
- 1385 73. Thorndike, R. L. Who belongs in the family? *Psychometrika* **18**, 267–276 (1953).
- 1386 74. Berg, J. J. & Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet.* **10**,
1387 e1004412 (2014).
- 1388 75. Frangos, C. C. & Schucany, W. R. Jackknife estimation of the bootstrap acceleration
1389 constant. *Comput. Stat. Data Anal.* **9**, 271–281 (1990).
- 1390 76. Sarmanova, A., Morris, T. & Lawson, D. J. *Population stratification in GWAS meta-analysis*
1391 *should be standardized to the best available reference datasets.*
1392 <http://biorxiv.org/lookup/doi/10.1101/2020.09.03.281568> (2020) doi:10.1101/2020.09.03.281568.
- 1393 77. McFadden, D. Conditional logit analysis of qualitative choice behavior. (1973).
- 1394 78. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **7**, 1–26 (1979).
- 1395 79. Kass, R. E. & Wasserman, L. A Reference Bayesian Test for Nested Hypotheses and its
1396 Relationship to the Schwarz Criterion. *J. Am. Stat. Assoc.* **90**, 928–934 (1995).
- 1397 80. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the

1398 medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).

1399

