1 Long runs of homozygosity are correlated with marriage 2 preferences across global population samples

3 Samali Anova Sahoo¹, Arslan A. Zaidi^{1,*}, Santosh Anagol² & Iain Mathieson^{1,*}

4 March 4, 2021

5 1. Department of Genetics, Perelman School of Medicine, University of Pennsylvania

6 2. Business Economics & Public Policy, Wharton School of Business, University of Pennsylvania

7 * Correspondence to AZ (aazaidi@pennmedicine.upenn.edu) or IM (mathi@pennmedicine.upenn.edu)

8 **Abstract**

9 Children of consanguineous unions carry long runs of homozygosity (ROH) in their
10 genomes, due to their parents' recent shared ancestry. This increases the burden of
11 recessive disease in populations with high levels of consanguinity and has been heavily
12 studied in some groups. However, there has been little investigation of the broader ef-
13 fect of consanguinity on patterns of genetic variation on a global scale. Here, we collect
14 published genetic data and information about marriage practices from 396 worldwide
15 populations and show that preference for cousin marriage has a detectable effect on
16 the distribution of long ROH in these samples, increasing the expected number of
17 ROH longer than 10Mb by a factor of 1.5 ($P=2.3 \times 10^{-4}$). Variation in marriage
18 practice and consequent rates of consanguinity is therefore an important aspect of de-
19 mographic history for the purposes of modeling human genetic variation. However,
20 marriage practices explain a relatively small proportion of the variation in ROH dis-
21 tribution and consequently the ability to predict marriage practices from population
22 genetic samples (for example of ancient populations) is limited.

23 Introduction

24 Marriage practices have consequences for human genetic variation. One extensively debated
 25 and regulated practice is consanguinity – the union between closely related individuals. An
 26 estimated 10.4 percent of the world’s population is made up of couples who are second-
 27 cousins or closer and their offspring [1] (see [2] and consang.net for additional estimates of
 28 worldwide consanguinity). The children of consanguineous marriages have longer runs of
 29 homozygosity [3] and therefore carry a greater recessive disease burden [4] equivalent to an
 30 increase in child mortality or severe disease of approximately 3-4% [5–7]. It is, however,
 31 unclear to what extent ethnographic assessments (either via surveys or direct observation)
 32 of marriage practices at the population level translate to the genetic signature associated
 33 with cousin marriage.

34 Although actual levels of consanguinity must be correlated with patterns of genetic vari-
 35 ation, there are three reasons why ethnographic assessments of cousin marriage prevalence
 36 might not. First, different cultures’ understanding of “cousin” marriage could cover different
 37 concepts. For example, some might have a clear distinction between first and second cousin
 38 marriage, while others consider all of these types the same. Second, cultural preferences for
 39 marriage may change rapidly. Ethnographic data may be outdated, or preferences may not
 40 persist long enough to have a detectable effect on patterns of genetic variation diagnostic
 41 of ongoing consanguinity. Finally, genomic signatures of consanguinity may be obscured by
 42 other demographic factors such as endogamy.

43 For these reasons, populations that practice cousin marriage may in fact demonstrate
 44 relatively few of the negative genetic consequences associated with increased homozygos-
 45 ity. For example, many marriages might occur between “cousins” who are actually quite
 46 distantly related. Alternatively, in many regions of the world, cousin marriage co-occurs
 47 with endogamy (marriage within defined sub-groups such as castes) [8]. Since endogamy
 48 can itself lead to excess recessive disease burden [9], the marginal effects of consanguinity
 49 might be small. Finally, even though both endogamy and consanguinity, by increasing

ROH, increase the risk of recessive disease in the short term, the same process exposes such alleles to selection, purging them more effectively from the population in the long term.

Most previous work on the relationship between consanguinity and genetic variation has focused on a limited number of populations. McQuillan et al. 2008 correlated pedigree based measures of marriage practices to ROH in the Orcadian population [10]. Kang et. al. (2016) found a correlation between long ROH in recent individuals and inbreeding coefficients for nine Jewish populations, where rates of consanguinity were based on survey data [11]. Arciero et al. (2020) found a strong correlation between survey measured consanguinity and long ROH in a British Pakistani cohort [12]. In the most similar study to ours [13], Pemberton et al. (2014) found a correlation of .349 between inbreeding coefficients estimated from genetic data and survey-based consanguinity data. However, this analysis was based on only 26 populations and did not correct for correlations among populations due to shared demographic history.

We therefore set out to test whether a cultural preference for cousin marriage is detectable in genetic data from a worldwide sample of 3,859 individuals from 396 populations. We categorize these populations into those prohibiting (43%), allowing (12%), and preferring (44%) cousin marriage using publicly available ethnographic sources. We find that ethnographic measures of preferences for cousin marriage are detectably related to the distribution of long ROH, increasing the expected number of ROH longer than 10Mb by $1.5\times$ in populations preferring cousin marriages over those that prohibit it ($P = 2.31 \times 10^{-4}$), after controlling for ten principal components of genetic variation, shorter runs of homozygosity, and heterozygosity (collectively serving as proxies for demographic events such as bottlenecks, admixture and endogamy). This effect corresponds to an approximate increase of 0.25 (in proportion of unions that are cousins) in populations where ethnographic data indicated a preference for the practice. Further, we find that populations preferring consanguinity show greater long ROH when compared to geographically and genetically "close" consanguinity-prohibiting populations in a matched-pair design.

77 Methods

78 Simulations

79 First, we carried out simulations to investigate the effect of degree (proportion of cousin
80 unions) and duration (number of generations) of consanguinity on the distribution of runs of
81 homozygosity. We also used these simulations to calibrate the difference in consanguinity
82 between populations that we classified as prohibiting and preferring cousin unions (see
83 below). We used cousin-sim [14] to generate pedigrees, each of which was initialized with
84 500 unrelated founders and 500 individuals in subsequent generations for either 10 or 50
85 generations. In each generation, ($c \in \{0, 0.25, 0.5, 0.75\}$) of all unions were between cousins,
86 of which 62% were first-, 23% were second-, and 15% were third-degree cousins [15]. We
87 used the pedigree as input to *ped-sim* [16] with a Poisson crossover model and a refined
88 genetic map of the human genome from [17] to simulate chromosomal segments and retained
89 only those that were identical by descent (IBD) within an individual. We refer to these
90 segments as runs of homozygosity (ROH) to distinguish them from inter-individual IBD.

We also calculated the expected number and total length of long runs of homozygosity analytically using the model described in [18]. Briefly, there are $m = 2n + 4$ meioses separating the two chromosomes of an offspring of n^{th} cousins. For such an offspring, the number of segments of length x in the i^{th} chromosome of length l_i (in Morgans) has the following density (Equation S8 of [18]):

$$f(x) = \frac{4}{2^m} \{ (l_i - x)m^2 e^{-xm} + 2me^{-xm} \} \quad (1)$$

91 The integrals $\int_{0.1}^{l_i} f(x)dx$ and $\int_{0.1}^{l_i} xf(x)dx$ yield the expected number and total length,
92 respectively, of long ($>10\text{cM}$ or 0.1M) runs of homozygosity in the i^{th} chromosome. This
93 genetic length corresponds to approximately 10Mb physical distance, on average across
94 the genome. We then computed the expected autosome-wide number and total length as
95 $\sum_{i=1}^{22} \int_{0.1}^{l_i} f(x)dx$ and $\sum_{i=1}^{22} \int_{0.1}^{l_i} xf(x)dx$, respectively, using the sex-averaged genetic map
96 from [17] for chromosomal lengths (in Morgans).

97 Genetic data collection and processing

98 We collected genotype data generated with the Affymetrix Human Origins array from 6
 99 papers [9, 19–23], and merged this with whole-genome sequence data from three further
 100 papers [24–26]. We merged all datasets and removed 1) apparent duplicate samples (iden-
 101 tified using *plink --genome*) 2) samples marked as questionable in the original publications
 102 3) samples where the population label was too generic to be useful. We retained the popu-
 103 lation labels from the publications except where there was an obvious typo or synonymous
 104 population. Our genetic dataset contained 4544 individuals from 488 populations (median
 105 number of individuals per-population 7, range 1-71). We restricted to autosomes, leaving
 106 469,421 SNPs.

107 For each individual, we calculated heterozygosity (at genotyped SNPs) and the number
 108 of ROH greater than 1,2,5 and 10 Mb (using the *--het* and *--homozy* commands in *plink*).
 109 We then subtracted the number of ROH so that our variable NROH1 represents the number
 110 of ROH longer than 1 but smaller than 2 Mb, NROH5 representing the number of ROH
 111 longer than 5 but smaller than 10 Mb, and so on. Finally, we computed principal compo-
 112 nents of the merged dataset, and also separately of individuals belonging to the 237 South
 113 Asian populations. We used *plink* v1.90 [27] and *smartpca* v16000 [28] for all analyses.

114 Ethnographic data collection and processing

115 We collected data on marriage practices from the following sources:

- 116 • The electronic Human Relations Area Files (eHRAF) World Cultures database [29],
 117 which catalogs a large set of ethnographic writing describing cultural and social as-
 118 pects of different groups.
- 119 • The Ethnographic Atlas [30], which describes the cultural practices of 1,291 societies.
- 120 • Focused online searches for marital practices in specific groups. Sources include
 121 Google Scholar, Google books, as well as other papers and books. To search for

specific populations, we searched for terms such as ‘marriage’, ‘cousin marriage’, ‘endogamy’, ‘exogamy’, ‘consanguinity’, and ‘consanguineous’ along with the name of the population.

We encoded the level of consanguinity in a population categorically as ‘prohibited’ (coded as 0), ‘permitted’ (1), or ‘preferred’ (2). When quantitative measures (e.g. prevalence of cousin unions in the population) were available, we designated populations to the above three categories based on the percent of unions between first cousins and first cousins once removed. In such cases, a prevalence of 0 to 2.5% was encoded as prohibited, 2.5% to 20% was encoded as ‘permitted’, and 20% and above was encoded as ‘preferred’. When quantitative measures were not available, we analyzed ethnographic records for terms indicating the preference for cousin unions. Groups that stated cousin marriage was “common”, “practiced”, “encouraged” or equivalent were classified as ‘preferred’ category. Meanwhile, groups where cousin marriage was described as being “allowed”, “occasional”, “present but uncommon” or equivalent were coded as ‘permitted’. Lastly, groups where cousin unions were “forbidden”, “barred” or marriage was “exogamous” were classified as ‘prohibited’. We classified groups where information was unclear or unavailable as ‘missing’. We gathered marriage practice information for 522 populations, which was reduced to 396 (and 3,859 individuals) after merging with the genetic data.

To validate our consanguinity scoring, we tested whether our assignments were correlated with individual survey data on marriage practices from the 2011 India Health and Development Survey (IHDS) [31]. We matched populations of individuals from the IHDS survey to population names in our genetics data using caste names provided in the IHDS data (N = 198 for the matched sample of Indian populations). Our consanguinity assignments are strongly correlated with the proportion of IHDS respondents answering yes to the following prompt: “Now, I would like to ask you some questions about marriage customs in your community (jati) for a family like yours. Is it permissible to marry a girl to her cousin?” (Fig. 1).

Figure 1: Our consanguinity score corresponds well with self-reports on whether cousin marriage is permitted in an individual's community. The x-axis shows the consanguinity score assigned to each population and the y-axis shows the proportion of survey participants in the 2011 Indian Health and Development survey (IHDS) who responded that in their community it was permissible for a girl to marry their cousin. Each point is a population (matched using the caste name variable from the IHDS survey, $N = 198$) and red points represent the median for that category.

149 Effect of consanguinity on long runs of homozygosity

150 We used Poisson regression to model the effect of consanguinity on the number of long
151 ($>10\text{cM}$) runs of homozygosity (NROH). The consanguinity score of the population was
152 treated as a categorical variable with three levels ('prohibited' - 0, 'permitted' - 1, and
153 'preferred' - 2). In all models, we included genome-wide heterozygosity, ten genetic principal
154 components (gPC 1-10), and the number of shorter (1-2cM and 2-5cM) runs of homozygosity
155 (hereafter NROH1 and NROH2, respectively) as covariates to correct for other demographic
156 events such as population bottlenecks.

157 We fitted models where each population was an observational unit as well as models
158 where each individual was treated as an observational unit. When each population was
159 treated as an observational unit, we used the mean value of the variable (e.g. NROH and
160 gPCs) in the population and fitted both weighted (weights proportional to the sample size
161 from each population) and unweighted models. When each individual was treated as the
162 observational unit, we used individuals' values of NROH and gPC and population-wide
163 consanguinity scores. Finally, because our densest sampling is from South Asia, we also
164 fitted models restricting to populations from this region.

165 We evaluated model fit based on the dispersion parameter and the distribution of stan-
166 dardized residuals. In the main text, we present results from the Poisson model, which
167 assumes that the mean of the residuals is equal to the variance. Indeed, the residuals
168 showed no evidence of over-dispersion (Pearson statistic = 259.62, p-value under χ^2 with
169 378 degrees of freedom ≈ 1).

170 Paired analysis

171 To further ensure that the relationship between consanguinity and NROH10 was not con-
172 founded by long-term demographic history, we identified population pairs that were matched
173 genetically and geographically but were different in that cousin unions were preferred in one
174 population and prohibited in the other. To maximize genetic similarity and geographic prox-

175 imity in selecting population pairs, we calculated genetic (Euclidean) distance using PC1-10
 176 and geographic distances (using longitude and latitude) between populations that prefer
 177 consanguinity (score = 2) and those that either prohibit it or permit it (score $\in \{0, 1\}$).
 178 Then, we divided each distance matrix by its median (to account for differences in scale)
 179 and averaged the two to calculate a single distance matrix. We selected population pairs in
 180 ascending order of the distance between them such that each population was part of only
 181 one pair. We used a maximum distance cutoff between populations (determined visually) to
 182 ensure that remaining populations did not pair with populations that were genetically and
 183 geographically distant. We used a Wilcoxon signed-rank test to test for a mean difference
 184 in NROH10 between population pairs.

185 Predictive model

186 To predict consanguinity score from genetic data, we fitted a logistic regression model where
 187 consanguinity score (2 versus 0 or 1) was treated as the response and NROH10, NROH5,
 188 NROH2, NROH1, and heterozyosity were used as predictor variables. The performance of
 189 the predictive model was evaluated using a leave-one-out approach.

190 Results

191 Expected effect of consanguinity on ROH

192 Based on the analytical result derived in [18], we expect, on average, ≈ 6 ROH longer than
 193 10cM in a population where all unions were between cousins; ≈ 9 if all of them are first
 194 cousins. Our simulations confirm that both the average total length and the number of ROH
 195 (of all lengths) increase as a function of increasing degree of consanguinity as well as the
 196 number of generations for which the practice continues. The simulation results also match
 197 well with the theoretical expectation derived for a single generation of cousin mating (black
 198 points in Fig. 2). In fact, the number of generations (either 1, 10, or 50) has little impact

on the average number or total length of long (>10 cM) ROH (ROH10) (Fig. 2). This is because long runs of homozygosity arise primarily due to cousin unions in the recent past as long segments of the genome do not survive for many generations because of recombination. Therefore, long runs of homozygosity are only informative about consanguinity occurring in the very recent past. In contrast, the number of generations has a much larger impact on the number and length of all ROH, which are sensitive to long-term demographic history [3]. The degree of consanguinity (% unions that are between cousins) affects both long and short runs of homozygosity (Fig. 2).

Geographic variation in cultural preference of consanguinity

We show the geographic variation in cultural preference for consanguinity in Fig. 3. It is interesting to note that in India, preferences for consanguinity appear to markedly change when comparing northern populations to southern populations (Figure 3B). Consanguinity is more prevalent in southern India compared with northern India. More formally, using linear regression, where the consanguinity score was used as a continuous variable, we show that prevalence of consanguinity varies significantly (ANOVA p-value = 1.22×10^{-6}) between broad geographic regions (e.g. East Asia, Northern Africa, and Western Europe). Multinomial ordinal regression, where consanguinity score was treated as an ordinal variable, showed a similar result (p-value = 1.58×10^{-6}). Two regions showed statistically significantly different preferences for consanguinity compared to the average. The Middle East and North Africa predicted a greater preference toward consanguinity (mean consanguinity is 1.67 on our scale from 0 to 2) whereas Western Eurasian populations on average were expected to have a lower preference for consanguinity (mean consanguinity of 0.53).

Genetic footprint of cousin marriages

The number of long (>10 cM) runs of homozygosity (NROH10) was positively associated with the degree of consanguinity when each individual was treated as an observational unit

Figure 2: Distribution of the mean number (A and C) and length (B and D) of runs of homozygosity observed in pedigrees simulated with varying degree of consanguinity (proportion of unions between cousins) and number of generations for which the practice persisted — either 10 (red) or 50 (blue). Colored points and whiskers represent the mean and 95% bootstrap confidence interval, respectively, for (A and B) all runs of homozygosity and (C and D) for long ($>10cM$) runs of homozygosity. Black points represent the theoretical expectation from one generation of consanguinity.

A) Consanguinity Preferences by Region

B) Consanguinity Preferences by Population

Figure 3: Graphical summary of prevalence of consanguinity across the world.

	region2	0	1	2	missing
1	Americas	8		8	2
2	Central Asia & Siberia	12	5	14	8
3	East Asia	11	5	11	1
4	Middle East & North Africa	3	5	22	1
5	Oceania	2		2	
6	South Asia	64	25	93	54
7	Sub-Saharan Africa	33	3	16	11
8	Western Eurasia	37	6	11	14

Table 1: Consanguinity prevalence by region (0 = prohibited, 1 = accepted, 2 = preferred).

(N = 3,859 individuals, $\beta_{score=2} = 0.39$, p-value = 3.02×10^{-14} , Table 2, Fig. 4A). This coefficient can be interpreted as a $1.5\times$ increase in the number of long runs of homozygosity in populations with a strong preference for cousin marriage compared to populations where such unions are prohibited. The estimated effect is consistent if we treat populations (as opposed to individuals) as observational units (N = 396 populations, unweighted $\beta_{score=2} = 0.52$, p-value = 2.31×10^{-4} ; weighted $\beta_{score=2} = 0.60$, p-value = 1.0×10^{-4} ; Fig. 4B; Tables S1 and S2). It is also consistent when we restrict the analysis to populations within South Asia, a region that is very diverse in marriage practices and where we have the largest density of observations (individual model with N = 1363 individuals, $\beta_{score=2} = 0.59$, p-value = 8.68×10^{-10} ; Tables S3, S4, S5; population model with N = 234 populations, unweighted $\beta_{score=2} = 0.49$, p-value = 0.0415; weighted $\beta_{score=2} = 0.62$, p-value = 0.034).

235

Because the covariates in our regression models may not completely capture the effects and interactions of shared ancestry and culture, we confirmed this result with a matched-pair analysis. We identified pairs of populations matched for both genetic similarity and

A) Population model

B) Individual model

Figure 4: Relationship between consanguinity and NROH10. Each point is a population in the top row (A) and an individual in the bottom row (B). Panels in the first column show raw values of NROH10 on the y-axes and panels in the second column show the residuals of a reduced model where consanguinity score has been left out. Black points represent individual observations and red points represent median (point) and mean (cross) and horizontal bars represent the 95% CI.

geographic proximity, but discordant for cousin marriage practice, and then tested for a difference in NROH within pairs (Methods). We find that NROH10 is significantly different between such matched populations ($P = 0.007$, $N = 72$ pairs) (Fig. 5) whereas other variables that are more sensitive to long-term demographic history such as NROH1, NROH2, NROH5, and HET are not (Table 3), consistent with results from the linear models. Removing populations with mean NROH10 > 5 did not change this result (Table S6).

Predicting marriage practice from genetic data

Finally, we asked to what extent it is possible to predict marriage practices based on the genomic distributions of ROH. We predict the binary variable of prohibited versus permitted or preferred cousin marriage using logistic regression using NROH1,2,5 and 10, and heterozygosity. We do not include region or principal components in the regression, which is conservative in this context. Using leave-one-out cross-validation we find that classification power is low (area under the curve, or AUC = 0.62) but well calibrated for predicted probabilities $> 50\%$ (Fig. 6).

Discussion

The negative genetic consequences of marriage between close relatives have been well-known for centuries, yet the practice continues amongst millions of households around the world. Consanguinity creates long runs of homozygosity in the genome, increasing the risk of recessive disease in the offspring. We find a statistically significant relationship between ethnographic assessments of cousin marriage practice and long ROH in 3,859 individuals across 396 worldwide populations. Based on our simulations, this suggests that populations that prefer cousin marriage have, on average, actual first cousin marriage rates of approximately 25 percent (if we assume populations that prohibit cousin marriage have

Figure 5: Paired comparison of NROH10 between populations where consanguinity is prevalent vs populations where it is not ($N = 77$ pairs). (A) Population pairs matched on genetic similarity and geographic similarity. Two pairs (one observed in Central America and one in North Africa) are not shown to maximize resolution in regions with a high density of pairs. (B) The mean number of long ($>10\text{cM}$) runs of homozygosity (NROH10) is greater in populations that prefer cousin unions compared to populations that prohibit them. The black lines demonstrate the comparison between pairs of populations that are geographically and genetically similar while the red and blue lines show the mean and median trend, respectively.

Figure 6: Results from predictive model. The model is well-calibrated (left) and the area under the ROC as the prediction cutoff varies is 0.62 (right).

263 zero percent cousin marriages). This translates into an increase in child mortality or severe
 264 genetic disease of less than 1%—a cost that may in many cases be outweighed by the social
 265 and economic benefits of cousin marriage.

266 Indeed, social scientists have primarily focused on the benefits of cousin marriage, such
 267 as keeping land within the family line [32, 33], creating greater assurance that a marrying
 268 daughter will receive better treatment in her new household [7, 34], marrying a daughter
 269 to her cousin as barter for obtaining a cousin-bride for a son [35], or parents having a
 270 strong preference for the social status of their in-laws [36]. Less emphasis has been placed
 271 on integrating these benefits in to a broader cost-benefit framework that ultimately deter-
 272 mines the emergence, persistence and potential decline of cousin marriage practice at the
 273 population level. Our results suggest that obtaining the benefits of cousin marriage does
 274 not appear to come with a large population level genetic cost. This is consistent with the
 275 long-standing prevalence of cousin marriage preference throughout the world, and suggests

276 that future work incorporate genetic data, which might provide more accurate estimates of
277 actual prevalence of consanguinity, as a cost-metric for this practice.

278 Our study has a number of limitations—some unavoidable. Our classification of mar-
279 riage practices is based on literature search and many of our classifications may be incorrect
280 or fail to reflect complexity or structure within named groups. Similarly, the genetic data
281 we assembled typically have little information about sampling. Group labels may be mis-
282 leading or incorrect and we may have linked them incorrectly to the ethnographic data. In
283 addition since we are, to first order, measuring whether the individuals in our sample have
284 consanguineous parents, our results could be biased if the sampled individuals were not rep-
285 resentative of the wider populations. With more detailed data, many of these limitations
286 can be avoided (e.g. [11]), although at the cost of limiting the geographic and cultural scope
287 of the study. Despite these limitations, we do detect the expected associations, indicating
288 that we are capturing an important part of the effect, although qualitative estimates of
289 magnitude could be underestimated.

290 Finally, we also briefly explored the potential of using the present-day relationship
291 between long ROH and population level assessments of cousin marriage to predict the
292 marriage practices of ancient populations using ancient DNA. This has the potential to
293 track the prevalence of consanguinity through time, especially for populations with sparse
294 archaeological and historical records. Our results show that while high levels of ROH predict
295 cousin marriage well, intermediate and low levels had little predictive power. This suggests
296 that it may be difficult to predict population level preferences for cousin marriage using
297 small samples of ancient DNA. Larger samples, more accurate predictors and integration of
298 archaeological data might improve this prospect. In general, this kind of broad-scale survey
299 of genetic data can be integrated with other types of information to address further questions
300 around the existence, history, and the health and economic effects of consanguinity.

301 Acknowledgements

302 This work was supported by NIGMS award number R35GM133708. The content is solely
303 the responsibility of the authors and does not necessarily represent the official views of the
304 National Institutes of Health.

305 References

- 306 1. Bittles, A. H. & Black, M. L. Consanguinity, human evolution, and complex diseases.
307 *Proceedings of the National Academy of Sciences* **107**, 1779–1786 (2010).
- 308 2. Romeo, G. & Bittles, A. H. Consanguinity in the contemporary world. *Human heredity*
309 **77**, 6–9 (2014).
- 310 3. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of
311 homozygosity: windows into population history and trait architecture. *Nature Reviews*
312 *Genetics* **19**, 220–234 (2018).
- 313 4. Szpiech, Z. A. *et al.* Long runs of homozygosity are enriched for deleterious variation.
314 *The American Journal of Human Genetics* **93**, 90–102 (2013).
- 315 5. Bittles, A. H. & Neel, J. V. The costs of human inbreeding and their implications for
316 variations at the DNA level. *Nature Genetics* **8**, 117–21 (1994).
- 317 6. Sheridan, E. *et al.* Risk factors for congenital anomaly in a multiethnic birth cohort:
318 an analysis of the Born in Bradford study. *Lancet* **382**, 1350–9 (2013).
- 319 7. Mobarak, A. M. *et al.* Estimating the health and socioeconomic effects of cousin mar-
320 riage in South Asia. *Journal of Biosocial Science* **51**, 418–435 (2019).
- 321 8. Wall, J. D. *et al.* South Asian Patient Population Genetics Reveal Strong Founder Ef-
322 fects and High Rates of Homozygosity–New Resources for Precision Medicine. *bioRxiv*
323 doi: 10.1101/2020.10.02.323238 (2020).
- 324 9. Nakatsuka, N. *et al.* The promise of discovering population-specific disease-associated
325 genes in South Asia. *Nature Genetics* **49**, 1403–1407 (2017).
- 326 10. McQuillan, R. *et al.* Runs of homozygosity in European populations. *American Journal*
327 *of Human Genetics* **83**, 359–72 (2008).

- 328 11. Kang, J. T., Goldberg, A., Edge, M. D., Behar, D. M. & Rosenberg, N. A. Consanguin-
329 ity rates predict long runs of homozygosity in Jewish populations. *Human Heredity*
330 **82**, 87–102 (2016).
- 331 12. Arciero, E. *et al.* Fine-scale population structure and demographic history of British
332 Pakistanis. *bioRxiv* doi: 10.1101/2020.09.02.279190 (2020).
- 333 13. Pemberton, T. J. & Rosenberg, N. A. Population-genetic influences on genomic esti-
334 mates of the inbreeding coefficient: a global perspective. *Human heredity* **77**, 37–48
335 (2014).
- 336 14. Finke, K. *et al.* Ancestral Haplotype Reconstruction in Endogamous Populations using
337 Identity-By-Descent. *PLOS Computational Biology* **2**, e1008638 (2021).
- 338 15. National Institute of Population Studies (NIPS) & ICF. *Pakistan Demographic and*
339 *Health Survey Demographic and Health Survey* tech. rep. (2019).
- 340 16. Caballero, M. *et al.* Crossover interference and sex-specific genetic maps shape iden-
341 tical by descent sharing in close relatives. *PLoS Genetics* **15**, e1007979 (2019).
- 342 17. Bherer, C., Campbell, C. L. & Auton, A. Refined genetic maps reveal sexual dimor-
343 phism in human meiotic recombination at multiple scales. *Nature Communications* **8**,
344 1–9 (2017).
- 345 18. Ringbauer, H., Novembre, J. & Steinrücken, M. Human parental relatedness through
346 time - detecting runs of homozygosity in ancient DNA. *bioRxiv* doi: 10.1101/2020.05.31.126912
347 (2020).
- 348 19. Pickrell, J. K. *et al.* The genetic prehistory of southern Africa. *Nat Commun* **3**, 1143
349 (2012).
- 350 20. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for
351 present-day Europeans. *Nature* **513**, 409–13 (2014).
- 352 21. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near
353 East. *Nature* **536**, 419–24 (2016).

- 354 22. Lipson, M. *et al.* Ancient genomes document multiple waves of migration in Southeast
355 Asian prehistory. *Science* **361**, 92–95 (2018).
- 356 23. Jeong, C. *et al.* The genetic history of admixture across inner Eurasia. *Nature Ecology*
357 *and Evolution* **3**, 966–976 (2019).
- 358 24. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse
359 populations. *Nature* **538**, 201–206 (2016).
- 360 25. Pagani, L. *et al.* Genomic analyses inform on migration events during the peopling of
361 Eurasia. *Nature* **538**, 238–242 (2016).
- 362 26. Fan, S. *et al.* African evolutionary history inferred from whole genome sequence data
363 of 44 indigenous African populations. *Genome Biology* **20**, 82 (2019).
- 364 27. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and
365 richer datasets. *Gigascience* **4**, 7 (2015).
- 366 28. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS*
367 *Genetics* **2**, e190 (2006).
- 368 29. *eHRAF World Cultures Database* <https://ehrafworldcultures.yale.edu/ehrafe/>.
369 Accessed: Mar-Jun 2020.
- 370 30. Murdock, G. P. Ethnographic atlas: a summary. *Ethnology* **6**, 109–236 (1967).
- 371 31. Desai, S. & Vanneman, R. *India human development survey-ii (IHDS-ii), 2011-12*
372 (Inter-university Consortium for Political and Social Research Ann Arbor, MI, 2015).
- 373 32. Anderson, N. F. Cousin marriage in Victorian England. *Journal of Family History* **11**,
374 285–301 (1986).
- 375 33. Stone, L. *Family, Sex and Marriage in England, 1500–1800* (New York: Harper &
376 Row, 1977).
- 377 34. Do, Q.-T., Iyer, S. & Joshi, S. The economics of consanguineous marriages. *Review of*
378 *Economics and Statistics* **95**, 904–918 (2013).

- 379 35. Edlund, L. Cousin marriage is not choice: Muslim marriage and underdevelopment.
 380 *AEA Papers and Proceedings* **108**, 353–57 (2018).
- 381 36. Edlund, L. Son preference, sex ratios, and marriage patterns. *Journal of Political*
 382 *Economy* **107**, 1275–1304 (1999).

383

384

385

386

387

388

	β	SE(β)	z value	Pr(> z)
(Intercept)	19.4099	0.4849	40.03	$< 2 \times 10^{-16}$
N.ROH1	-0.0230	0.0020	-11.76	$< 2 \times 10^{-16}$
N.ROH2	-0.0105	0.0032	-3.25	0.0012
HET	-53.2682	0.9573	-55.64	$< 2 \times 10^{-16}$
Consanguinity score = 1	0.1834	0.0625	2.93	0.0033
Consanguinity score = 2	0.3945	0.0519	7.60	3.02×10^{-14}
Region, Central Asia & Siberia	-0.0794	0.3244	-0.24	0.8066
Region, East Asia	-0.7236	0.3421	-2.12	0.0344
Region, Middle East & North Africa	-0.6708	0.3439	-1.95	0.0511
Region, Oceania	0.3227	1.0099	0.32	0.7493
Region, South Asia	-0.4938	0.3396	-1.45	0.1459
Region, Sub-Saharan Africa	0.1513	0.5604	0.27	0.7872
Region, Western Eurasia	-1.1983	0.3351	-3.58	0.0003

Table 2: Poisson regression model (individual) with NROH10 used as response. Coefficients for consanguinity score is shown relative to the score of 0 (consanguinity prohibited). Coefficients of 1 and 2 represent populations which permit and prefer cousin unions, respectively. Regional coefficients are shown relative to the Americas. Ten genetic PCs (calculated for the full sample) were also included in the model but their coefficients are not shown.

	Variable	Statistic	Pvalue
1	NROH10	1405.50	0.007
2	NROH5	1369.50	0.459
3	NROH2	1280.00	0.851
4	NROH1	1118.50	0.274
5	HET	1025.00	0.105

Table 3: Genetic differences between pairs of populations preferring and prohibiting consanguinity (N = 72 pairs)

	β	SE(β)	z value	Pr(> z)
(Intercept)	15.4059	1.6104	9.57	$< 2 \times 10^{-16}$
N.ROH2	0.0185	0.0155	1.19	0.2332
N.ROH1	-0.0351	0.0093	-3.78	0.0002
HET	-42.3276	4.4278	-9.56	$< 2 \times 10^{-16}$
Consanguinity score = 1	0.3102	0.1966	1.58	0.1146
Consanguinity score = 2	0.5183	0.1435	3.61	0.0003

Table S1: Poisson regression model fitted for NROH10 as response (population - unweighted). Ten genetic PCs (calculated for the full sample) were also included in the model but their coefficients are not shown.

	β	$SE(\beta)$	z value	$Pr(> z)$
(Intercept)	13.3861	2.1550	6.21	5.24×10^{-10}
N.ROH2	0.0594	0.0172	3.45	0.0006
N.ROH1	-0.0509	0.0110	-4.63	3.73×10^{-6}
HET	-36.4380	5.6782	-6.42	1.39×10^{-10}
Consanguinity score = 1	0.3202	0.1872	1.71	0.0871
Consanguinity score = 2	0.5995	0.1550	3.87	0.0001

Table S2: Poisson regression model fitted for NROH10 as response (population - weighted). Ten genetic PCs (calculated for the full sample) were also included in the model but their coefficients are not shown.

	β	$SE(\beta)$	z value	$Pr(> z)$
(Intercept)	27.1465	3.2748	8.29	$< 2 \times 10^{-16}$
N.ROH2	0.0277	0.0381	0.73	0.4668
N.ROH1	-0.0343	0.0208	-1.65	0.0983
HET	-75.5652	9.0313	-8.37	$< 2 \times 10^{-16}$
Consanguinity score = 1	0.1002	0.3389	0.30	0.7675
Consanguinity score = 2	0.4922	0.2415	2.04	0.0415

Table S3: Poisson regression model fitted for NROH10 in South Asia (population - un-weighted). Ten genetic PCs (calculated in South Asians only) were also included in the model but their coefficients are not shown.

	β	$SE(\beta)$	z value	$Pr(> z)$
(Intercept)	25.1564	4.5711	5.50	3.73×10^{-08}
N.ROH2	-0.0163	0.0536	-0.30	0.7612
N.ROH1	-0.0224	0.0216	-1.04	0.3003
HET	-70.1002	12.2784	-5.71	1.13×10^{-08}
Consanguinity score = 1	0.4152	0.3693	1.12	0.2608
Consanguinity score = 2	0.6208	0.2923	2.12	0.0337

Table S4: Poisson regression model fitted for NROH10 in South Asia (population - unweighted). Ten genetic PCs (calculated in South Asians only) were also included in the model but their coefficients are not shown.

	β	$SE(\beta)$	z value	$Pr(> z)$
(Intercept)	19.8837	0.5459	36.43	$< 2 \times 10^{-16}$
N.ROH1	-0.0207	0.0038	-5.51	3.6×10^{-08}
N.ROH2	-0.0170	0.0058	-2.91	0.0036
HET	-56.3922	1.4404	-39.15	$< 2 \times 10^{-16}$
Consanguinity score = 1	0.3082	0.1212	2.54	0.0110
Consanguinity score = 2	0.5992	0.0977	6.13	8.68×10^{-10}

Table S5: Effect of consanguinity on NROH10 in South Asia. Poisson regression model (individual-level) with NROH10 used as response. Coefficients for consanguinity score is shown relative to the score of 0 (consanguinity prohibited). Coefficients of 1 and 2 represent populations which permit and prefer cousin unions, respectively. Ten genetic PCs (calculated in South Asians only) were also included in the model but their coefficients are not shown.

	Variable	Statistic	Pvalue
1	NROH10	1155.50	0.04
2	NROH5	1155.50	0.75
3	NROH2	1191.00	0.91
4	NROH1	1091.00	0.62
5	HET	969.00	0.21

Table S6: Genetic differences between pairs of populations preferring and prohibiting consanguinity (outliers removed, N = 68 pairs)