PNAS

Supporting Information for
Increased homozygosity due to endogamy results in fitness consequences in a human population

N.A. Swinford, S.P. Prall, S. Gopalan, C.M. Williams, J. Sheehama, B.A. Scelza, B.M. Henn

Brenna Henn
Email: bmhenn@ucdavis.edu

This PDF file includes:

Supporting Methods
Figures S1 to S7
Supporting References

Supporting Methods

Bottleneck Inference

To determine the effective population size through time, we first identified a subset of unrelated individuals (n=120) and extracted them from the larger, already-quality-controlled (see Genetic Data) dataset. To identify a set of unrelated individuals, we used a custom python script that takes in PONDEROSA (1) and KING (2) output files and a user-specified maximum degree of relatedness (in our case, fifth degree). The script selects the individual who is unrelated to the highest number of other samples and adds him or her to the list, then identifies the next individual who is unrelated to the most other people and adds him or her to the list given that he or she is not already in the list or related to anyone already in the list. We used a custom pipeline similar to that in Gopalan et al (3) to run IBDNe. First, we phased our data with SHAPEIT and used GERMLINE with the --w_extend and the --haploid flags as well as using 'bits' 75 and err_hom set to 1 to identify IBD segments between all pairs of individuals (4). We then joined IBD segments that may have been split due to errors, defined by those that were separated by less than a 0.6 cM gap and that had no more than one discordant SNP. We also filtered out segments with low SNP density which were defined as 1Mb windows that fell in the fifth percentile for SNP count compared to all other 1Mb windows across the genome. Finally, we ran IBDNe (5) specifying a minimum centimorgan length of 4 cM and limited the number of generations before present to calculate N_e for to 100 (gmax 100). The output was graphed in R.

Our msprime simulations followed Gopalan et al (3) and specified a starting N_e of 4000, a final N_e in the present of 450, the generation at which to begin the bottleneck (60 or 6 ga), the number of haploid individuals to simulate (n=120), and a mutation rate of 1e-8. This changing N_e simulation uses a coalescent model with recombination ("Hudson" model) until 100 ga, then switches to a discrete Wright-Fisher model (3). Our starting and final N_e values were 4000 and 450, respectively, to reflect the approximate starting and minimum N_e values inferred by IBDNe in the real data. We chose to simulate a bottleneck beginning 60 ga to reflect what was inferred from our actual data, as well as a more recent historical bottleneck beginning 6 ga to reflect our original hypothesis. We simulated 120 diploid individuals to match the number of Himba individuals used to infer the bottleneck in our actual data. Because simulated data do not have the same issues as actual data (i.e. they are perfectly phased and without genotype errors or missing SNPs), we did not need to use the same custom pipeline to optimize the parameters to infer IBD and run IBDNe. Instead, we processed the vcf output from the simulations, identified IBD using hap-ibd run with default parameters, and ran IBDNe with default parameters.

To run HapLD, we merged 99 unrelated Zulu individuals with our 120 unrelated Himba samples. The Zulu individuals were taken from the AGR dataset, lifted over to hg38, converted to plink format, combined with the Himba, and then filtered for missingness (--geno 0.05) and MAF (--maf 0.00047). We ran HapLD on both the Himba and the Zulu using the HapMap Phase 2 GRCh37 genetic map and default parameters (6).

To run ASCEND, we first converted our merged Himba and Zulu dataset into eigenstrat format using CONVERTF. We ran ASCEND using default parameters on the Himba and specified the Zulu as the outgroup (7).

Pedigrees

For pedigree reconstruction, we used PONDEROSA, an algorithm that infers pedigree relationships and is especially suited for populations with elevated IBD sharing (1) to identify all pedigree relationships in the data. To run PONDEROSA, we phased our Himba genotype data with SHAPEIT and identified IBD segments using GERMLINE with the --haploid and --w_extend flags as well as specifying a minimum match length of 5 cM and a maximum of 3

mismatching SNPs per 100 SNP window (4). PONDEROSA also takes a KING (2) file as input to parent-offspring pairs as well as proportions of IBD1 and IBD2. We ran PONDEROSA specifying a maximum of a 30-year age gap for maternal half sibling relationships, a minimum of a 15-year age gap for parent-offspring relationships, and a minimum of a 30-year age gap for grandparent-grandchild relationships. We also specified that PONDEROSA should not use inferred sibling relationships from KING and used the default parameters of a maximum of 1 discordant SNP and a 1 cM gap to stitch together IBD segments. The Kinship2 R package was used to plot the relationships identified by PONDEROSA.

To estimate the prevalence of close reticulations, we used Ped-Sim (8) to simulate many multi-relationship types: double cousins (co-co), double half-cousins (hco-hco), half-sibling/cousins (hs-co), half-sibling/half-cousins (hs-hco), and half-sibling/second-cousins (hs_sco; e.g., paternal half-siblings whose mothers are first cousins) relationships as well as the standard relationships (all 2nd degree relatives, first cousins, and full-siblings). To do this, we used the crossover interference model and a sex-specific recombination map as in Cabellero et al (8). We seeded the model with Himba individuals and then ran KING on the output files to get the IBD1 and IBD2 values. It is important to note that Ped-Sim can output IBD1 and IBD2 values but this only uses the exact IBD segments that are simulated and would not account for background IBD sharing and would underestimate these IBD values. We also took simulated hs, hs-co, hs-hco, and half-sibling/second-cousins (hs_sco; e.g., paternal half-siblings whose mothers are first cousins) and used them to train a linear discriminant analysis classifier, which we used to classify real Himba half-siblings as either hs only, hs-co, hs-hco or hs-sco.

Supplemental Figures

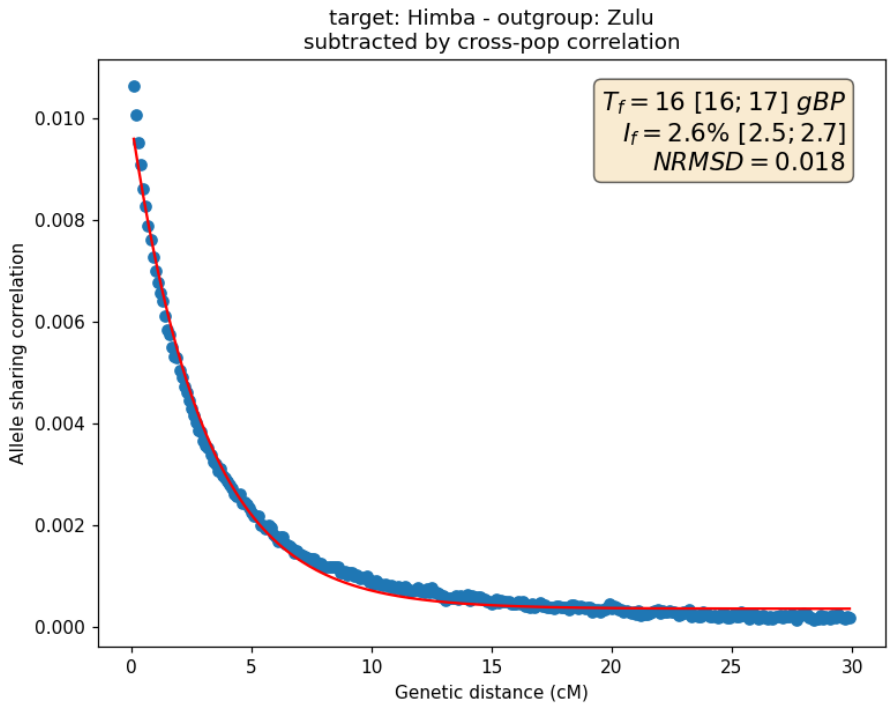

Figure S1. The ASCEND method predicts a founder event 16 generations ago in the Himba with an intensity (calculated as the duration of the bottleneck divided by twice the effective population size during the bottleneck) of 2.6%. 95% confidence intervals for these metrics are shown in brackets. The plot, automatically output by the ASCEND program, displays the correlation for allele sharing decay with increasing genetic distance (blue points) and the fit exponential model (red line).

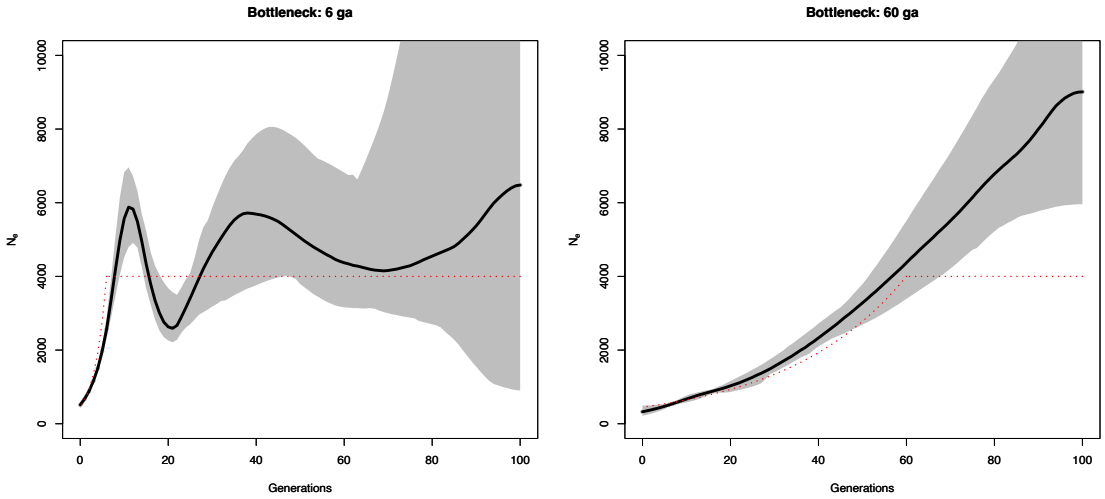

Figure S2. To help validate the results of our inferred bottleneck, we simulated two populations experiencing a bottleneck beginning either 6 ga or 60 ga in msprime and used IBDNe to infer the simulated bottleneck. The black line and 95% confidence intervals (shown in gray) are inferred by IBDNe, and the red dashed lines represent the simulation parameters.

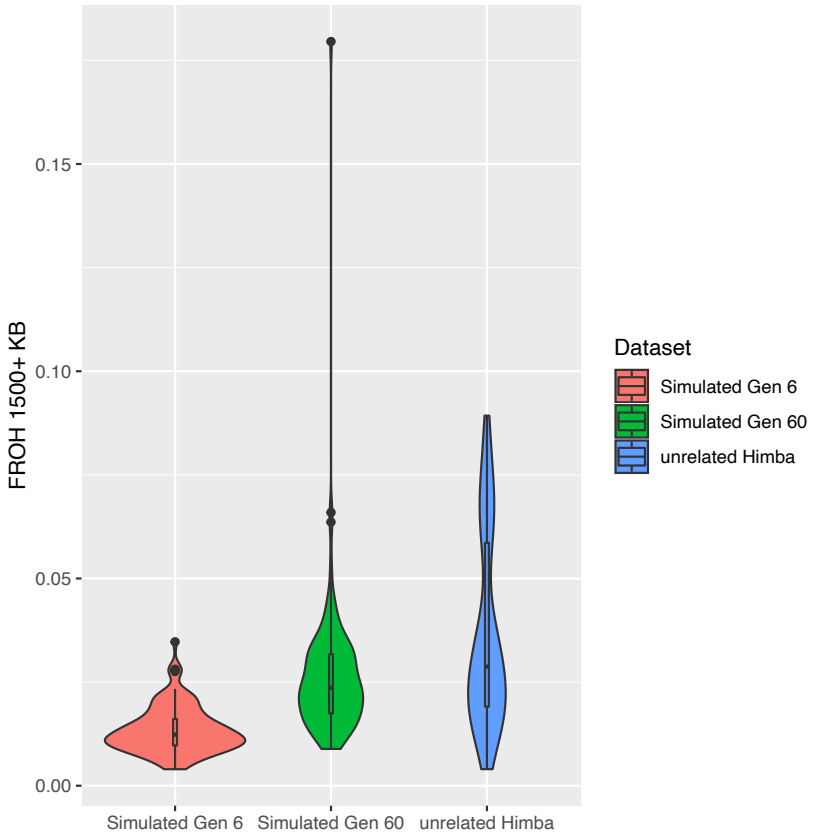

Figure S3. To understand how bottlenecking beginning at different time points produces different distributions of F_{ROH} and validate the timing of our inferred bottleneck, we compared F_{ROH_1500} distributions between simulated and actual individuals. The set of unrelated (n=120) Himba individuals are shown in blue and the simulated individuals with bottlenecks beginning at 6 and 60 generations ago are shown in red and green, respectively.

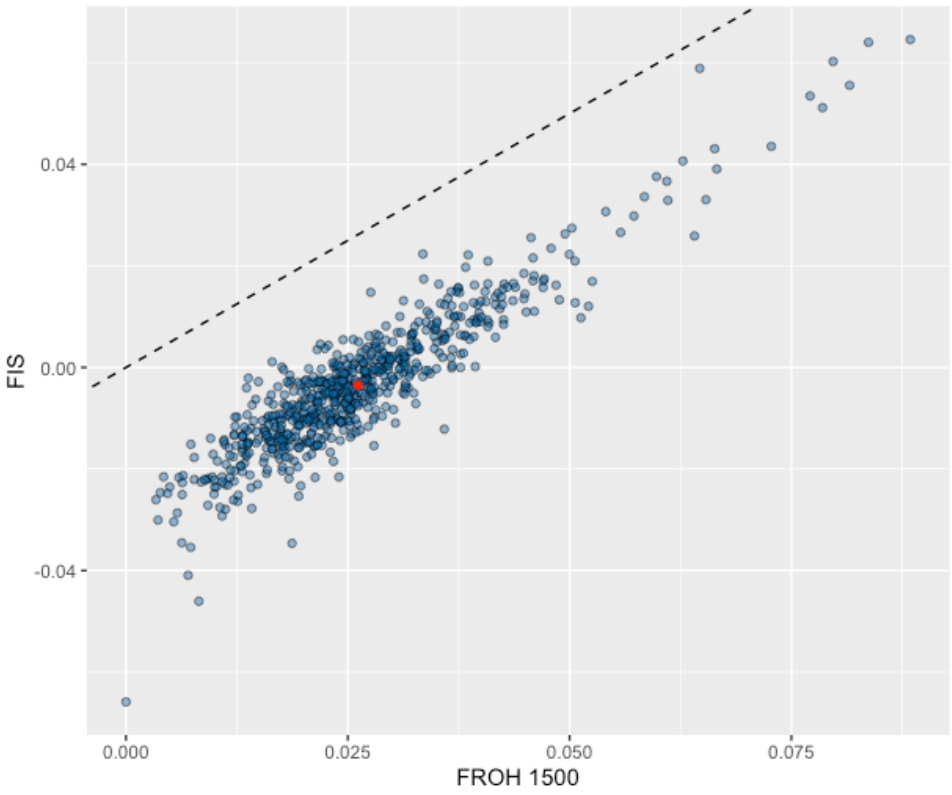

Figure S4. F_{ROH} 1500 vs. F_{IS} is plotted for all 681 Himba individuals (blue points) and is well correlated with a Pearson's correlation coefficient value of 0.92. The dashed line indicates F_{ROH}=F_{IS} and the red point represents the population average of F_{ROH} and F_{IS}. On average, the population has a negative F_{IS} near 0 (FIS = -0.0035) and F_{ROH} 1500 is greater than F_{IS} for all individuals, suggesting that low N_e rather than recent consanguinity is responsible for elevated F_{ROH}.

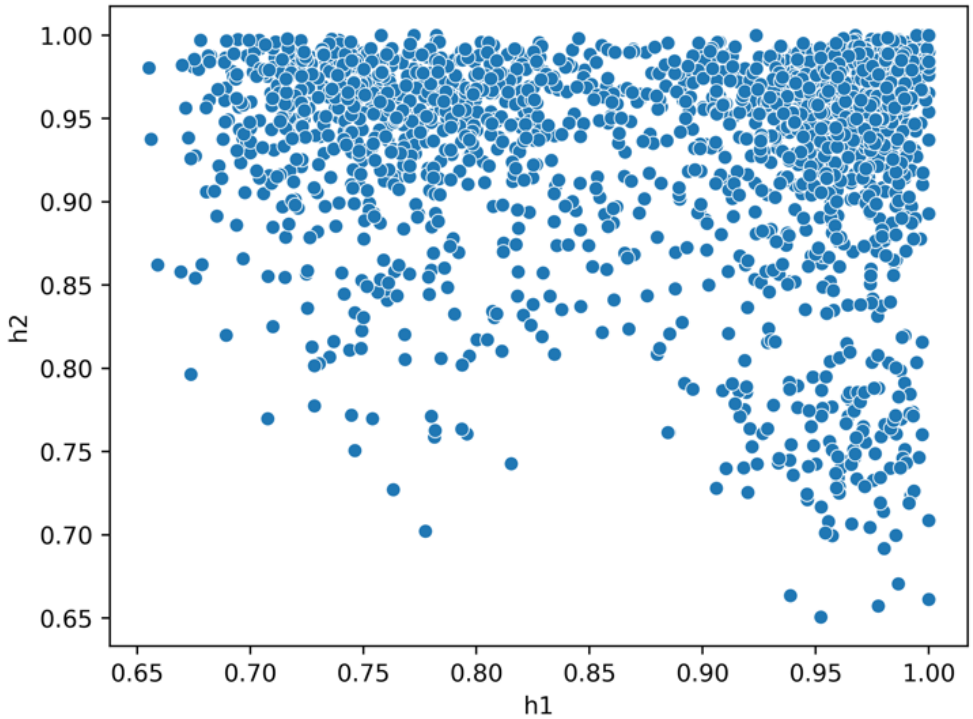

Figure S5. The Himba are very well phased. We plotted the haplotype scores (Williams et al 2020) for 2500 pairs of 2nd degree Himba relatives (blue points). The relationship between these scores (h1 and h2) describes IBD sharing across parental haplotypes. For one member of a 2nd degree relationship (half-sibling, niece/nephew, grandchild), all IBD segments should be on the same parental chromosome, resulting in a haplotype score of 1. For the second member in avuncular and grandparental relationships (i.e. the aunt/uncle or grandparent) the haplotype score will be less than 1 but no more than ½. For half-sibling pairs, both individuals should have haplotype scores of 1. Because at least one member of the 2nd degree relationship (h1 or h2) should have a haplotype score of 1, the points should fall on the top and right borders. If not well phased, shared IBD segments that should be inherited from a single parent will appear broken up between parental chromosomes. However, we do not see this in this plot where only 7% of these pairs have h1 and h2 values <0.90.

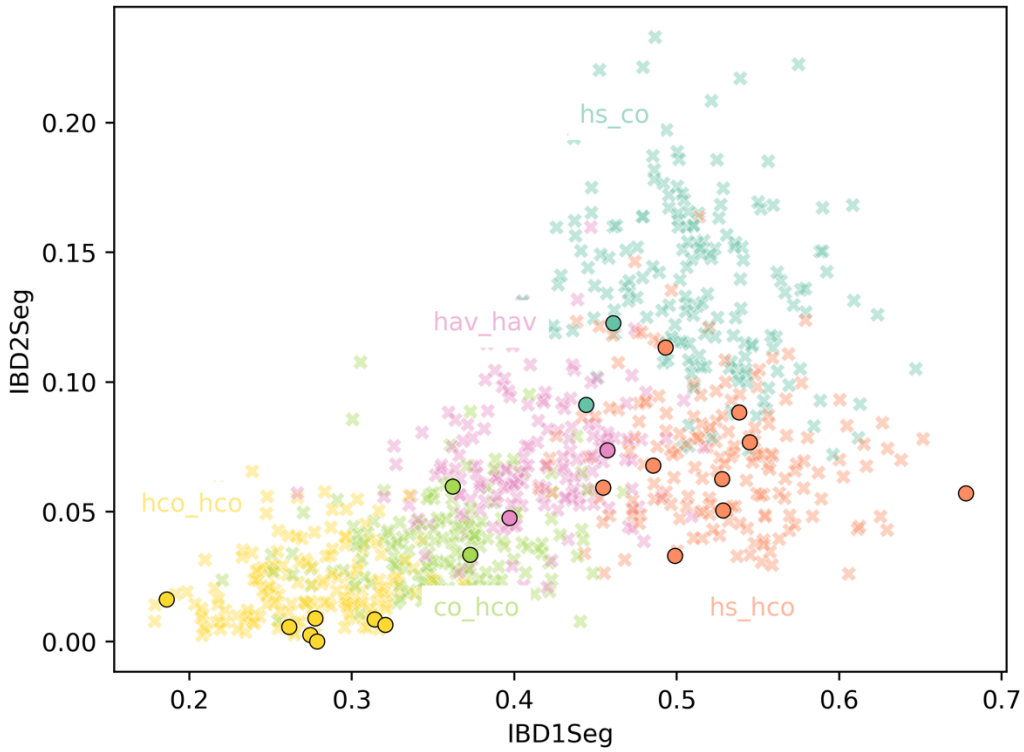

Figure S6. Many Himba relative pairs exhibit high levels of IBD2 sharing that indicate that they are related through both parents. There are several confirmed cases: double cousins (co-co), double half-cousins (hco-hco), half-siblings/half-cousins (hs-hco), half-siblings/cousins (hs-co), cousin/half-cousin (co-hco), and even double half-avuncular pairs (hav-hav). There are 22 of these relationships confirmed, but there are likely more as these can only be confirmed in families with four generations of genotyped individuals. We used Ped-Sim to simulate co-co, hco-hco, hs-co, and hs-hco relationships as well as the standard relationships (all 2nd degree relatives, first cousins, and full-siblings). The 22 confirmed relationships (solid points) are plotted along with the simulated pairs (x's).

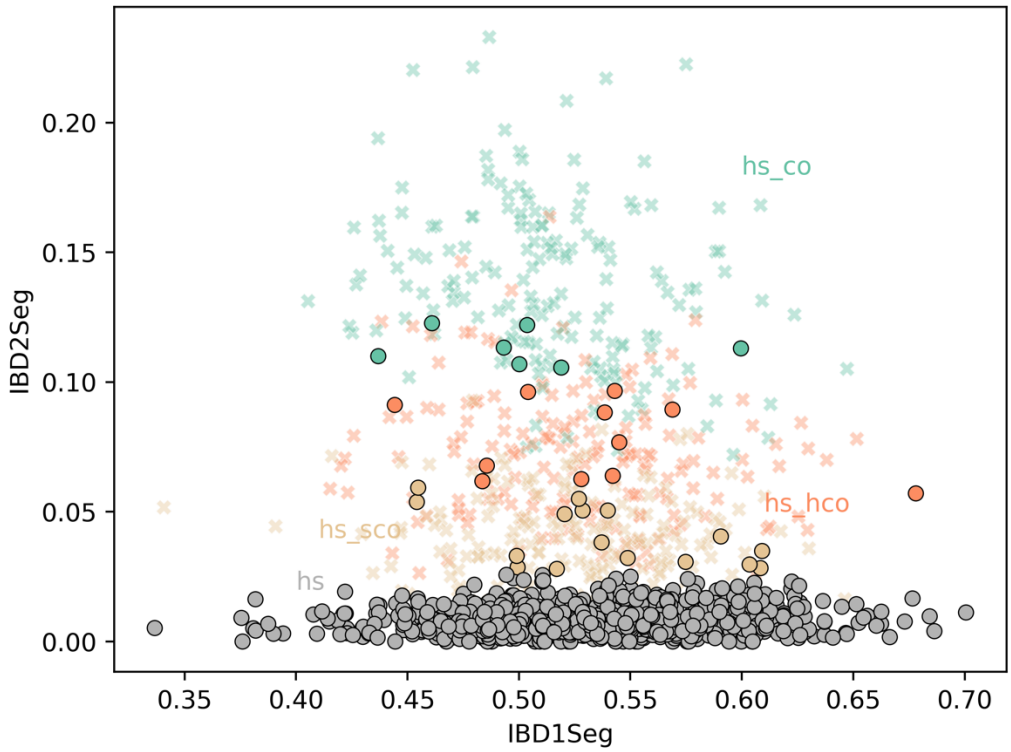

Figure S7. Some half-siblings have close reticulations in their pedigrees. We took Ped-Sim simulated hs, hs-co, hs-hco, and half-sibling/second-cousins (hs_sco; e.g., paternal half-siblings whose mothers are first cousins) and used them to train a linear discriminant analysis classifier, which we used to classify real Himba half-siblings (solid points) as either hs only (gray), hs-co (green), hs-hco (orange), or hs-sco (gold). We classified 34 of the 835 half-siblings as being either hs-co (7) or hs-hco (11) or hs-sco (16). Simulated individuals are shown as x's.

Supporting Methods

1. C. M. Williams, *et al.*, A rapid, accurate approach to inferring pedigrees in endogamous populations. *bioRxiv*, 2020.02.25.965376 (2020).

2. A. Manichaikul, *et al.*, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

2. S. Gopalan, *et al.*, Hunter-gatherer genomes reveal diverse demographic trajectories during the rise of farming in Eastern Africa. *Current Biology* **32**, 1852–1860 (2022).

3. A. Gusev, *et al.*, Whole population, genome-wide mapping of hidden relatedness. *Genome Res* **19**, 318–326 (2009).

4. S. R. Browning, B. L. Browning, Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am J Hum Genet* **97**, 404–418 (2015).

5. R. Fournier, D. Reich, P. F. Palamara, Haplotype-based inference of recent effective population size in modern and ancient DNA samples. *bioRxiv* (2022) https:/doi.org/10.1101/2022.08.03.501074.

6. R. Tournebize, G. Chu, P. Moorjani, Reconstructing the history of founder events using genome-wide patterns of allele sharing across individuals. *PLoS Genet* **18** (2022).

8. M. Caballero, *et al.*, Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLoS Genet* **15**, e1007979 (2019).