The American Journal of Human Genetics, *91*

Supplemental Data

Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History

Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er

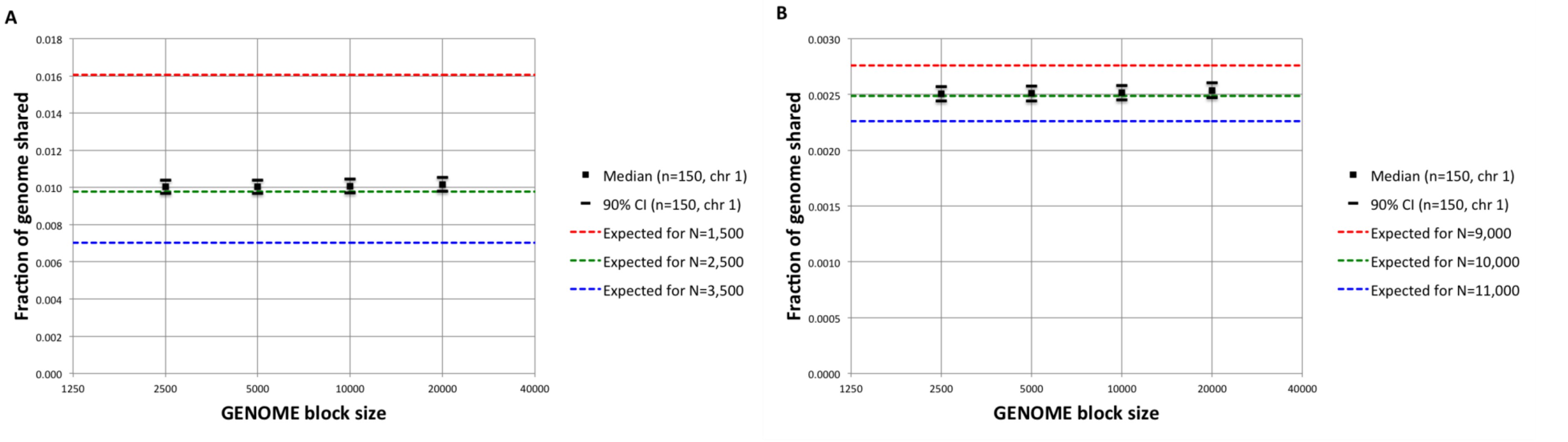

Figure S1. Impact of Simulation Parameters

The GENOME algorithm allows recombination events to only occur at the boundaries of non-recombining blocks of a specified length. This approximation results in substantial run-time and memory improvement, while introducing a potential bias to the length of simulated IBD segments. We note however that when sufficiently long IBD segments are considered (i.e. ≥ 1 cM), effects of such bias are negligible for reasonably short non recombining blocks. To demonstrate this, we simulated several instances of synthetic populations, sampling a realistic Chromosome 1 for 150 diploid individuals from populations of size N=2,500 and N=10,000 (Panel A and B, respectively). We repeated this simulation 1,000 times for block sizes of 2,500; 5,000; 10,000 and 20,000 Kb with recombination rate of 1 cM/Mb. Comparing empirical values to respective theoretical expectations, we observe any introduced bias to be negligible. Note that other minor sources of deviation from the theoretical expectation (e.g. presence of segments truncated by chromosome boundaries) may also affect this analysis.

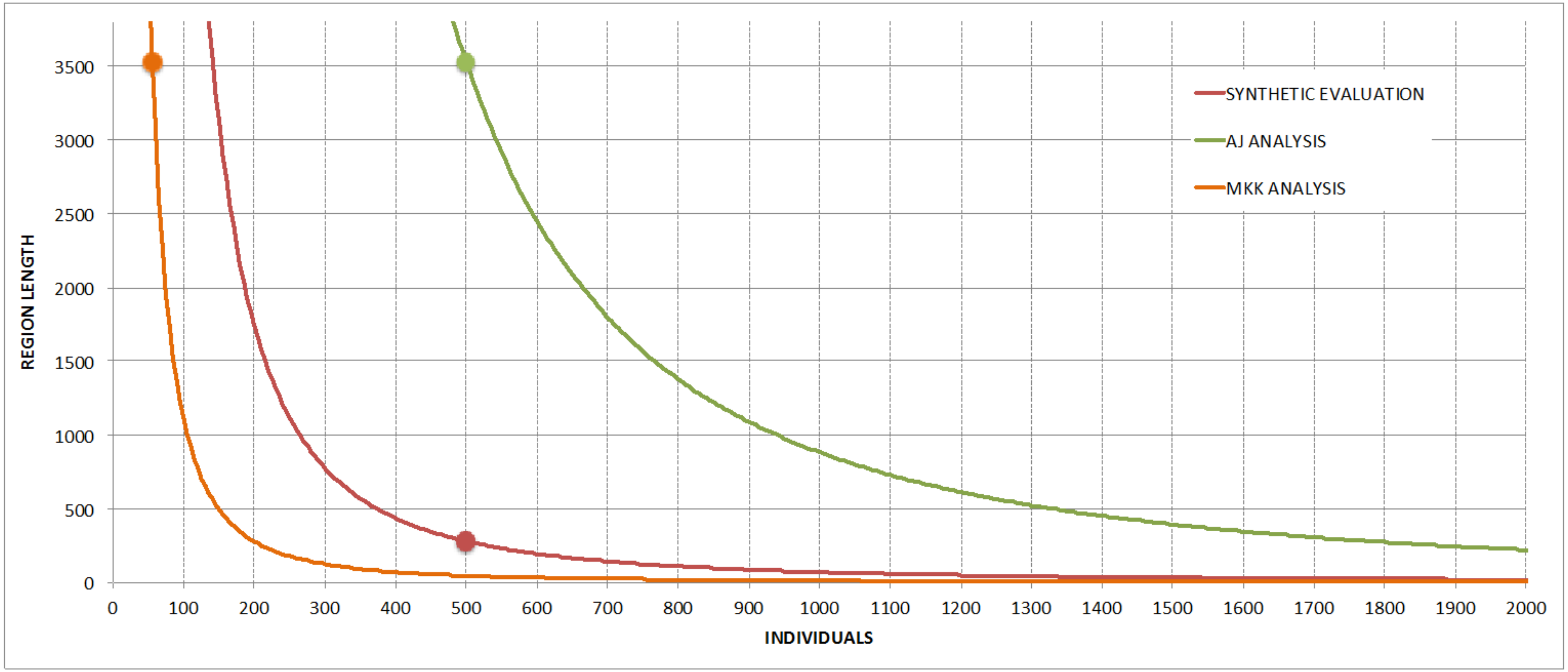

Figure S2. Relationship among Inference Accuracy, Length of Genomic Regions Analyzed, and Number of Samples Considered

We vary the number of samples (x axis) and the length of the considered genomic region (y axis) so that the denominator of Equation 18 remains unchanged. The figure shows a quadratic relationship between the length (in centiMorgans) and number of diploid samples for a fixed number of observed IBD segments. The red dot represents the configuration used for the reported evaluation on synthetic data. Equivalent results are obtained if ~140 diploid samples are analyzed genome wide (~3,500 cM for the autosomal genome). Similar curves are shown for the number of samples and size of the genomic region analyzed for the Maasai (orange) and Ashkenazi (green) populations.

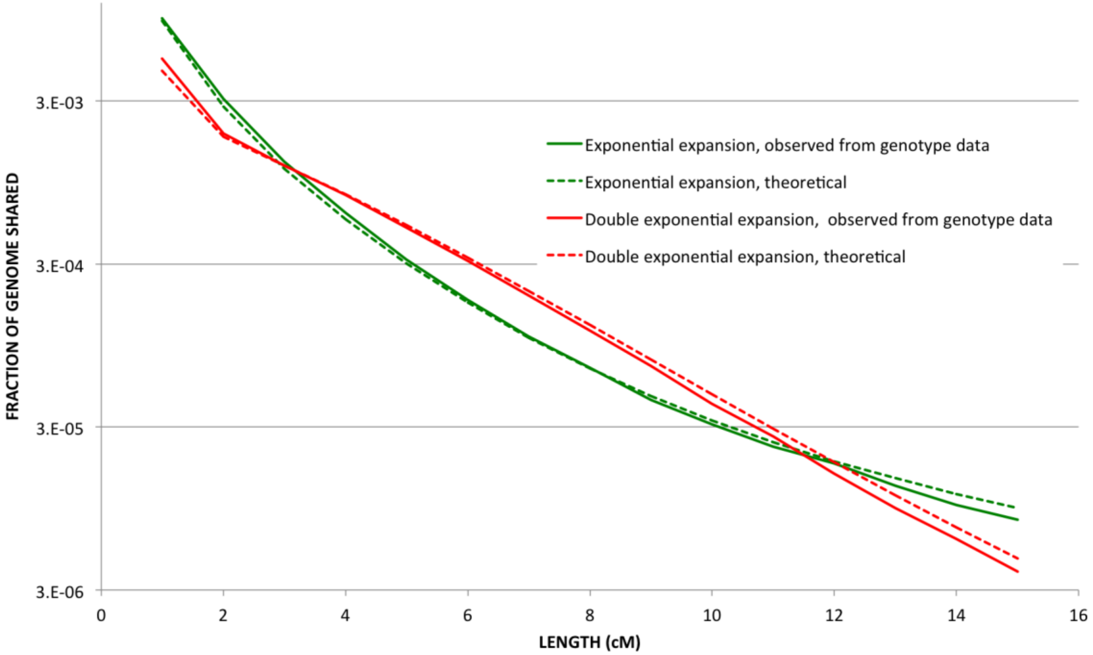

Figure S3. Quality of IBD Discovery

We inferred IBD segments from genotype data of two synthetic datasets of 500 diploid individuals each (22 autosomal chromosomes, using the same GERMLINE parameters as in the AJ population, see Methods). The simulated demographic scenarios are exponential expansion (model $\mathcal{M}_E$, green lines, from 2,500 individuals to 50,000 in 40 generations) and double exponential expansion, separated by a founder event (model $\mathcal{M}_{EFE}$, red lines, Figure 1, model D, using the parameter values inferred for the AJ population, see Results). In both cases we compared the distribution of IBD sharing obtained from computationally phased data (solid lines) to the theoretical distribution for the demographic model considered (dashed lines). Long segments tend to be underestimated, reflecting greater frequency of phasing errors for such long ranges. On the other hand, short segments tend to be overestimated, possibly as a result of longer segments being divided into shorter haplotypes. The analytical procedure applied to IBD obtained from real data will therefore tend to overestimate recent population size, while underestimating more ancient population size, as observed when refining analytical results for AJ demography.

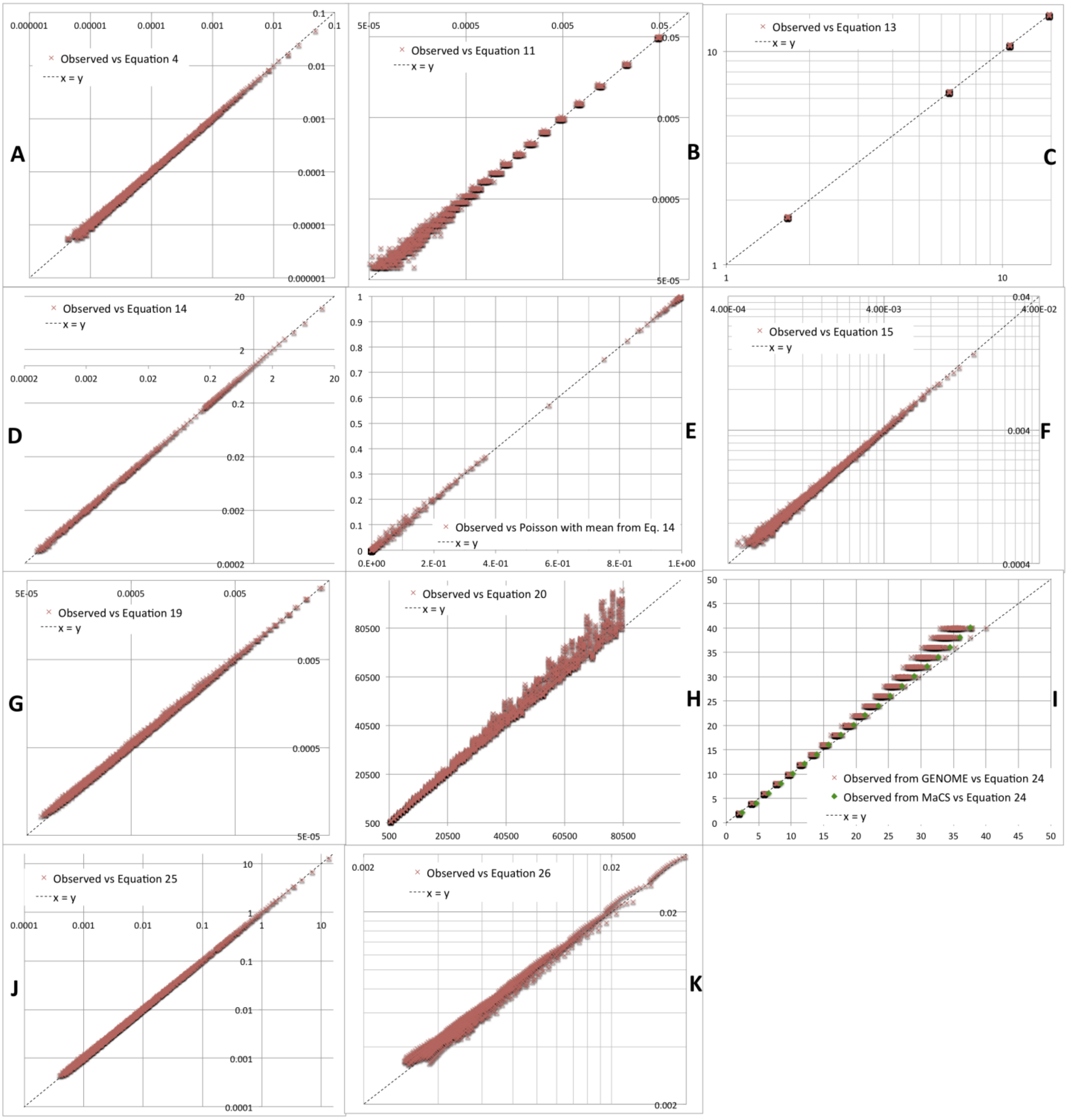

Figure S4. **Simulation-Observed vs. Theoretically Predicted Values for Several Features of IBD Sharing**

We evaluated the presented predictions of IBD sharing quantities (x axis) with the corresponding values empirically obtained from simulated data (y axis). In all cases the observed values were obtained from simulation of populations of constant size from 500 to 40,000 diploid individuals (steps of 500), using a realistic chromosome 1 for 500 diploid samples. The values presented were obtained for length intervals of width between 1 and 4 cM, or for all segments greater than a minimum length threshold. We obtained good correspondence between predicted and observed values for all quantities (panel A through K). Equation 20 (panel H, reporting haploid values) predicted inflated values for population size where larger values of u were used (e.g. minimum observed segment length of 10 cM), while providing accurate predictions for smaller values of u. Similarly, expected length of segments greater than a given threshold u (panel I) tended to be higher than observed as u was increased from 1 to 20 cM. However, we note that these discrepancies could be partially explained by the resolution of 0.01 cM used in simulations based on the GENOME software package. When we simulated a population of 500 diploid individuals using the MaCS simulator (Chen, Marjoram, Wall, *Genome Research* 2009) a smaller inflation was observed. The approximated standard deviation for the fraction of genome shared through segments greater than a minimum threshold u (panel K) performed best for u ranging between 1 and 3 cM.

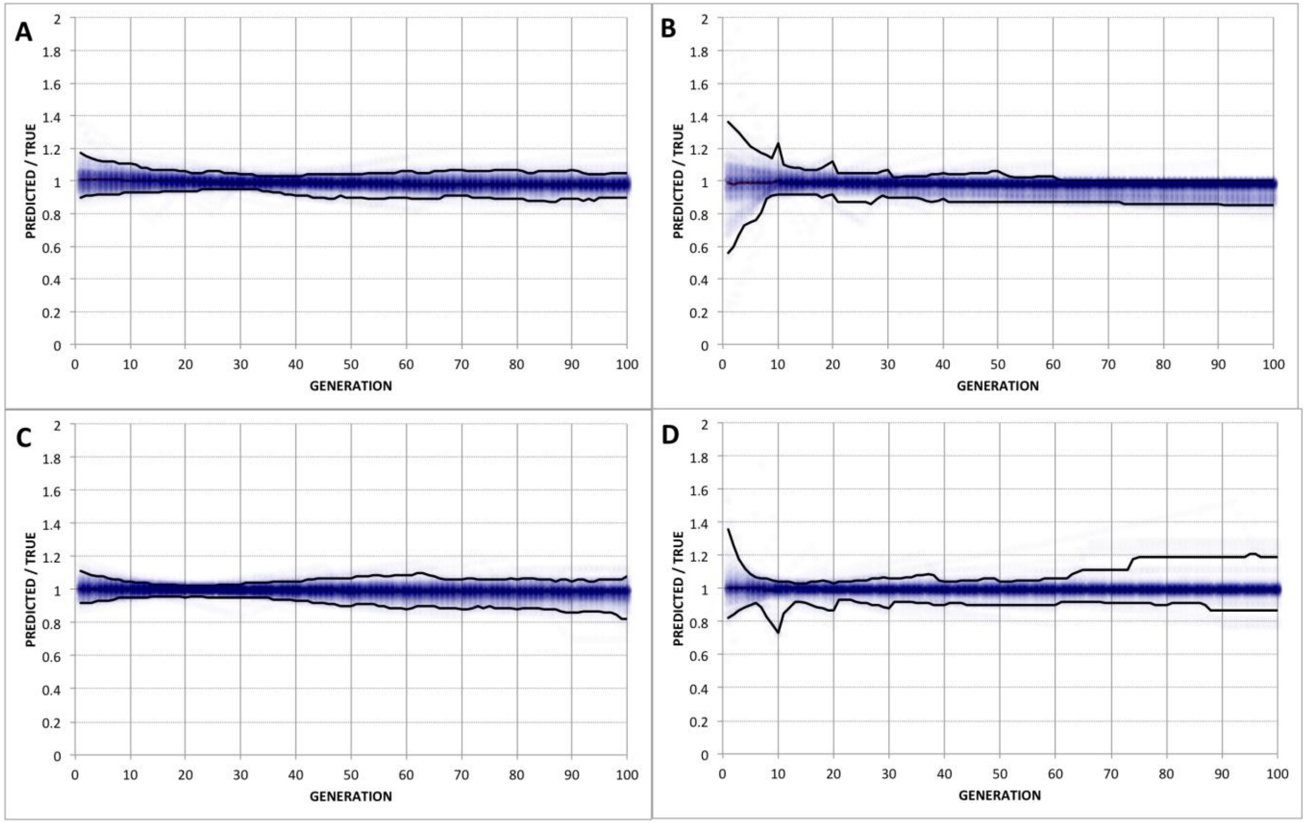

Figure S5. Accuracy of the Demographic Inference for Exponentially Expanding/Contracting Populations

To assess the accuracy of the proposed inference procedure we divided simulated demographic scenarios into mild expansions ($0.046 \leq r$, panel A) strong expansions ($0 < r < 0.046$, panel B), mild contractions ($-0.046 < r < 0$, panel C) and strong contractions ($r \leq -0.046$, panel D). We observed strongest fluctuations in correspondence of recent generations for strong expansions, and ancient generations for strong contractions. In all cases a realistic Chromosome 1 was simulated across 500 diploid samples (comparable results would be obtained analyzing ~140 diploid samples along the entire autosomal genome).

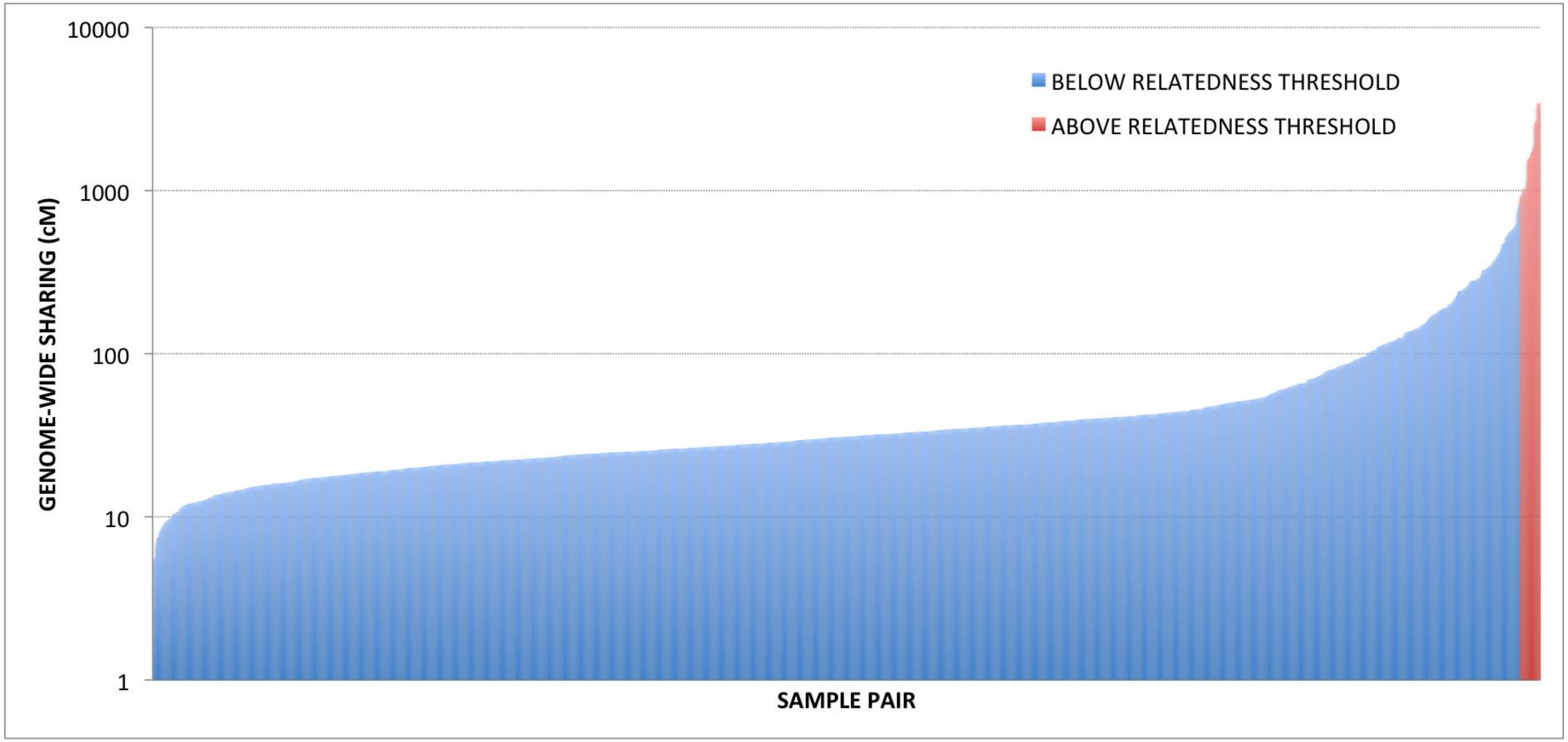

Figure S6. Population-wide IBD Sharing in the MKK Cohort

High levels of IBD sharing are detected across MKK samples. Some pairs share more than 25% of their genome, suggesting unreported relationships of first degree cousin or closer among the samples (red). While this (previously reported) level of IBD sharing is surprisingly high, we find it to be ubiquitous in the cohort, beyond the top-sharing pairs. This suggests that presence of unreported close relatives is not due to specific mistakes in sampling procedures, but rather reflective of a demographic phenomenon across the Maasai population

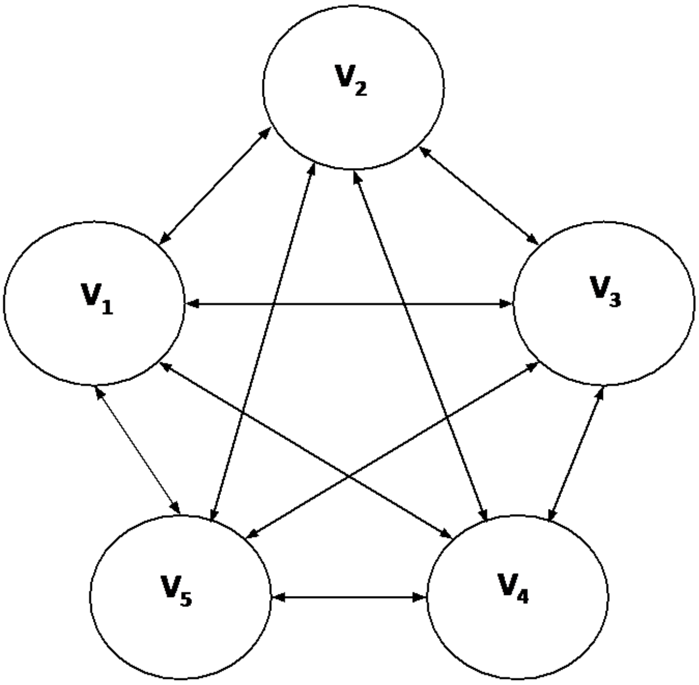

Figure S7. “Village Model” Used for the MKK Analysis

A number of small demes (five villages in this example) of equal, constant population size exchange individuals at a uniform migration rate.

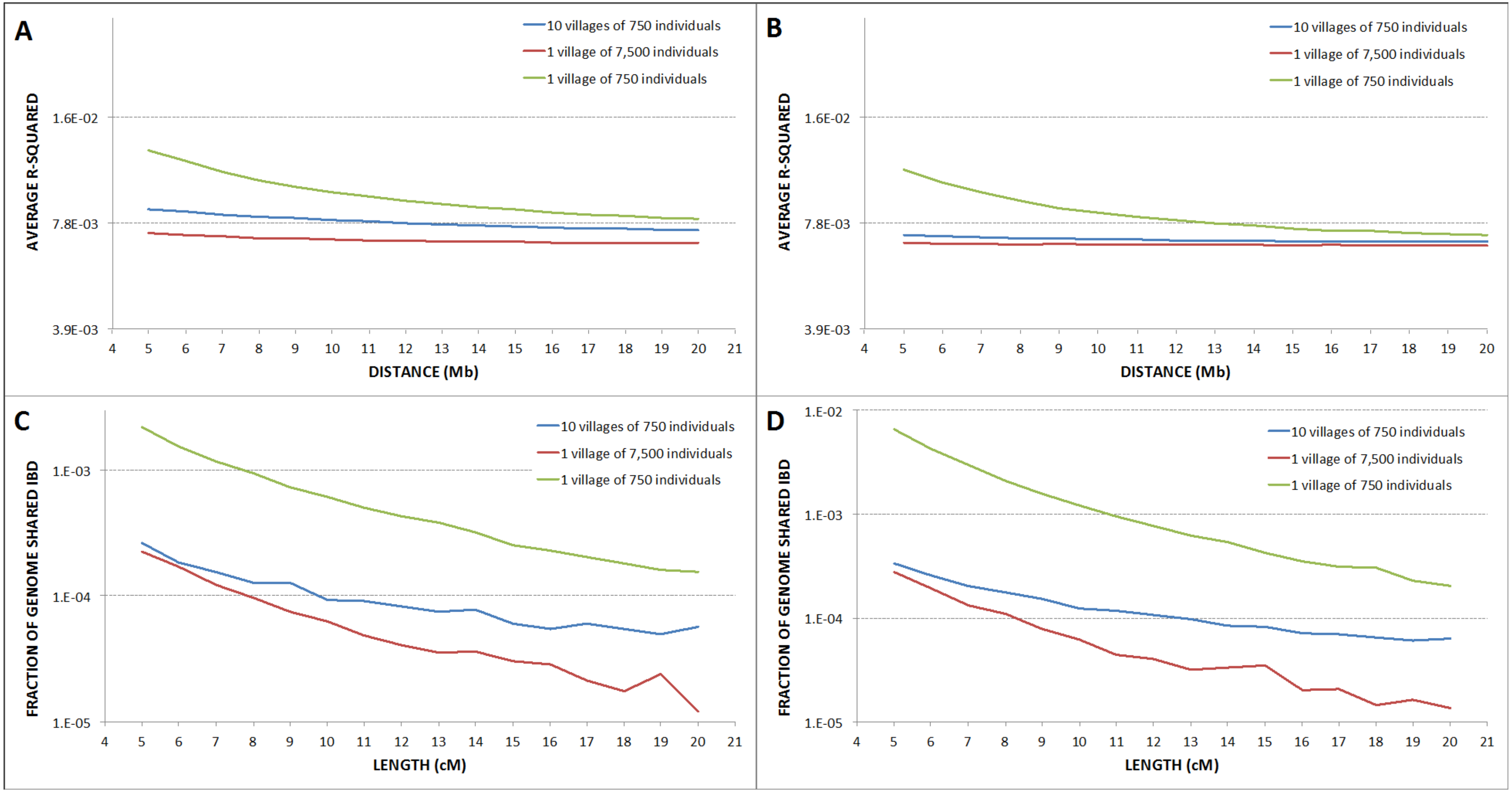

Figure S8. Comparison of IBD and LD Information in the Presence of Phasing Errors

The abundance of long range haplotypes observed in the case of a reduced recent effective population size is reduced (as in the case of the MKK analysis) is likely to have detectable effects on linkage disequilibrium. However, such signature may be difficult to observe due to limitations of statistical phasing. While accuracy of LD measures is strongly dependent on phasing accuracy, current approaches to IBD discovery

(GERMLINE for this study) are robust to occasional phasing errors, and provide a more viable mean to expose long-range linkage of genetic markers. Here we simulated an instance of the “village” model used for MKK analysis, studying the effect of phase inaccuracy on reconstructed IBD and LD information in the range of 4 to 20 cM or Mb (recombination rate was assumed uniform at 1 cM/Mb). LD was computed considering all markers with $MAF \geq 0.05$, by evaluating mean r-squared $r^2 = (f_{AB} - f_A f_B)^2/(f_A f_B f_a f_b)$ for markers with alleles Aa and Bb at specified distances. 600 diploid individuals were extracted from a realistic chromosome 1 in all cases. When perfect phase information is available (A and C) both LD and IBD information can be reliably extracted and used to discriminate between scenarios of a single village of 750 individuals, a single village of 7,500 individuals, or 10 villages of 750 individuals with high migration rates (0.1 per individual per generation). When phase information is computationally obtained (using Beagle), information content of LD decays, while IBD detection using GERMLINE remains of comparably high quality (B and D).

Table S1. Demographic Parameters

Contraction			Range of Grid Search			
N_a	G	N_c	**Parameter**	**Start**	**Interval**	**End**
5000[1]	10	500	N_a	2,500	2,500	100,000
20000	20	2000	G	5	5	200
35000	30	3500	N_c	500	500	20,000
50000	40	5000				
65000	50	6500				
80000	60	8000				
95000	70	9500				
	80					
	90					
	100					
Expansion			**Range of Grid Search**			
N_a	G	N_c	**Parameter**	**Start**	**Interval**	**End**
500	10	5000[2]	N_a	500	500	20,000
2000	20	20000	G	5	5	200
3500	30	35000	N_c	2,500	2,500	100,000
5000	40	50000				
6500	50	65000				
8000	60	80000				
9500	70	95000				
	80					
	90					
	100					

Parameters used to evaluate reconstruction accuracy for exponential expansion/contraction scenarios. We simulated populations with all combinations of demographic parameters listed. The initial grid search step of the inference procedure was performed across the reported range of demographic parameters.

[1] values for which $N_a \leq N_c$ are not considered.
[2] values for which $N_c \leq N_a$ are not considered.

Table S2. Summary of the AJ Demographic Analysis

STEP	DESCRIPTION	NOTES
1	Grid search in parameter space for models $\mathcal{M}_E$, $\mathcal{M}_{FE}$ and $\mathcal{M}_{EFE}$ minimizing Equation 22.	Model $\mathcal{M}_E$ was excluded form further analysis, as it provided a poor fit while suggesting a current population size larger than 10^9 individuals. Grid points for models $\mathcal{M}_{EFE}$, $\mathcal{M}_{FE}$ described in Supplementary Table 3.
2	Gradient-driven refinement of solution obtained from grid search for models $\mathcal{M}_{FE}$ and $\mathcal{M}_{EFE}$.	
3	Local search for maximum likelihood using rejection sampling. Initialized at best-fit point detected in step 2.	
4	Comparison of maximum likelihood parameters for models $\mathcal{M}_{FE}$ and $\mathcal{M}_{EFE}$, using AIC.	
5	Refinement of the maximum likelihood solution obtained for model $\mathcal{M}_{EFE}$ using coalescent simulations to account for phasing uncertainty.	

An algorithmic summary of the steps performed for the AJ analysis is reported.

Table S3. Summary of Parameters Used for AJ Demographic Analysis

PROCEDURE	PARAMETERS	NOTES
Grid search minimization of Equation 22. Grid values.	**Model $\mathcal{M}_{EFE}$:** N_C from 5,013,000 to 50,038,000 interval 75,000 G_1 from 28 to 42 interval 1 N_{A1} from 201 to 275 interval 2 N_{A2} from 50,000 to 210,000 interval 2,500 N_{A3} from 894 to 1,844 interval 50 **Model $\mathcal{M}_{FE}$:** N_C from 200,000,000 to 600,000,000 interval 500,000 G from 18 to 45 interval 1 N_{A1} from 125 to 625 interval 2 N_{A2} from 2,500 to 12,500 interval 50	For both models a preliminary grid search in larger parameter space was performed with coarser resolution. $G_2 = 200$ for all $\mathcal{M}_{EFE}$ points.
Grid search maximization of rejection-based likelihood. Number of sampled points.	**Models $\mathcal{M}_{FE}$ and $\mathcal{M}_{EFE}$:** 10,000 datasets for all points in the neighborhood. After converging, additional sampling up to 100,000 datasets for points with top 10 likelihoods.	

We report grid-search parameters and number of randomly sampled datasets obtained for rejection based likelihood analysis.

Table S4. Grid of Likelihood Values for Model $\mathcal{M}_{EFE}$

N_C G_1 N_{A1} N_{A2} N_{A3}	**Log Likelihood**
84001806-33-454-75620-3616	-2.99573
79801716-33-454-75620-3616	-3.05931
79801716-33-454-79401-3616	-3.11329
84001806-33-454-79401-3616	-3.15356
84001806-33-454-71839-3616	-3.15356
79801716-33-454-86963-3435	-3.23954
88201896-33-454-75620-3616	-3.27017
84001806-33-454-86963-3435	-3.30771
79801716-33-454-83182-3435	-3.3092
75601625-33-454-79401-3616	-3.3192
84001806-33-454-68058-3797	-3.32424
88201896-33-454-68058-3797	-3.37261
84001806-33-454-83182-3435	-3.38139
88201896-33-454-71839-3616	-3.40521
75601625-33-454-75620-3616	-3.42792
79801716-33-454-68058-3797	-3.43918
75601625-33-454-86963-3435	-3.45058
88201896-33-454-79401-3616	-3.47055
79801716-33-454-71839-3616	-3.4767
75601625-33-454-83182-3435	-3.48854
92401987-33-454-75620-3616	-3.56843
84001806-33-454-71839-3797	-3.56843
92401987-33-454-71839-3616	-3.58632
88201896-33-454-83182-3435	-3.60812
88201896-33-454-64277-3797	-3.62309
79801716-33-454-71839-3797	-3.64966
88201896-33-454-86963-3435	-3.70095
84001806-33-454-79401-3435	-3.70095
79801716-33-454-79401-3435	-3.70723
88201896-33-454-71839-3797	-3.7297
92401987-33-454-68058-3797	-3.78099
84001806-33-454-64277-3797	-3.80317
79801716-33-454-83182-3616	-3.80439
88201896-33-454-79401-3435	-3.80766
84001806-33-454-83182-3616	-3.85375
92401987-33-454-64277-3797	-3.86801
92401987-33-454-83182-3435	-3.88246
75601625-33-454-71839-3616	-3.89848
71401535-33-454-79401-3616	-3.92575
75601625-33-454-83182-3616	-3.98264
71401535-33-454-86963-3435	-3.99245

92401987-33-454-79401-3616	-4.03419
75601625-33-454-71839-3797	-4.05338
71401535-33-477-64277-3797	-4.08529
96602077-33-454-68058-3797	-4.09235
96602077-33-454-71839-3616	-4.13517
75601625-33-454-79401-3435	-4.14086
88201896-33-454-68058-3616	-4.14144
71401535-33-454-75620-3616	-4.16986
75601625-33-454-68058-3797	-4.18766
92401987-33-454-71839-3797	-4.22673
79801716-33-454-64277-3797	-4.23675
71401535-33-454-83182-3435	-4.28177
88201896-33-454-83182-3616	-4.28309
92401987-33-454-86963-3435	-4.30507
84001806-33-454-68058-3616	-4.30507
92401987-33-454-79401-3435	-4.32754
84001806-33-454-64277-3978	-4.32754
92401987-33-454-68058-3616	-4.36615
84001806-33-454-75620-3797	-4.37406
71401535-33-454-83182-3616	-4.38567
96602077-33-454-75620-3616	-4.39816
96602077-33-454-64277-3797	-4.43122
88201896-33-454-64277-3978	-4.48295
79801716-33-454-75620-3797	-4.51816
71401535-33-477-68058-3616	-4.54347
96602077-33-454-79401-3435	-4.57561
92401987-33-454-64277-3978	-4.58537
79801716-33-454-64277-3978	-4.59612
71401535-33-477-64277-3616	-4.64221
79801716-33-454-68058-3616	-4.68068
79801716-33-454-86963-3616	-4.71053
75601625-33-454-75620-3797	-4.73094
96602077-33-454-83182-3435	-4.733
79801716-33-454-75620-3435	-4.76236
96602077-33-454-68058-3616	-4.82831
88201896-33-454-75620-3797	-4.82831
84001806-33-454-75620-3435	-4.84089
75601625-33-454-86963-3616	-4.863
96602077-33-454-79401-3616	-4.87961
96602077-33-454-71839-3797	-5.02069
71401535-33-477-71839-3616	-5.08614
92401987-33-454-75620-3435	-5.116
71401535-33-454-71839-3616	-5.22811
79801716-33-454-68058-3978	-5.24521

75601625-33-454-64277-3978	-5.35441
71401535-33-454-79401-3435	-5.37383
84001806-33-454-68058-3978	-5.44914
75601625-33-454-75620-3435	-5.49899
92401987-33-454-75620-3797	-5.68398
75601625-33-454-68058-3978	-6.04755
88201896-33-454-86963-3616	-6.07485

We obtained approximate likelihoods for several parameter values surrounding the maximum likelihood point. We sampled all parameter values across a grid defined as follows (integer rounded haploid values for population size reported): N_C from 71,401,535 to 96,602,077 interval 4,200,090 (MLE±15%); G_1 from 32 to 34 interval 1 (MLE±3%); N_{A1} from 409 to 499 interval 22 (MLE±10%); N_{A2} from 64,277 to 86,963 interval 3,781 (MLE±15%); N_{A3} from 3,074 to 4,158 interval 180 (MLE±15%); $G_2 = 200$ for all points. We used the tolerance threshold of $\delta \simeq 0.04$. We selected all parameter values for which at least 4 out of 1,000 samples passed such threshold. We report log-likelihood values at these points after sampling additional datasets, for a total of at least 10,000 per point.