SUPPLEMENTAL INFORMATION:

Fine-scale resolution of human recombination using topological data analysis

Pablo G. Camara[1,2*], Daniel I. S. Rosenbloom[1,2], Kevin J. Emmett[1,3], Arnold J. Levine[4], Raul Rabadan[1,2*]

[1] Department of Systems Biology,

[2] Department of Biomedical Informatics,

Columbia University College of Physicians and Surgeons,

1130 St. Nicholas Ave., New York,

[3] Department of Physics,

Columbia University, New York, NY 10032

[4] The Simons Center for Systems Biology,

Institute for Advanced Study, Princeton, NJ 08540.

- Corresponding authors contact: pg2495@cumc.columbia.edu

rr2579@c2b2.columbia.edu

Supplemental Figures

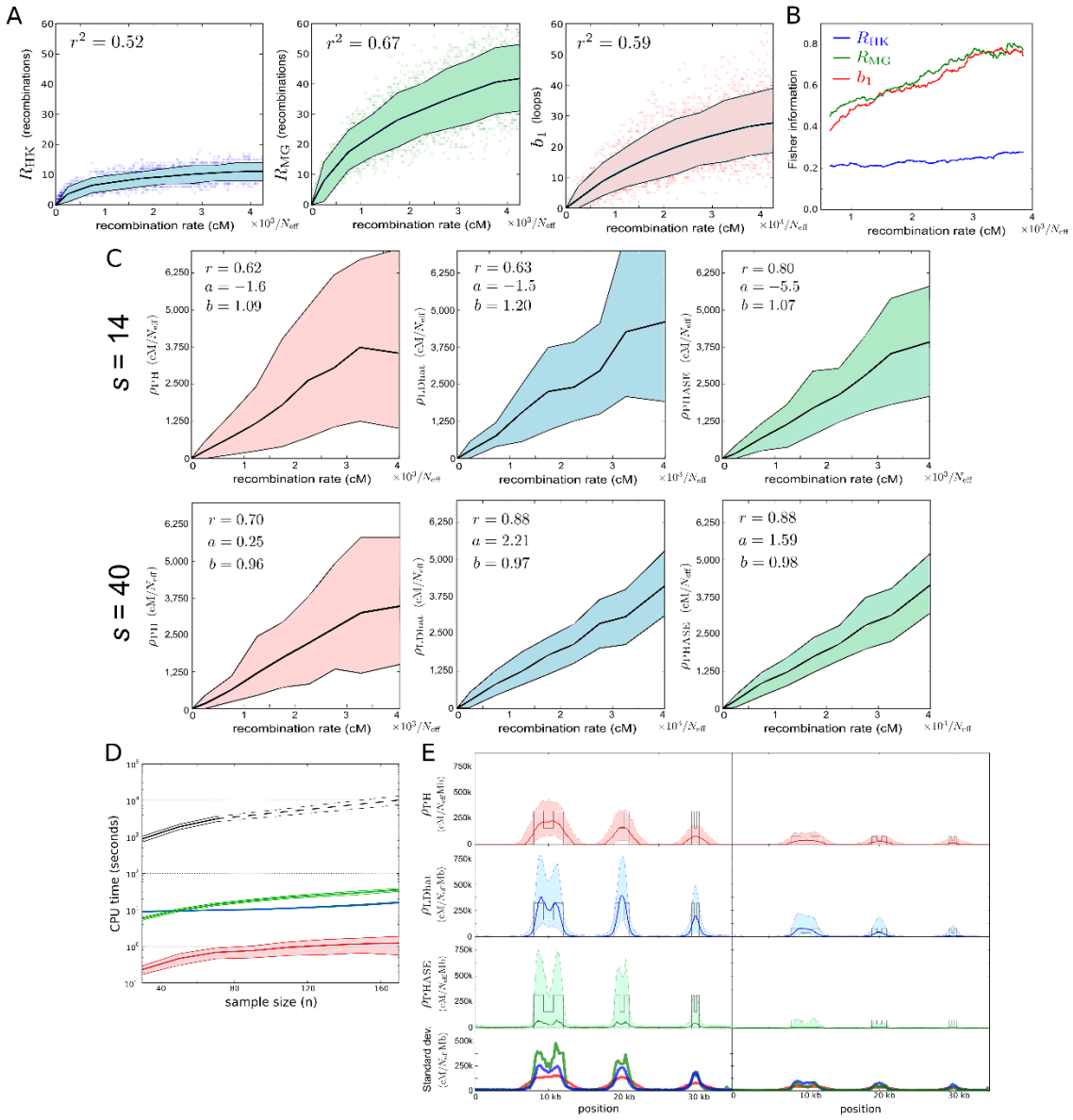


Figure S1. Persistent homology estimator of recombination (related to Figure 2).

(**A**) Dependence of Hudson-Kaplan (left), Myers-Griffiths (center) and b_1 (right) summaries of recombination on the recombination rate for simulations at fixed number of segregating sites (s =

40). Each plot is based on 4,000 coalescent simulations of a sample of 160 sequences. Colored bands represent the interdecile range and central line corresponds to the mean. Squared Pearson's correlation coefficient is shown in each case.

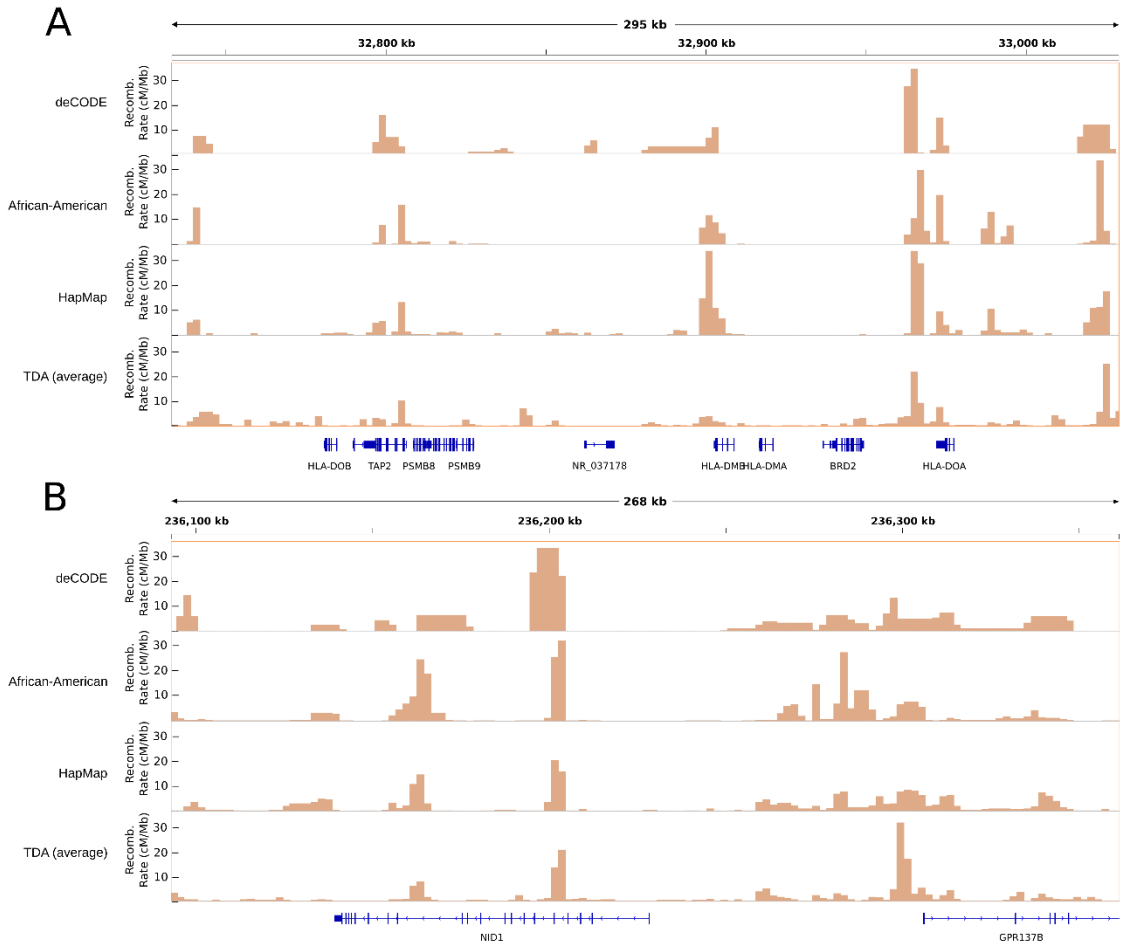
(**B**) Fisher information for each of the 3 summaries in (A) as a function of the recombination rate. Information was computed in increments of $12.5/N_{\text{eff}}$ cM. A smoothed trend is plotted by averaging windows of 101 computed values, weighted by the number of simulations.

(**C**) ρ_{PH} (left), `LDhat interval` (center) and `PHASE` (right) estimates of the recombination rate, for simulated samples of 160 sequences at constant recombination rate and fixed number of segregating sites (s = 14 (top) and s = 40 (bottom)). Central lines correspond to the mean and colored bands represent the interdecile range of the estimates. Linear regression parameters and Pearson's correlation coefficient are shown in each case.

(**D**) Performance of ρ_{PH} (red), `LDhat interval` (blue), `LDhat pairwise` (black) and `PHASE` (green) for simulated samples at constant recombination rate and 14 segregating sites. Average computing time in a CPU of a standard modern desktop is represented against the number of sampled sequences. `LDhat pairwise` was run for $n < 80$. Vertical axis is in logarithmic scale.

(**E**) ρ_{PH}, `LDhat rhomap` and `PHASE` estimates of the recombination rate for simulated samples of 160 sequences 35 kbp long with background recombination rate $500/N_{\text{eff}}$ cM/Mb and six recombination hotspots of widths 4 kbp, 2 kbp and 1 kbp. Local recombination rate is enhanced at hotspots by a factor 640 (left) and 160 (right). Intra-hotpot recombination rate variation is also simulated, with a ½ decay of the local recombination rate at the central region of hotspots.

Central lines correspond to the median and colored bands denote the interdecile range of the estimates. ρ_{PH} was computed using a sliding window of variable size with fixed number (s = 14) of segregating sites moved in steps of 7 segregating sites. The standard deviation of the three estimators (red: ρ_{PH}, blue: `LDhat`, green: `PHASE`) is shown at the bottom, with ρ_{PH} having the lowest variance.



C

TDA	deCODE	African-American	Hapmap
Phase 1 data (s = 14)	0.35 – 0.39	0.44 – 0.50	0.48 – 0.54
Phase 1 data (s = 40)	0.43 – 0.48	0.53 – 0.61	0.54 – 0.63
Phase 3 data (s = 14)	0.44 – 0.46	0.57 – 0.61	0.62 – 0.66

Figure S2. Comparison of deCODE, HapMap, African-American and TDA recombination maps (related to Figure 3).

(**A**, **B**) Comparison across ~300 kbp regions within the major histocompatibility locus (A) and the minisatellite MS32 region (B). All maps were binned at 2 kbp in this figure. We took the sex-averaged version of deCODE map and the average of the seven TDA recombination maps

considered in this work. An average whole-genome recombination rate of 1.16 cM/Mb, observed in genetic linkage experiments (Kong et al., 2010), has been used to normalize the TDA recombination map.

(**C**) Whole-genome Spearman correlation between 10 kbp binned recombination maps, using data from phase 1 and phase 3 releases of 1,000 Genomes Project for the TDA recombination maps.

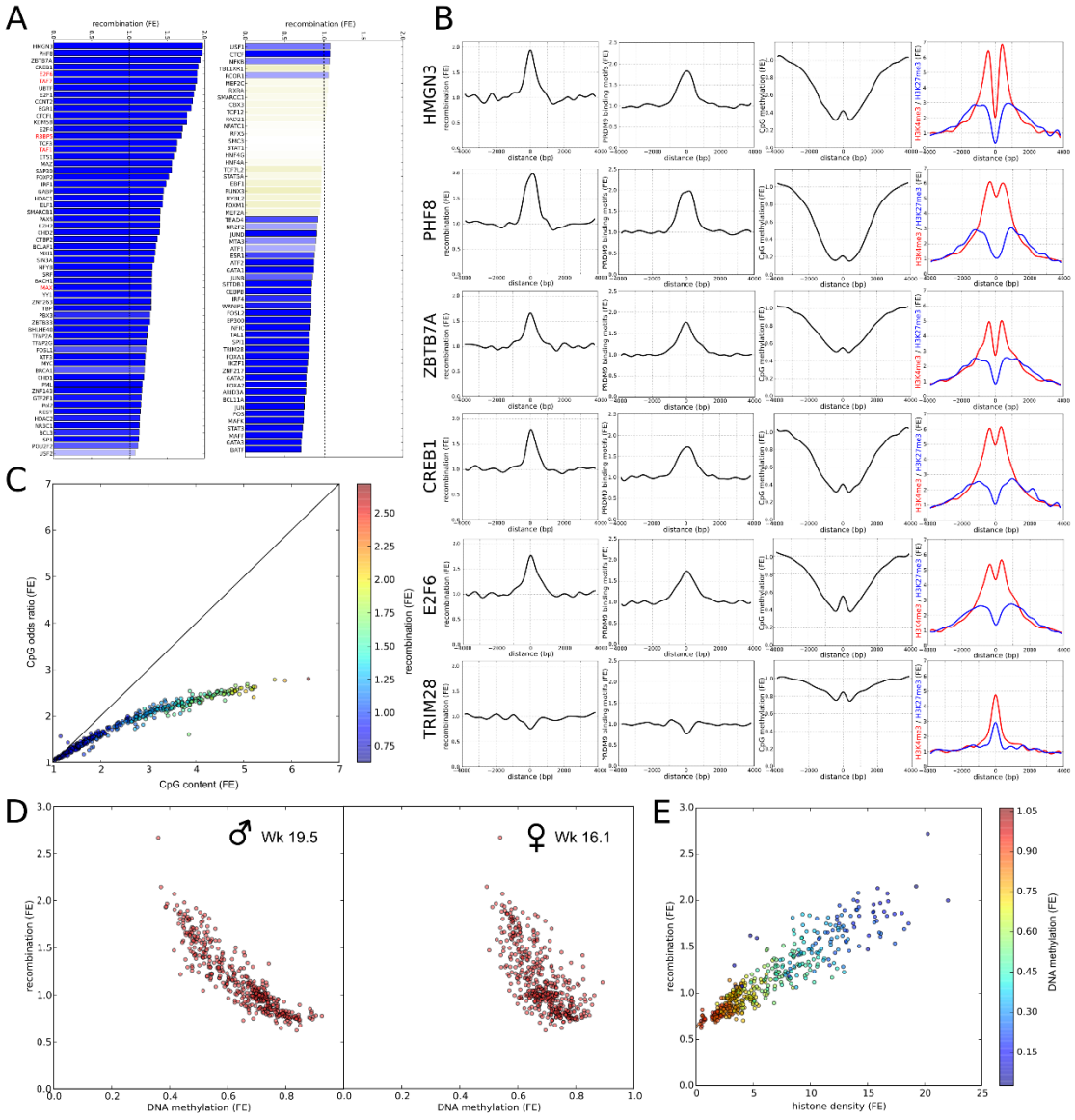


Figure S3. Recombination enrichment at TF binding sites (related to Figure 4).

(**A**) Recombination enrichment with respect to the whole-genome average for the TF binding sites. Binding sites are based on data from ENCODE (Gerstein et al., 2012). In total 118 TFs and 91 cell lines were considered. ChIP-seq peaks of each TF were merged across all cell lines. TFs

that do not show a significant (q < 0.05, Benjamini-Hochberg) enrichment or depletion of recombination are shaded. Recombination enrichments are based on the GBR population. TFs that may for part of MLL1/2 complexes (indicated in red) are generally enriched for recombination.

(**B**) Enrichment for recombination, predicted PRDM9 binding sites (defined by the motif CCNCCNTNNCCNC), and sperm CpG methylation, H3K4me3 and H3K27me3 marks as functions of the distance to several recombination enriched TF binding sites. Recombination enrichments were computed using the 500bp recombination map of the GBR population for several TF binding sites considered in (A). For comparison, the analysis of binding sites of TRIM28, which do not exhibit any recombination enrichment, is also shown.

(**C**) CpG odds ratio versus CpG content fold enrichment at TF binding sites considered in Figure 4A. Each point corresponds to a different combination of TF and cell line. Binding sites are based on ChIP-seq data from ENCODE. In total 118 TFs and 91 cell lines were considered. The observed CpG enrichment at TF binding sites is only partially explained by an enhancement of the CpG odds ratio (for most TFs, CpG odds ratio enhancement ≤ CpG enrichment), indicating that part of the CpG enrichment is simply due to an enrichment for GC content.

(**D**) Recombination enrichment against CpG methylation enrichment at the loci of TF binding sites in human PGCs of male 19.5 week (left) and female 16.1 week (right) embryos, for the same TFs and cell lines considered in Figure 4E. Recombination enrichments were estimated using the 500bp recombination map of the GBR population.

(**E**) Recombination enrichment against histone enrichment at the loci of TF binding sites in human sperm, for the same TFs and cell lines considered in Figure 4E. Based on MNase-seq data from (Hammoud et al., 2009). Color scale represents CpG methylation enrichment in sperm.

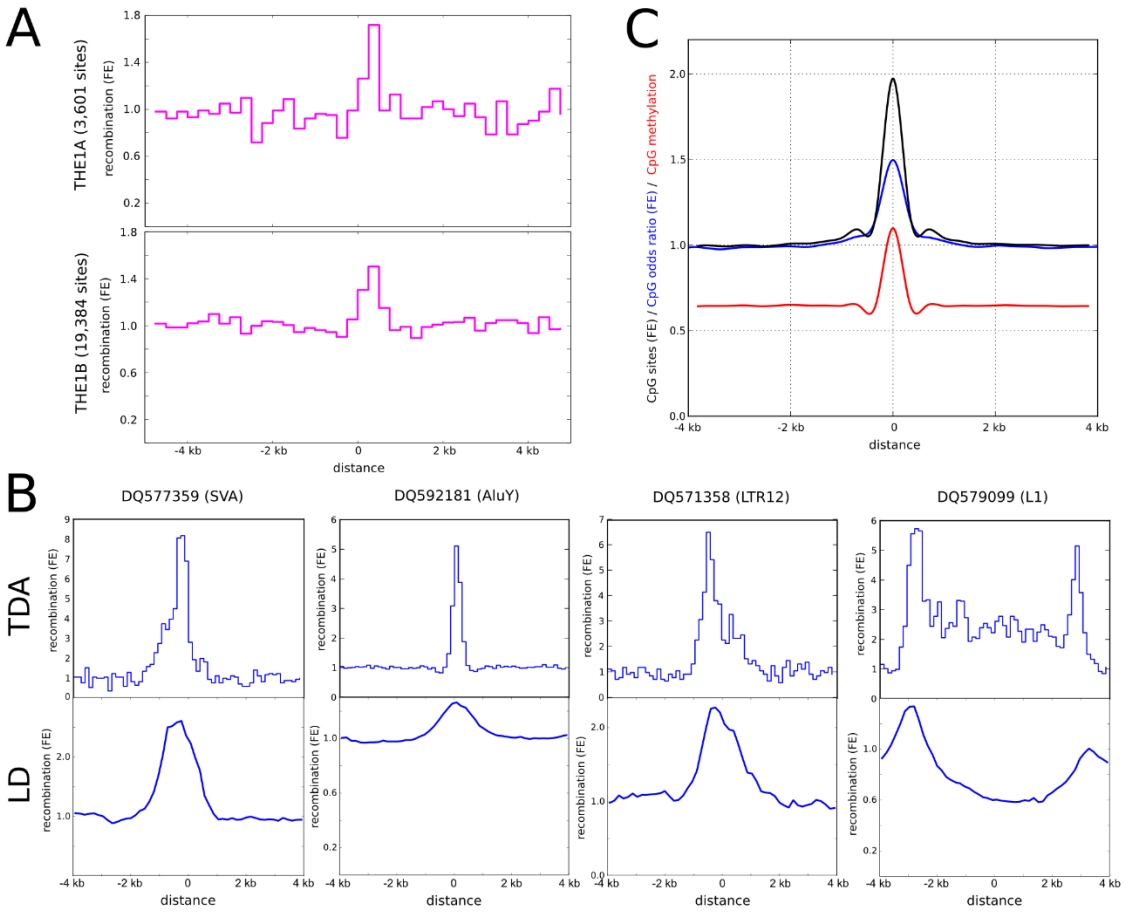


Figure S4. Recombination enrichments at repeat-derived loci matched by piRNA (related to Figure 5).

(**A**) Distribution of recombination enrichment around THE1A/B elements. Recombination enrichments were estimated using the 500bp recombination map of the GBR population.

(**B**) Enrichment for recombination for loci matched by four specific repeat-derived piRNA (piRNA-Bank accession numbers DQ577359, DQ592181, DQ579099 and DQ571358), estimated using the TDA (500bp) recombination map of the GBR population (top) and `LDhat` on

1,000 Genomes Project data (bottom). The origin of coordinates corresponds to the location of the piRNA-matched sequence.

(**B**) Relative CpG abundance, CpG odds ratio and fraction of methylated sites for piRNA-matched loci. The origin of coordinates corresponds to the location of the piRNA-matched sequence. The observed CpG enrichment at piRNA-matched loci is only partially explained by an enhancement of the CpG odds ratio (CpG odds ratio enhancement ≤ CpG enrichment), indicating that part of the CpG enrichment is due to an enrichment for GC content.

Supplemental Tables

Table S1 (Provided as a separate spreadsheet). Estimates of $\lambda = E[b_1]$ for different number of sampled sequences (n), population recombination rate (ρ) and mutation rate (θ) in simulated Wright-Fisher models with recombination (related to Figure 2).

Table S2. Summary of the 1,000 Genomes Project populations considered and their estimated effective population size using ρ_{PH} on a sliding window with fixed number of segregating sites (s = 40) (related to Figure 3). Effective population sizes are determined using the formula $N_{eff} = E[\rho]/(4rL)$, with r = 1.16 cM/Mb measured at genetic linkage experiments (Kong et al., 2010) and L expressed in number of nucleotides.

		Samples			
Acronym	Description	Males	Females	Total	N_{eff}
CEU	Utah residents with Northern and Western European ancestry	44	40	84	27,700
CHB	Han Chinese in Bejing, China	44	53	97	31,300
FIN	Finnish in Finland	36	58	94	26,400
GBR	British in England and Scotland	41	48	89	30,000
JPT	Japanese in Tokyo, Japan	50	38	88	28,700
LWK	Luhya in Webuye, Kenya	48	49	97	43,500
TSI	Toscani in Italy	50	48	98	31,500
Total:		313	334	647	

Table S3. Association between recombination enrichment at TF binding sites, epigenetic markers in sperm and predicted PRDM9 binding sites, using data from phase 1 and phase 3 releases of 1000 Genomes Project (related to Figure 4). Statistical *p*-values are smaller than 10^{-50} for all associations.

Association	Pearson's *r*, phase 1	Pearson's *r*, phase 3
Recombination – CpG content	0.95	0.73
Recombination – CpG methyl.	-0.92	-0.67
Recombination – H3K4me3	0.82	0.70
PRDM9 – CpG methyl.	-0.90	-0.89
PRDM9 – H3K4me3	0.90	0.84

Supplemental Experimental Procedures

Persistent homology estimators of recombination

We considered a Wright-Fisher coalescent model with recombination, characterized by the population-scaled mutation rate $\theta = 4uN_{\text{eff}}$ and the population-scaled recombination rate $\rho = 4rN_{\text{eff}}$, where N_{eff} denotes the effective population size, and u and r are respectively the probabilities of mutation and recombination per individual and generation. Pairwise distances within a set of n sequences sampled from that population are given by the number of segregating sites between each pair of sequences, normalized by the mutation rate u. To each distance matrix we can associate a filtration of simplicial complexes (Edelsbrunner et al., 2002; Zomorodian and Carlsson, 2005), which determines the topologies that are compatible with the distance matrix at any given genetic scale ε (also known as *filtration value*). The first homology group $H_1(\varepsilon)$ of the filtration is a topological invariant that characterizes the 1-dimensional loops associated to the simplicial complex at scale ε. The number b_1 of generators of $H_1(\varepsilon)$, known as *persistent first Betti number*, is thus expected to be proportional to the number of irreducible recombination events associated to the sampled set of sequences (Chan et al., 2013).

The first Betti number of a set of sequences sampled from a Wright-Fisher population is a random variable that follows a Poisson distribution with parameter $\lambda = E[b_1]$, where $E[\,\cdot\,]$ denotes expected value. A closed analytic expression for λ as a function of n, θ and ρ is not known. Our approach was to model λ statistically using an educated ansatz. For that aim, we simulated a large number of coalescent trees and haplotypes from such population for different values of n, θ and ρ, as described in the main manuscript, and estimated λ for each configuration

of values of the above parameters. Simulated data were correctly modeled by equation (1) of the main manuscript (Figure 2A).

To derive equation (4), we expanded ρ_{PH} in powers of b_1

$$\rho_{PH} = \sum_{j=0}^{\infty} m_j \, (b_1)^j$$

and we determined the coefficients m_j by solving the following equation

$$E[\rho_{PH}] = \sum_{j=0}^{\infty} m_j \mu_j = \rho$$

making use of equation (1). In this expression μ_j are the moments of the Poisson distribution around the origin, given by

$$\mu_j = \sum_{i=1}^{j} \lambda^i \left\{ {j \atop i} \right\}$$

and the Stirling numbers of the second kind are defined as

$$\left\{ {j \atop i} \right\} = \frac{1}{i!} \sum_{k=0}^{i} (-1)^{i-k} \binom{i}{k} k^j$$

Solving for m_j and summing over j leads to

$$\rho_{\mathrm{PH}} = \; g \sum_{j=1}^{b_1} \sum_{k=1}^{j} k! \binom{b_1}{j} \left\{ {j \atop k} \right\} \frac{h^{k-1}}{f^j},$$

Equation (4) in the main manuscript results from taking $h = 0$ in this expression.

For completeness we also computed the variance of ρ_{PH}, obtaining

$$\mathrm{Var}[\rho_{PH}] = \frac{g^2}{(1+h)^2}\left\{-\frac{e^{\frac{2E[b_1]}{f}}}{\left[1-h\left(e^{\frac{E[b_1]}{f}}-1\right)\right]^2}+\sum_{j,k=1}^{\infty}\frac{h^{j+k-2}}{(1+h)^{j+k}}e^{\frac{E[b_1]}{f}\left(j+k+\frac{jk}{f}\right)}\right\}$$

Comparison of b_1 to other recombination summaries

We compared b_1 to R_{HK} and R_{MG} using the program `RecMin` (Myers and Griffiths, 2003) with default parameters. We used two sets of 4,000 neutral population simulations of samples with n = 160 sequences, with 14 and 40 segregating sites, and ρ taking values in the range 0 – 160 (corresponding to recombination rates in the range 0 – 160 cM/Mb, for a genomic interval of 1 kb and an effective population size of N_{eff} = 25,000 individuals). To evaluate the performance of each summary as a function of ρ, we used Fisher information,

$$I(\rho) = \int \left(\frac{d}{d\rho}\log f(x|\rho)\right)^2 f(x|\rho)\, dx,$$

where the likelihood function f is the probability of observing value x, given ρ, and the integral is understood to range over all possible values of x. To approximate the information from simulation results, the likelihood f was replaced with $f(x|[\rho-1, \rho+1])$, the probability of observing value x, given that ρ falls in the interval. The derivative was discretized and computed from $\rho - 0.1$ to $\rho + 0.1$. The likelihood was computed considering only the largest range in

which all (integer) x are observed in all three intervals $[\rho - 1.1, \rho + 0.9]$, $[\rho - 1, \rho + 1]$, and $[\rho - 0.9, \rho + 1.1]$. The integral was approximated as a sum over all x in this range.

Comparison to linkage methods at constant ρ

We used the programs `ms` and `seq-gen v1.3.3` to perform 6,500 neutral population simulations of samples with 14 segregating sites. We took constant ρ, taking values in the range 0 – 160 (corresponding to recombination rates in the range 0 – 160 cM/Mb, for a genomic interval of 1 kb and an effective population size of N_{eff} = 25,000 individuals), and n in the range 30 – 170 sampled sequences. The length of the sequences produced by `seq-gen` was chosen such that the expected θ was 0.001 per nucleotide (corresponding to a recombination rate of 10^{-8} mutations per generation per nucleotide, for an effective population size of N_{eff} = 25,000 individuals). We computed ρ_{PH} and run `LDhat v2.2` (http://ldhat.sourceforge.net/) and `PHASE v2.1.1` on each sample. Specifically, the command `lkgen` was used to generate pre-calculated lookup likelihood tables for the populations, followed by the command `interval` with 1,500,000 Markov-Chain Monte Carlo (MCMC) iterations, sampling every 4,000 iterations, using block penalty 25 and discarding the first 100,000 iterations. For samples with $n < 80$ sequences `LDHat pairwise` was also run. `PHASE` was run with the option `-MR3`. Only the 96 % best estimates were kept to discard outliers. Additionally, we performed 1,000 simulations of n = 160 samples with 40 segregating sites, constant ρ, taking values in the range 0 – 160 and run a similar comparison between `LDHat`, `PHASE` and ρ_{PH} based on these simulations.

Comparison to linkage methods at non-constant ρ

We used the software `msHOT` and `seq-gen v1.3.3` to simulate 370 samples of 160 sequences, 35 kbp long, from a neutral population. The background effective population recombination rate was $\rho = 0.00002$ per nucleotide (corresponding to a recombination rate of 0.02 cM/Mb, for an effective population size of $N_{\text{eff}} = 25{,}000$ individuals). We simulated six recombination hotspots of widths 4 kbp, 2 kbp and 1 kbp, and two different intensities, with the local recombination rate enhanced by a factor 640 or 160. Intra-hotpot recombination rate variation was also simulated, with a ½ decay of the local recombination rate at the central region of the hotspot. θ was 0.001 per nucleotide (corresponding to a mutation rate of 10^{-8} mutations per generation and bp, for an effective population size of $N_{\text{eff}} = 25{,}000$ individuals).

We computed ρ_{PH} on a sliding window of size 14 segregating sites moved in steps of 7 segregating sites. The command `rhomap` from `LDhat v2.2` was run with 1,500,000 MCMC iterations, sampling every 2,500 iterations, and discarding the first 100,000 iterations. `PHASE v2.1.1` was run with default parameters. Additionally, 2,000 samples were produced by the above method and ρ_{PH} was computed on a sliding window of fixed size $L = 500$ bp and step 250 bp.

1,000 Genomes Project data

We built recombination maps using data from the 1,000 Genomes Project. The associations described in this work were reproducible using both phase 1 and phase 3 data. Overall, we found that phase 1 data offers better agreement with whole-genome pedigree-based recombination rate estimates as well as stronger associations between recombination enrichment at TF binding sites,

epigenetic markers and predicted PRDM9 binding sites (Table S3). Phase 3 data presents slightly higher correlations with other existing recombination maps in the literature (Figure S2). All statistics presented refer to phase 1 data.

Recombination similarity across human populations

We binned at 10 kbp ρ_{PH} estimates performed with the s = 40 segregating sites window for each of the 7 populations. We only considered bins with an average recombination rate of at least 25,000/N_{eff} cM/Mb in each of the seven maps. We computed pairwise Spearman's correlation coefficients on these bins and built a dendrogram using nearest neighbor algorithm.

We converted to hg19 coordinates and binned Hapmap, African-American and sex-averaged deCODE recombination maps at 10 kbp. We computed pairwise Spearman's correlation for bins that were non-zero in both deCODE and African-American maps.

Recombination map annotation

Genomic coordinates of exons and introns were obtained from the University of California Santa Cruz (UCSC) Genes Track, assembly GRCh37/hg19. TF binding sites were defined by merging the complete set of narrow peak calls for the 188 transcription factors and 91 cell lines analyzed by ENCODE (Gerstein et al., 2012) as of May 2013. Inter-genic regions were defined as genomic regions covered by the 1,000 Genomes Consortium, excluding exons, introns and TF binding sites from them. The coordinates of piRNAs were obtained from piRNA-Bank database (Sai Lakshmi and Agrawal, 2008). piRNAs matching repeated elements were identified by

intersecting piRNA-Bank database with UCSC RepeatMasker Track. piRNA producing clusters were taken from (Ha et al., 2014), keeping only clusters with RPKM counts larger than 15.0. PRDM9 protein binding sites were predicted by searching for the motif CCNCCNTNNCCNC in both strands of the GRCh37/hg19 human genome assembly. This sequence includes binding motifs of the common PRDM9 alleles A and B (Berg et al., 2010).

Processed data on H3K4me3, H3K27me3 marks and histone retention in sperm were taken from (Hammoud et al., 2009) (GEO database accession number GSE15690), whereas sperm methylation profiles and hypo-methylated regions were taken from (Molaro et al., 2011) (accession number GSE30340). Methylation profiles for primordial germ cells were taken from (Gkountela et al., 2015) (accession number GSE63393). Methylated CpG di-nucleotides were defined with a lower threshold of 70 % of the reads corresponding to the methylated state.

BED files were merged and/or intersected when needed by making use of `BEDTools v.2.19`. When required, coordinates were converted to GRCh37/hg19 coordinates by making use of the UCSC tool `LiftOver`.

Estimation of relative recombination rates

Recombination enrichments at genomic and epigenetically marked regions (Figures 4B, 5A, S3A and Table 1) were estimated by counting (with the 500bp window) maxima of ρ_{PH} within the region of interest and normalizing by its nucleotide length. We used the GBR population for that aim. Statistical significances and error bars were estimated by counting maxima that lie inside (N_{in}) and outside (N_{out}) the region of interest, and performing a log-likelihood ratio test under the

assumption that both counts (in and out) are Poisson distributed with exposure equal to their total nucleotide lengths (L_{in} and L_{out}, respectively),

$$D = 2\log\frac{\Pr(X=N_{in}|N_{in})\Pr(Y=N_{out}|N_{out})}{\Pr(X=N_{in}|\mu L_{in})\Pr(Y=N_{out}|\mu L_{out})} \quad (5)$$

$\Pr(X=k|\lambda)$ denotes the Poisson probability mass function, $\mu = (N_{\text{in}}+N_{\text{out}})/(L_{\text{in}}+L_{\text{out}})$ and D approximately follows a χ^2 distribution with one degree of freedom.

Recombination enrichments at loci with respect to neighboring regions (Figures 4A, 4C, 4E, 4F, 5B, 5C, S3B, S3D, S3E, S4A and S4B) were obtained by measuring the density of maxima of ρ_{PH} within a region 500 bp wide around the center of the corresponding elements (TF binding sites or piRNAs), and comparing with the density of maxima within regions 500bp wide located at a distance of 4kbp away from the center of the element. Statistical significance was assessed by means of Student's t-test. Only combinations of ENCODE transcription factors and cell lines with at least 6,000 binding sites in the cell line were considered. Statistical significances were adjusted for multiple testing using Benjamini-Hochberg procedure for controlling the false discovery rate.

Relative CpG odds ratio enhancement

We defined the ratio

$$r = \frac{\#(\text{CpG sites})}{\#(\text{G sites})\#(\text{C sites})} \quad (6)$$

To estimate the relative CpG odds ratio enhancement (Figures S3C and S4C), we computed *r* within a region 500 bp wide around the center of the corresponding elements (TF binding sites or piRNA), and compared with the value of *r* within 500 bp regions located at a distance of 4 kbp away from the center of the element.

Recombination maps availability

Recombination maps are available at http://rabadan.c2b2.columbia.edu/cgi-bin/hgGateway?hgsid=256902&clade=mammal&org=Human&db=0.

Supplemental References

Berg, I.L., Neumann, R., Lam, K.W., Sarbajna, S., Odenthal-Hesse, L., May, C.A., and Jeffreys, A.J. (2010). PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet *42*, 859-863.