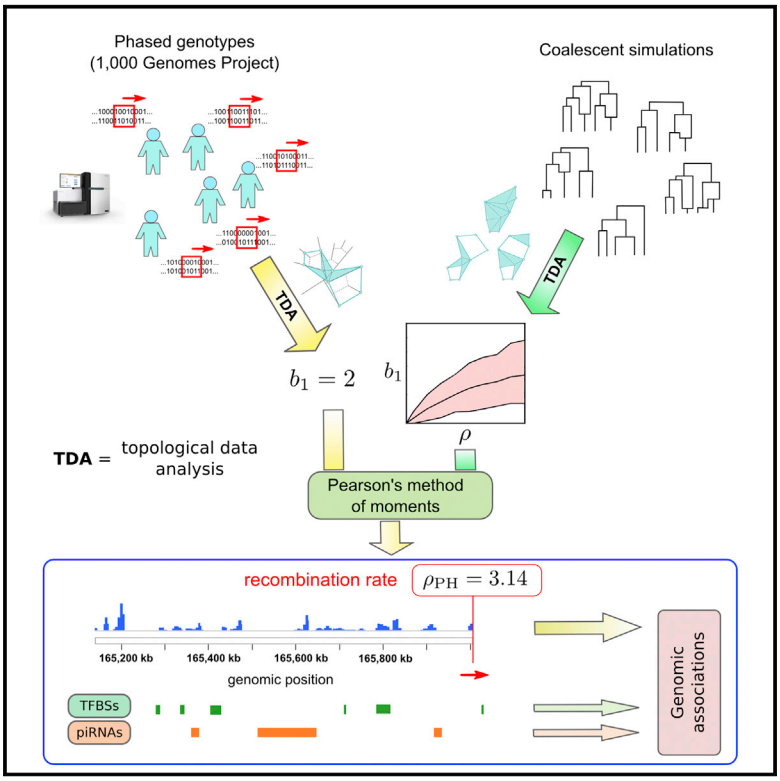
Article

Cell Systems

Topological Data Analysis Generates High-Resolution, Genome-wide Maps of Human Recombination

Graphical Abstract



Highlights

- Topological data analysis captures recombination from large genomic samples
- High-resolution recombination maps of seven human populations are presented
- Binding sites of specific transcription factors are enriched for recombination
- Repeat-derived loci matched by piwi-interacting RNAs are enriched for recombination

Authors

Pablo G. Camara,
Daniel I.S. Rosenbloom,
Kevin J. Emmett, Arnold J. Levine,
Raul Rabadan

Correspondence

pg2495@cumc.columbia.edu (P.G.C.),
rr2579@cumc.columbia.edu (R.R.)

In Brief

Camara et al. introduce a new method to estimate recombination rates from large genomic samples and present high-resolution recombination maps of seven human populations. Using these maps, they show evidence of previously unreported associations of recombination with binding sites of specific transcription factors and with repeat-derived loci matched by piRNAs.

Camara et al., 2016, Cell Systems *3*, 83–94
July 27, 2016
http://dx.doi.org/10.1016/j.cels.2016.05.008

Cell Systems
Article

Topological Data Analysis Generates High-Resolution, Genome-wide Maps of Human Recombination

Pablo G. Camara,[1,2,*] Daniel I.S. Rosenbloom,[1,2] Kevin J. Emmett,[1,3] Arnold J. Levine,[4] and Raul Rabadan[1,2,*]
[1]Department of Systems Biology
[2]Department of Biomedical Informatics
Columbia University College of Physicians and Surgeons, 1130 St. Nicholas Avenue, New York, NY 10032, USA
[3]Department of Physics, Columbia University, New York, NY 10027, USA
[4]The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540, USA
*Correspondence: pg2495@cumc.columbia.edu (P.G.C.), rr2579@cumc.columbia.edu (R.R.)
http://dx.doi.org/10.1016/j.cels.2016.05.008

SUMMARY

Meiotic recombination is a fundamental evolutionary process driving diversity in eukaryotes. In mammals, recombination is known to occur preferentially at specific genomic regions. Using topological data analysis (TDA), a branch of applied topology that extracts global features from large data sets, we developed an efficient method for mapping recombination at fine scales. When compared to standard linkage-based methods, TDA can deal with a larger number of SNPs and genomes without incurring prohibitive computational costs. We applied TDA to 1,000 Genomes Project data and constructed high-resolution whole-genome recombination maps of seven human populations. Our analysis shows that recombination is generally under-represented within transcription start sites. However, the binding sites of specific transcription factors are enriched for sites of recombination. These include transcription factors that regulate the expression of meiosis- and gametogenesis-specific genes, cell cycle progression, and differentiation blockage. Additionally, our analysis identifies an enrichment for sites of recombination at repeat-derived loci matched by piwi-interacting RNAs.

INTRODUCTION

The maintenance of genetic diversity in a species can promote survival during times of unpredictable environmental change. Germline mutations, inherited by the offspring, are the raw material of genetic diversity in sexually reproducing organisms. Meiotic recombination enables a population to explore and maintain this genetic diversity by allowing for the rapid generation of new allele combinations. An excessive amount of genetic linkage due to insufficient meiotic recombination can preclude removal of deleterious variants from the genome over successive generations, leading to a substantial fitness reduction.

Meiotic recombination is initiated by the induction of programmed DNA double-strand breaks (DSBs) during meiosis. These initiating lesions can be repaired through various pathways involving the formation of heteroduplex DNA. Consequently, meiotic recombination is usually accompanied by GC-biased gene conversion tracts (Duret and Galtier, 2009). Additionally, some of the repair pathways lead to the formation of chromosomal crossovers, required for proper chromosomal disjunction (Koehler et al., 1996). The aggregate effect of all these biochemical processes over evolutionary time defines the recombination landscape of the genome.

Studies of the recombination landscape in eukaryotes have revealed that recombination is highly regulated, with ~80% of recombination events in humans occurring at narrow (~2 kb) regions known as recombination hot spots (Crawford et al., 2004; Kauppi et al., 2004; McVean et al., 2004; Myers et al., 2005). The enrichment of non-allelic homologous recombination (NAHR) variants observed at recombination hot spots (Mills et al., 2011) suggests that the programmed DSBs required for recombination initiation can also serve as a source for NAHR. Regulation of the location and frequency of recombination can therefore potentially reduce the impact of DSB repair problems and target recombination to genomic regions where genetic diversity is more advantageous.

The specific biological mechanisms that regulate meiotic recombination are largely unknown. In mammals, the meiosis-specific histone 3 lysine 4 (H3K4) tri-methyltransferase PRDM9 binds chromosomes at recombination hot spots through a tandem-array of C2H2 zinc-fingers that recognizes a specific DNA binding motif (Baudat et al., 2010; Myers et al., 2010; Parvanov et al., 2010). In *Prdm9* knockout mice, meiotic DSBs occur at different locations than in wild-type, suggesting that this gene determines the location of meiotic DSBs (Brick et al., 2012). Accordingly, variants of the *Prdm9* gene coding different numbers of zinc-finger domains are associated with variation in hot spot location, both between human populations and between mammalian species (Baudat et al., 2010; Myers et al., 2010), as well as between human individuals (Pratto et al., 2014). It is an open question whether factors other than PRDM9 modulate recombination in mammals.

Several methods have been proposed to study the landscape of recombination (Kirkness et al., 2013; Lu et al., 2012; McVean

et al., 2004; Pan et al., 2011; Pratto et al., 2014; Smagulova et al., 2011; Wang et al., 2012). Population-based recombination maps capture the recombination history of populations using genome-wide genomic data and have become a valuable tool in the study of human recombination during the last decade (Hinch et al., 2011; Frazer et al., 2007; Kong et al., 2010; Myers et al., 2005). Sub-kilobase scale mapping and annotation of human recombination is now possible due to the large number of genomes published by consortia such as the 1,000 Genomes Project (Abecasis et al., 2012) and ENCODE (ENCODE Project Consortium, 2012). Nucleotide-resolution data sets, such as those obtained by chromatin immunoprecipitation (ChIP-seq), bisulfite, or RNA sequencing methods, reveal a gamut of biological features associated to small genomic regions, often spanning mere handfuls of bases. How these fine-scale nucleotide-level features influence the structure and position of recombination hot spots is not understood.

A key step toward this understanding is the development of methods that can accurately estimate fine-scale meiotic recombination rates genome-wide, so that relationships with narrow (and often clustered) biological features of the genome can be assessed statistically. Such high-resolution recombination maps are only attainable through the analysis of large numbers of sequences and segregating sites, becoming an important challenge for current methods of recombination rate estimation. Widely used methods (Crawford et al., 2004; Hudson, 2001; Li and Stephens, 2003; McVean et al., 2002) are based on the non-random association of alleles at different loci, that is, linkage disequilibrium. Analysis based on these methods, however, becomes computationally expensive when the number of sequences is on the order of 100. New mathematical and computational approaches are needed to meet this challenge.

Topological data analysis (TDA) is a new branch of applied topology that extracts global features from large data sets. TDA has been successfully utilized in cross-sectional studies of complex genetic diseases (Li et al., 2015) and cancer (Nicolau et al., 2011). Persistent homology, a framework within TDA for deriving and classifying topological features associated to data (discussed in detail below), has been shown to capture instances of recombination and reassortment in viral populations (Chan et al., 2013). These results suggest that it may be also used for quantifying recombination in human populations.

Here, we introduce an estimator of recombination rates at fine scales (0.5–1 kb) that uses persistent homology and is tailored to the analysis of very large genomic samples. We make use of this estimator to build fine-scale recombination maps of seven human populations sequenced by the 1,000 Genomes Project. Comparison of these recombination maps with recent fine-scale annotations of the human genome (Gerstein et al., 2012; Sai Lakshmi and Agrawal, 2008) reveals that although transcription start sites are generally depleted for recombination (Coop et al., 2008; Lu et al., 2012), specific transcription factor (TF) binding sites are frequently associated with PRDM9 binding motifs and recombination. These include TFs that regulate expression of meiosis- and gametogenesis-specific genes, cell cycle progression, and differentiation blockage. We also observe that repeat-derived loci targeted by piwi-interacting RNAs (piRNAs), coding some recent families of transposable elements known to be expressed during gametogenesis (Guo et al., 2015) and early embryogenesis (Smith et al., 2014), are also enriched for recombination.

RESULTS

Topology and Evolution

Topology is the branch of mathematics concerned with properties of spaces that are preserved under continuous deformations (deformations that do not involve cutting or pasting), such as the number of loops or connected components of a space. For example, a "B"-shaped space can be continuously deformed into an "8"-shaped space without changing its topology (Figure 1A). A frequently used approach in topology is the replacement of the original space by a simpler one, known as a simplicial complex, which has the same topological features as the original space, but consists of a finite set of elements (Figure 1B). A simplicial complex is a generalization of a network that, in addition to nodes and edges, includes higher dimensional elements like triangles and tetrahedra. Simplicial complexes are powerful because they allow the implementation of algebraic operations to extract the topological features of the space. These topological features of a space can be arranged in homology groups, that is, algebraic structures that encompass and classify all gaps or holes in the space. Note that throughout this work, the term homology refers to topological homology, which is unrelated to the notion of sequence homology. Elements of the 0^{th} homology group correspond to disconnected parts of the space; elements of the first homology group correspond to loops; elements of the second homology group correspond to hollow spheres, and, in general, elements of the n^{th} homology group correspond to $(n+1)$-dimensional voids of the space. The number of independent elements of the n^{th} homology group is called the n^{th} Betti number. For instance, the first Betti number of an "8"-shaped space is 2 (Figure 1A). We refer to Ghrist (2014) and Hatcher (2002) for an extended introduction to the basic concepts of algebraic topology.

Motivated by the fact that actual data are rarely given in the form of topological spaces, recent mathematical developments have expanded the realm of algebraic topology to point cloud data, that is, any set of data points with a notion of distance between them (Carlsson, 2009). Starting from a set of points sampled from an unknown space, TDA aims to infer the topological features of the underlying space (Figure 1C). TDA provides the necessary tools to build simplicial complexes starting from point cloud data. One such construction builds a simplicial complex by taking balls of radius ε, centered on the data points. If two balls intersect, the points at the center of the balls are connected in the simplicial complex. In this way, there is a simplicial complex (and hence a set of topological features) associated to the data at each value of ε. Tracking how homology groups change with ε permits their generalization to point cloud data. The resulting mathematical structures are known as persistent homology groups (Edelsbrunner et al., 2002; Zomorodian and Carlsson, 2005).

TDA can be used to infer evolutionary relations from a sample of genomic sequences (Chan et al., 2013). We consider high-dimensional spaces where each point corresponds to a genomic sequence and distances between points are given by the genetic

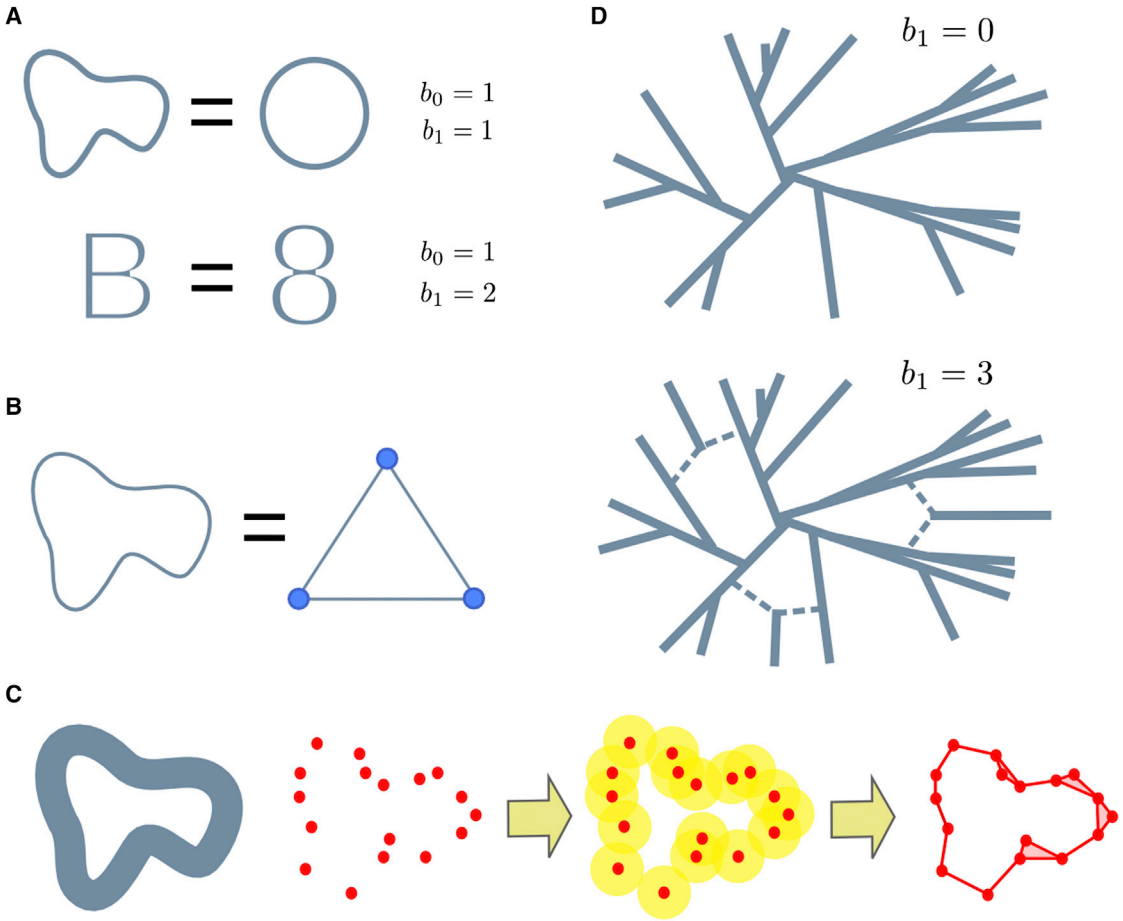


Figure 1. Topology and Evolution

(A) Topology is concerned with properties of objects that are invariant under continuous deformations. For instance, a "B"-shaped space can be continuously deformed into an "8"-shaped space. Both have one connected piece and two inequivalent loops. These topological invariants are counted by Betti numbers, b_n. Similarly, a circumference always has one connected component and a loop, no matter how it is deformed, as long as nothing is cut or pasted.

(B) A prominent tool in algebraic topology is simplicial complexes. These are finite set representations of the original space that share the same topology. Here, we present a simplicial complex that describes the topology of a circumference. The simplicial complex is given in terms of a finite set of elements (three points and three segments). Algebraic operations on the simplicial complex can extract the topological features of the original circumference.

(C) TDA infers the topological features of a space from a finite set of sampled points by assigning simplicial complexes to the data. One such construction consists of taking balls of fixed radius ε centered on the points. Points at the center of intersecting balls are connected in the simplicial complex. From the resulting complex, it is possible to extract topological features associated to the data at scale ε.

(D) In the context of evolution, genomic sequences can be represented as points in a high dimensional space, where the distance between points is given by the Hamming distance between the corresponding sequences. In the absence of back-mutation and recombination, subsequent mutations can only increase the distance between genomes, and the evolutionary space of the system does not have loops. When recombination events are present, the evolutionary space contains loops, whose presence can be inferred from the finite genomic sample using TDA methods.

distance (e.g., Hamming distance) between sequences. The evolutionary history of a sample of genomic sequences can be represented as such a space, consisting of all the genomic sequences that occur from the most common recent ancestor of the sample to the present. Assuming that each genomic site mutates at most once across the entire sample history, the genetic distance between two sequences can only increase with the acquisition of new mutations (Figure 1D). Hence, the only way of "closing" a loop in this space is by means of a recombination event. In populations evolving clonally without recombination or back-mutation, the first Betti number of the evolutionary space of the sample is zero (Chan et al., 2013), as a phylogenetic tree suffices to describe ancestry. More generally, the number of loops of the evolutionary space is related to the number of recombination events in the sample history. Since we only have access to a sample of points, we can make use of the persistent first Betti number (b_1) of the sample to infer the amount of recombination in the sample history. In what follows, we make use of this approach to build an estimator of recombination. As noted, our estimator does not rely on genetic linkage, but rather on the topology of spaces formed by genomic sequences.

Persistent Homology Estimator of Recombination: ρ_{PH}

A suitable approach to recombination rate estimation in very large data sets is the use of estimators based on summary statistics of the data (Wall, 2000). As we have argued, the persistent

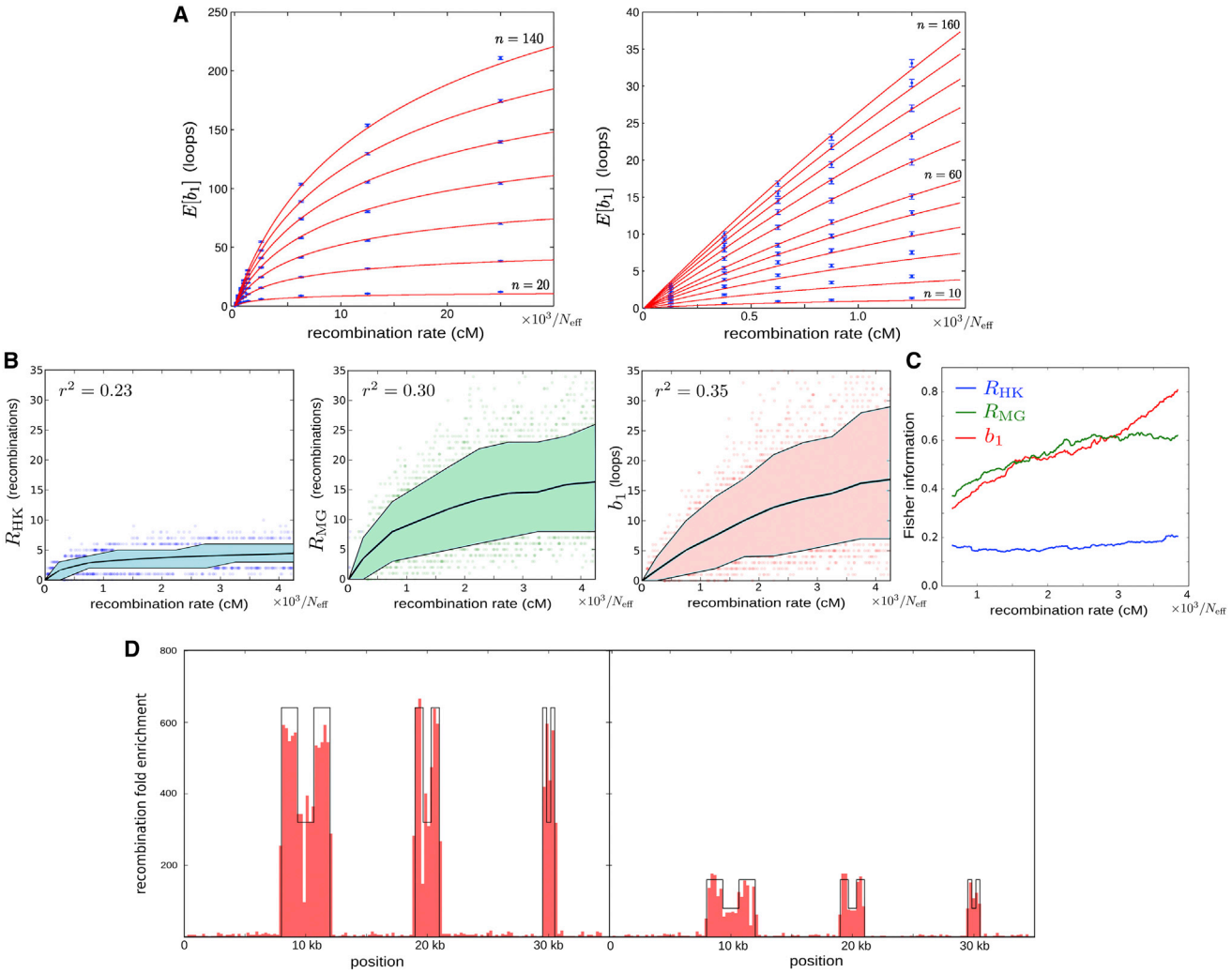


Figure 2. Persistent Homology Estimator of Recombination

(A) Estimated expected first Betti number $E[b_1]$ as a function of the recombination rate (expressed in terms of the effective population size N_{eff}) and sample size n. Each point is based on 500 simulations. The error bars represent 95% confidence level intervals. The red curves correspond to the best fit according to Equation 1 in Experimental Procedures.

(B) Dependence of Hudson-Kaplan (left), Myers-Griffiths (center), and b_1 (right) summaries of recombination on the recombination rate at a fixed number of segregating sites (s = 14). Each plot is based on 4,000 coalescent simulations of a sample of 160 sequences. The colored bands represent the interdecile range and the central lines correspond to the mean. The squared Pearson's correlation coefficient is shown in each case.

(C) Fisher information for each of the three summaries in (B) 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.

(D) Distribution of 500 bp segments with $\rho_{\text{PH}} > 0$ in simulated samples of 160 sequences, 35 kbp long. The background recombination rate is $500/N_{\text{eff}}$ cM/Mb. The six recombination hot spots of widths 4 kbp, 2 kbp, and 1 kbp are simulated. The local recombination rate is enhanced at hot spots by a factor 640 (left) and 160 (right). Intra-hot spot recombination rate variation is also simulated, with a 1/2 decay of the local recombination rate at the central region of hot spots.

See also Figure S1 and Table S1.

first Betti number of a sample of genomic sequences is expected to be a concise mathematical summary of recombination. To check this hypothesis, we performed extensive coalescent simulations with recombination (Table S1) and observed that the expected value of b_1 increases monotonically with the population recombination rate parameter ρ (Figure 2A), which is defined as four times the product of the effective population size (N_{eff}, the number of diploid individuals in a coalescent simulation that produces a level of genetic diversity similar to that of the population of interest) and the per-meiosis recombination rate. At low values of this parameter, b_1 is proportional to ρ. Intuitively, this behavior is expected, as the number of topology-changing events in coalescent models of evolution scales as $\rho \log(n)$ for large n, where n is the number of sequences in the sample (Hein et al., 2004). At large values, b_1 saturates due to the limit in the number of loops that a finite set of sequences can generate. We found that this behavior of b_1 is well described by a logarithmic function (Equation 1 in Experimental Procedures and Figure 2A).

To evaluate the utility of b_1 as a summary of recombination, we compared it to other summaries of recombination available in the

literature. Specifically, we considered the lower bounds on the minimum number of recombination events of a sample history developed in Hudson and Kaplan (1985) and Myers and Griffiths (2003). We refer to those as R_{HK} and R_{MG}, respectively. These summaries combine simple local bounds across different genomic regions to generate a more stringent global bound. Based on simulated samples of sequences with a small number of segregating sites (s = 14), we observed that among the three summaries, b_1 has the largest fraction of variance explained by the recombination rate (Figure 2B). In particular, b_1 has better sensitivity to high recombination rates, as it does not saturate as early as other summaries of recombination. This conclusion was also confirmed by comparing Fisher information of the three summaries as a function of the recombination rate (Figure 2C). For samples with a large number of segregating sites (s = 40), the performance of R_{MG} and b_1 was comparable, with R_{MG} having a larger fraction of explained variance than b_1 (Figure S1A), but similar Fisher information (Figure S1B).

To investigate whether saturation of b_1 occurs in practical applications, we assumed a mutation rate of 10^{-8} mutations per base per generation and an effective population size of N_{eff} = 25,000 individuals. With these assumptions, saturation effects for a sample of 200 sequences became important at genetic map distances above ~0.2 centimorgans (cM) over 2.6 kb (for s = 14) or ~0.4 cM over 7.5 kb (for s = 40). These recombination rates are rarely found in the human genome (Kong et al., 2010).

Taking these results together, we concluded that b_1 is a robust summary of recombination at fine scales, where the number of segregating sites is small, and used Pearson's method of moments to build an estimator of recombination rate, ρ_{PH}, based on b_1 (Supplemental Information).

Comparison to Linkage Methods

Currently, the most commonly used methods to estimate recombination rates are based on linkage disequilibrium, as defined above. Practical methods implement Markov chain Monte Carlo algorithms to approximate a likelihood function, built from the observed genetic linkage between pairs of sites (Hudson, 2001; McVean et al., 2002) or from the reconstructed segments of exchanged genetic material (Crawford et al., 2004; Li and Stephens, 2003). These methods have great accuracy, producing low-variance unbiased recombination rate estimates, but their applicability to very large data sets can be hindered by their computational cost.

We compared our estimator, ρ_{PH}, to the ones generated by widely used software packages LDhat (McVean et al., 2002, 2004) and PHASE (Crawford et al., 2004; Li and Stephens, 2003). The three estimators produced comparable results on simulated data at constant recombination rate and a fixed number (s = 14) of segregating sites (Figure S1C), observing some advantage of PHASE over the other two estimators in terms of accuracy. Our estimator, ρ_{PH}, however, was on average 12 times faster than LDhat and 30 times faster than PHASE (Figure S1D). At a large number (s = 40) of segregating sites, both LDhat and PHASE offered some advantage over ρ_{PH} in terms of precision (Figure S1C), although ρ_{PH} was still 2–6 times faster than these estimators.

We evaluated the power of ρ_{PH} to resolve variation in the recombination rate across narrow genomic loci, using simulations at non-constant recombination rate. Our estimator ρ_{PH} produced lower variance estimates than LDhat and PHASE on a sliding window of variable length and a fixed number (s = 14) of segregating sites (Figure S1E). To enhance the sensitivity of our estimator to variation in the recombination rate, we implemented a second sliding window with a fixed length (L = 500 bp). Counting the number of times that each 500 bp segment had $\rho_{PH} > 0$, over multiple simulations, was sensitive to relative fine-scale variation in the recombination rate, allowing for detection of sub-kilobase scale variations (Figure 2D). In these simulations, our persistent homology estimator was 150–1,500 times faster than LDhat and PHASE, making it uniquely suited to very large genomic samples. We exploited this property of our approach and built recombination maps across human populations from the 1,000 Genomes Project.

Recombination Maps of Seven Human Populations

We built recombination maps of seven human populations (one African, two Asian, and four of European ancestry; Table S2), using phased genotype data of ~38 million SNPs from the 1,000 Genomes Project (Abecasis et al., 2012). Our data set included a total of 647 individuals. We scanned the entire genome three times for each population, using two sliding windows with a fixed number of segregating sites (s = 14 and 40) and a sliding window with a fixed length (L = 500 bp). For each window and genomic position, a Hamming distance matrix was obtained, from which b_1 and ρ_{PH} were computed. Windows with a large number of segregating sites give accurate estimates of the recombination rate over relatively large (5–10 kbp) genomic intervals; whereas windows with a small number of segregating sites provide information about the precise genomic location of recombination events. As a first consistency check of our method, we detected no signature of recombination along the Y and mitochondrial chromosomes with either window, as expected from the predominantly uniparental inheritance of these chromosomes. Nevertheless, the relative dearth of SNPs in these chromosomes may also reduce detection sensitivity.

A snapshot of the output produced by this method is shown in Figure 3A. Median detected recombination rates for non-African populations are ~15,000/N_{eff} cM/Mb and the highest 10% of recombination rates are $\geq$100,000/N_{eff} cM/Mb (Table S2). Population recombination rates ρ were approximately doubled in the African population. This is consistent with a larger effective population size and an out-of-Africa human expansion model (Templeton, 2002). Our method was therefore able to capture the expected differences in population recombination rates due to known population structure effects.

To check the consistency of the TDA approach across different data sets, we performed a pairwise comparison between the seven maps. We found a high degree of consistency between the location and intensities of recombination peaks identified in distant populations (Figure 3). Position-dependent recombination rates were correlated between pairs of populations, with Spearman's r ranging from 0.53 to 0.78. Hierarchical clustering of populations based on these correlations followed known ancestral relationships, with African, Asian, and European populations grouped in different clusters (Figure 3B).

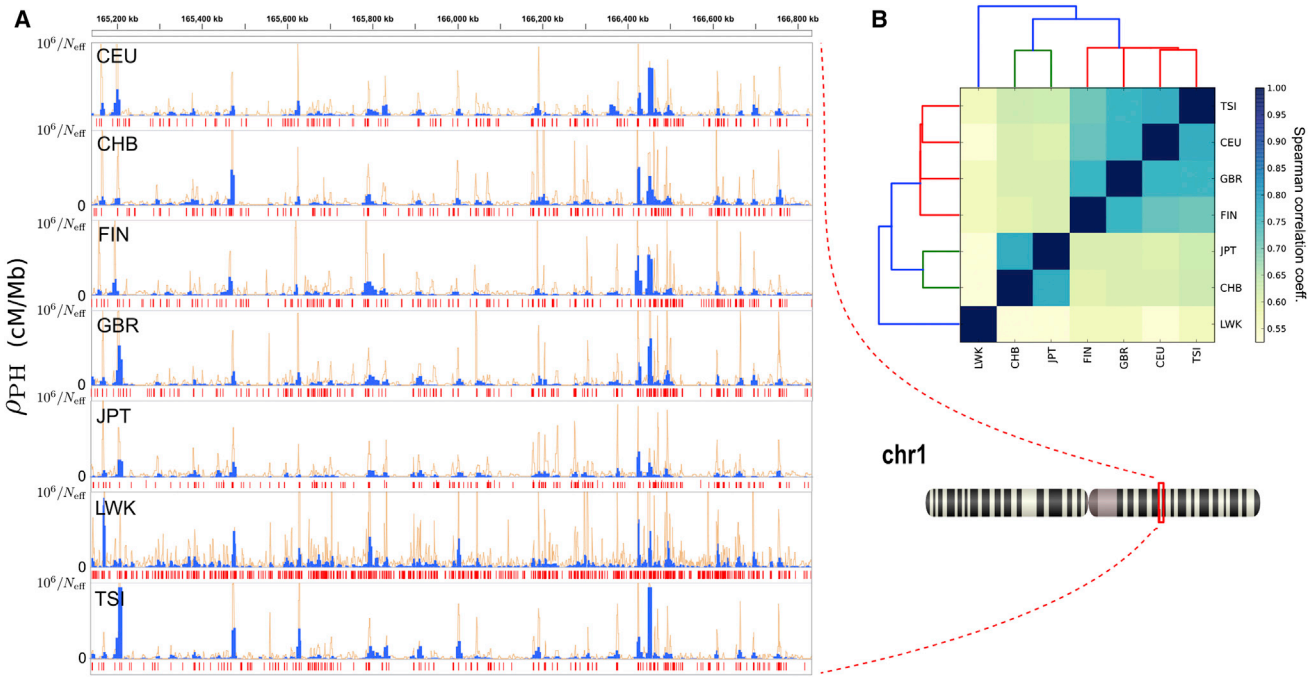


Figure 3. Recombination Rate Estimates across Distant Human Populations

(A) Position-wise recombination rates for each of the seven populations, for the cytogenic band 1q24.1. The blue and orange line plots represent recombination rates estimated with sliding windows with a fixed number of segregating sites (s = 40 and s = 14, respectively). Below each track, the red segments represent genomic regions where a 500 bp sliding window detects recombination ($b_1 > 0$).

(B) Spearman correlation matrix for the position-wise recombination rate of different populations across the entire genome. The maps were binned at 10 kbp and correlation was computed for bins with an average recombination rate of at least 25,000/N_{eff} cM/Mb in each of the maps. The tiles are colored according to the degree of correlation. Hierarchical clustering of the matrix components is also shown, with colored leaves corresponding to African (blue), Asian (green), and European (red) populations.

See also Figure S2 and Table S2.

We also compared our recombination maps to other maps in the literature. Specifically, we considered the deCODE map (Kong et al., 2010), based on a half million crossovers identified in 15,000 Icelandic meioses; the African-American (AA) map (Hinch et al., 2011), based on more than two million crossovers in 30,000 unrelated AA; and the HapMap recombination map (Frazer et al., 2007), based on linkage disequilibrium breakdown using three million SNPs genotype data. Although the nature, content, and underlying assumptions of each of these maps differ from each other, comparison with our recombination maps across ~300 kbp regions within the major histocompatibility complex and the MS32 mini-satellite loci revealed a large degree of consistency between different maps (Figures S2A and S2B). To perform a more quantitative comparison, we binned all maps at 10 kbp. Whole-genome Spearman's correlation with our recombination maps was in the range 0.54–0.63, 0.53–0.61, and 0.43–0.48, respectively for HapMap, AA, and deCODE maps. These correlations were comparable to those observed between HapMap and deCODE (r = 0.60) and between deCODE and AA (r = 0.62) maps. Furthermore, recombination rates at exons, introns, and intergenic regions matched those observed in pedigree-based studies (Table 1), providing additional consistency checks of our maps.

From these results, we inferred the merit of the high-resolution recombination maps produced by TDA.

Recombination Enrichment at TF Binding Sites

Because we hypothesized that the human recombination landscape is largely affected by the epigenome and transcriptome of meiotic germ cells, we focused our analysis on the loci of TF binding sites and piRNAs. We considered the DNA binding sites of 118 TFs detected by ChIP-seq in at least one of 91 cell lines studied by the ENCODE Analysis Working Group (Gerstein et al., 2012). For each cell line and TF, we computed the recombination enrichment across binding sites, observing little variation across cell lines. We found TF binding sites to be, on average, depleted of recombination with respect to the whole-genome average (fold enrichment [FE] = 0.96, $p < 10^{-37}$). This observation is consistent with previous work showing a depletion of recombination at transcription start sites (Coop et al., 2008; Lu et al., 2012).

Disaggregating TF binding sites by type of TF, we identified systematic differences across TF families (Figures 4A, 4B, S3A, and S3B). Several types of TF binding sites show significant recombination enrichments with respect to the whole-genome average (log-likelihood ratio test as described in Supplemental Information, Benjamini-Hochberg adjusted $p < 10^{-5}$). These binding sites are found to associate mostly with TFs that have a bias toward proximal promoters according to the ENCODE categorization (Gerstein et al., 2012). Some of those TFs are members of the E2F family, with key roles in the regulation of

Table 1. Recombination Rate Estimates across Different Genomic Loci and Comparison to Pedigree-Based Estimates

Type	Recombination Rate, TDA (FE)	Recombination Rate, deCODE (FE)[a]
Exon	0.84 ± 0.02	0.85
Intron	0.99 ± 0.01	1.02
Intergenic	1.02 ± 0.01	1.03

[a]Data from Kong et al. (2010).

cell cycle progression and differentiation blockage (DeGregori and Johnson, 2006). In particular, promoters containing binding sites of the transcriptional repressor E2F6, regulating the expression of meiosis-specific genes (Kehoe et al., 2008; Velasco et al., 2010), were found to be enriched for recombination (Figures 4B and 4C). In addition, binding sites of RNA polymerase II and regulatory subunits of MLL1/2 protein complexes were also found to be enriched for recombination (FE 1.28–1.68, p $\sim 10^{-2}$–10^{-7}) (Figure 4B). These results were independently confirmed using HapMap and AA recombination maps, despite the lower resolution of these maps (Figure 4C).

The above analysis demonstrates that instances of recombination are enriched within the binding sites of specific TF families. This observation is in accord with the local nucleotide sequence. We observed a strong association between recombination and CpG content at these sites (Pearson's $r = 0.95$, $p < 10^{-50}$); the observed CpG abundance within regions of recombination enrichment is partially explained by local enrichments for G and C nucleotides (Figure S3C). CpG enrichment at highly recombinant regions is thought to occur through biased gene conversion in the repair of meiotic DSBs (Duret and Galtier, 2009). In addition, we observed a statistically significant association of recombination at TF binding sites with predicted PRDM9 binding sites at these loci (Pearson's $r = 0.96$, $p < 10^{-50}$) (Figures 4D and S3B), suggesting that PRDM9 drives recombination toward these loci during meiosis.

Motivated by our findings on the type of TF binding sites that are enriched for recombination, we decided to interrogate whether these binding sites are part of active promoters in the germline. To that end, we assessed the DNA methylation state of these sites in sperm (Molaro et al., 2011) and human primordial germ cells (PGCs) (Gkountela et al., 2015). Our analysis revealed a statistically significant association (Pearson's $r = 0.92$, $p < 10^{-50}$) between recombination at TF binding sites and CpG hypo-methylation of these loci in sperm (Figures 4D, 4E, and S3B). Hypo-methylation was also detected at earlier stages of gametogenesis, with similar associations in PGCs of male 19.5-week- and female 16.1-week-old embryos (Figure S3D).

DNA hypo-methylated TF binding sites in germ cells and embryonic stem cells have been related to bivalent developmental gene promoters (Bernstein et al., 2006; Hammoud et al., 2009), characterized by simultaneous H3 Lys-4 and Lys-27 tri-methylation (H3K4me3 and H3K27me3) marks on their nucleosomes. The simultaneous presence of these marks defines a poised transcriptional state for the promoters, which remain off until the appropriate developmental stage. To find whether recombination-enriched TF binding sites are part of bivalent promoters in the germline, we considered the profile of these epigenetic marks in sperm (Hammoud et al., 2009). Whereas male germ cells are mostly depleted of nucleosomes after meiosis, we found that loci of recombination-enriched TF binding sites retain nucleosomes (Figure S3E). In addition, our analysis revealed a statically significant association between recombination at TF binding sites and simultaneous H3K4me3 (Pearson's $r = 0.82$, $p < 10^{-50}$) and H3K27me3 (Pearson's $r = 0.80$, $p < 10^{-50}$) marks in sperm (Figure 4F). Taken together, these results suggest that PRDM9 drives meiotic recombination toward certain active or poised promoters in germ cells.

Recombination Enrichment at Loci Targeted by piRNA

Inspired by the association between developmental promoters and PRDM9-mediated recombination at these loci, we decided to explore recombination rates at the loci of other important transcriptional regulators in the germline. piRNAs attain their broadest expression in germ cells and have a central role in post-transcriptional regulation (Watanabe et al., 2015) and transposon control (Aravin et al., 2007). We studied recombination across genomic loci with 100% sequence identity to known human piRNA sequences (Sai Lakshmi and Agrawal, 2008). We observed an enrichment for recombination (FE = 2.67, $p < 10^{-49}$) at these loci with respect to the whole-genome average. The enrichment was also present, although reduced (FE = 1.31, p $\sim 10^{-9}$), when restricting to piRNA-producing clusters as defined and annotated by Ha et al. (2014).

A large fraction of piRNAs are repeat-derived. We estimated the enrichment for recombination at piRNA-matched loci derived from repetitive elements (Figure 5A). All main families (LTR, LINE, and SINE) are enriched for recombination compared to neighboring genomic regions. In terms of specific transposable elements, some of the most recent L1 elements (L1Hs and L1PA1-4), human endogenous retroviruses (HERVK, HERVH, and HERVL), LTRs (LTR12C and MER11C), and most Alu and SVA elements present the highest recombination rates. These elements have been found to be expressed during germ cell development (Guo et al., 2015) and early embryogenesis (Smith et al., 2014). We observed a systematic pattern for the recombination rate at these loci, peaking in most cases at the 5′- and 3′-ends of the transposon, with the 5′-end presenting the highest recombination rate (Figure 5B). Our analysis also reproduced known recombination enrichments at THE1A/B repeat elements (Myers et al., 2005) (Figure S4A). These findings were also confirmed using linkage-based methods on 1,000 Genomes Project data (Figure S4B).

Being aware that repetitive elements are particularly prone to misalignment, we looked for additional evidence supporting the enrichment for recombination at these loci. We found that the enrichment is also accompanied by conserved PRDM9 binding motifs (Figures 5B and 5C). Furthermore, recombination at these loci is strongly correlated with CpG abundance (Figures 5B and S4C). As occurred with TF binding sites, a large fraction of the CpG abundance is explained by a local enrichment for G and C nucleotides, being suggestive of extensive biased gene conversion from meiotic DSB repair. All these features added strong support to the observed recombination enrichment at repeat-derived loci targeted by piRNA, and were different in the case of uniquely mapped piRNAs (single locus piRNA genes) (Figure 5C), suggesting that only repeat-derived loci are enriched for recombination.

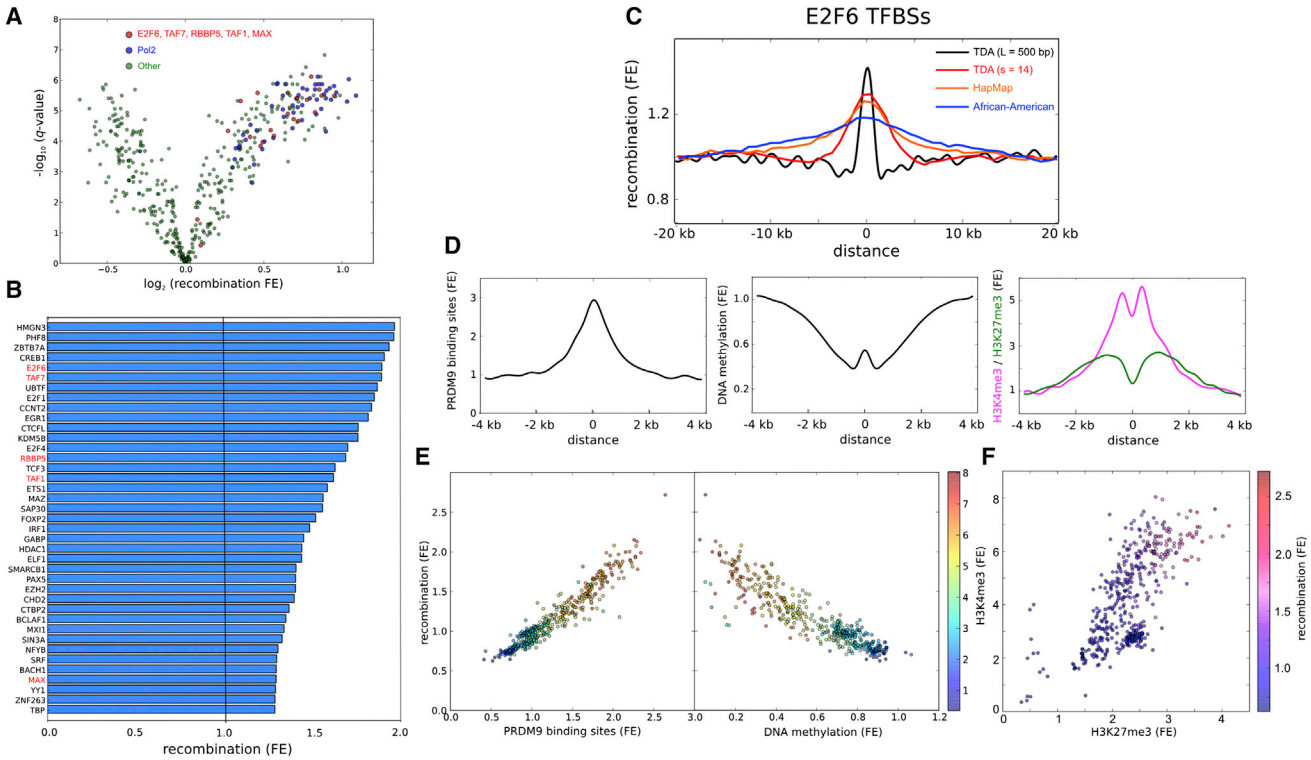


Figure 4. Relation of Recombination to TF Binding Sites

(A) Recombination enrichments at TF binding sites. Each point corresponds to a different combination of TF and cell line. The binding sites are based on ChIP-seq data from ENCODE (Gerstein et al., 2012). In total, 118 TFs and 91 cell lines were considered. Recombination enrichments were computed with respect to neighboring regions using the 500 bp recombination map of the British (GBR) population. The statistical significances are adjusted for multiple testing using Benjamini-Hochberg procedure. *Pol2* and TFs that may be part of MLL1/2 complexes are indicated in blue and red, respectively.

(B) Recombination enrichment with respect to the whole-genome average for the TF binding sites considered in (A). Only TFs with the highest enrichments are shown. ChIP-seq peaks of each TF were merged across all cell lines. TFs that may be part of MLL1/2 complexes (indicated in red) are generally enriched for recombination.

(C) Recombination enrichment at E2F6 binding sites as a function of the distance to the binding site, according to TDA ($L = 500$ bp and $s = 14$) recombination maps of the GBR population, as well as HapMap (Frazer et al., 2007) and AA (Hinch et al., 2011) recombination maps. E2F6 binding sites were obtained from ENCODE, merging ChIP-seq peaks across K562 and HeLa cell lines.

(D) Enrichment for predicted PRDM9 binding sites (defined by the motif CCNCCNTNNCCNC) and sperm CpG methylation, H3K4me3 and H3K27me3 marks as functions of the distance to E2F6 binding sites, for the binding sites considered in (C).

(E) Recombination enrichment at TF binding sites against enrichment for predicted PRDM9 binding sites (left) and sperm CpG methylation (right) for the TFs and cell lines considered in (A). The color scale represents enrichments for sperm H3K4me3 marks at the loci of these TF binding sites.

(F) Enrichments for sperm H3K4me3 and H3K27me3 marks at the loci of TF binding sites for the TFs and cell lines considered in (A). The color scale represents recombination enrichment at the loci of these TF binding sites. Higher recombination enrichments occur for bivalent TF binding sites.

See also Figure S3 and Table S3.

DISCUSSION

In the last few years, technological advances have allowed the identification of highly localized molecular features such as TF binding sites, methylation sites, loci of noncoding RNAs, and regions of open chromatin, at unprecedented resolution. Population-based methods tailored for the analysis of large data sets at the relevant fine genomic scales are needed to explore the relationship between recombination and these highly localized molecular features. We have presented an efficient method, based on TDA, to estimate recombination rates from very large samples of genomic sequences. Our method detects a topological signature of recombination distinct from the linkage-based information upon which existing methods depend. The performance and accuracy of this approach have allowed us to build high-resolution recombination maps of seven human populations (Figure 3). These maps can be used to establish new associations between recombination and genomic features, as we have demonstrated for TF binding sites (Figure 4) and piRNA loci (Figure 5). These associations are supported by enrichments for predicted PRDM9 binding motifs, CpG content, and epigenetic marks in germ cells. Overall, these findings provide confidence in the application of TDA to human genomic data and demonstrate its utility for overcoming some of the limitations of conventional methods and for generating hypotheses about fine-scale human recombination.

Our results raise broad questions about the control and evolution of recombination in humans. *Prdm9* knockout experiments in mice suggest that the protein encoded by this gene plays a

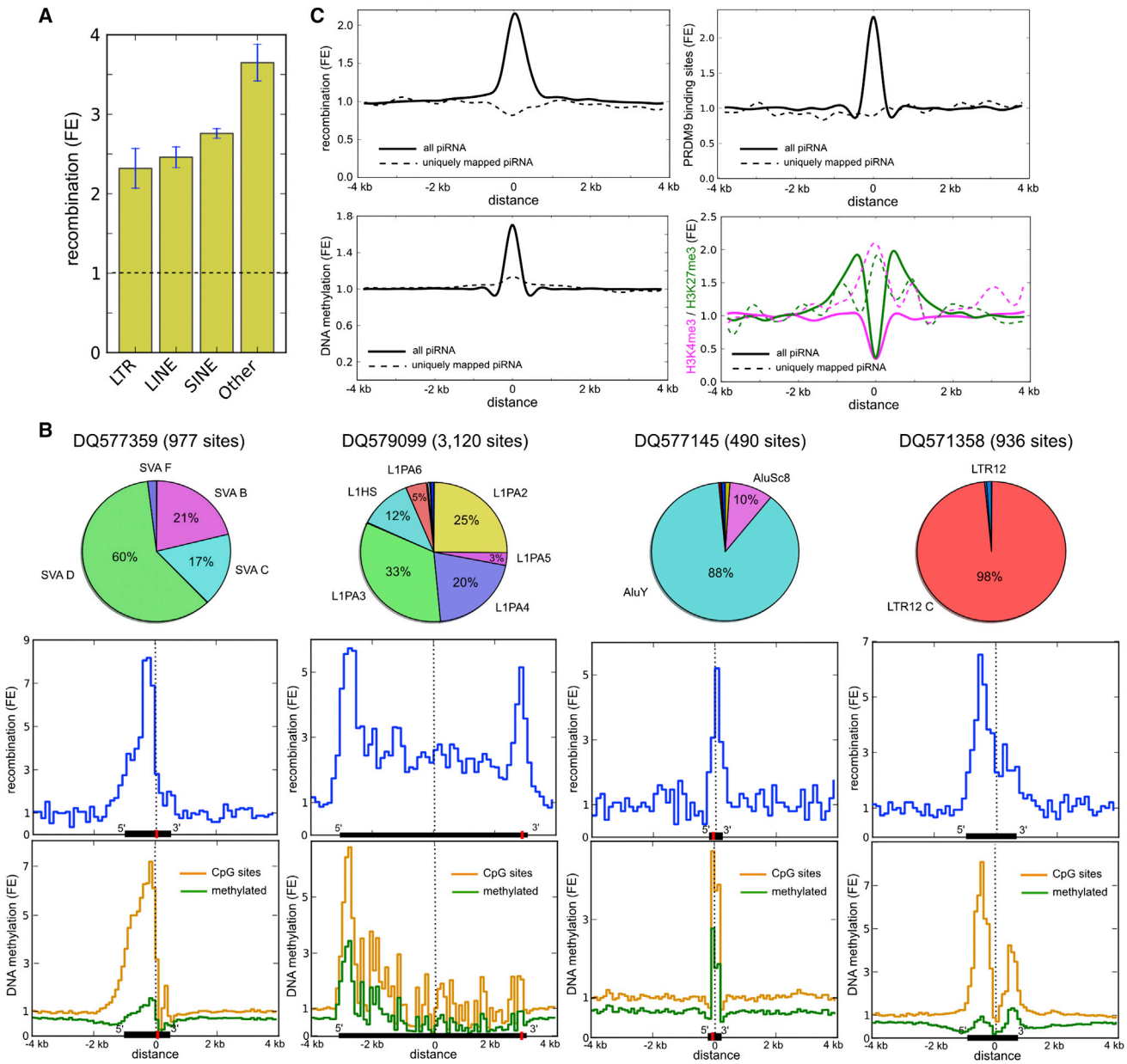


Figure 5. Relation of Recombination to piRNA Loci

(A) Recombination enrichment at repeat-derived piRNA-matched loci with respect to the whole-genome average. Repeat-derived loci with 100% identity to some sequence deposited in piRNA-Bank (Sai Lakshmi and Agrawal, 2008) were classified as belonging to SINE, LINE, SVA, or other family of repetitive elements. The recombination enrichments were computed using the British (GBR) recombination map.

(B) Enrichment for recombination, CpG sites, and sperm methylation for loci matched by four specific repeat-derived piRNA (piRNA-Bank accession numbers DQ577359, DQ579099, DQ577145, and DQ571358). The location of transposable elements and their 5′-and 3′-ends are shown in black. The predicted PRDM9 binding motifs conserved across different loci are indicated in red. The origin of coordinates corresponds to the location of the piRNA-matched sequence. In the four cases, the piRNA sequence is antisense to the transposable element.

(C) Enrichment for recombination (top left), predicted PRDM9 binding sites (top right) according to the motif CCNCCNTNNCCNC, sperm CpG methylation (bottom left), and sperm H3K4me3 and H3K27me3 marks (bottom right) for the loci considered in (A).

See also Figure S4.

role in sequestering recombination away from gene promoters and other functional genomic elements (Brick et al., 2012). We find that hypothesis to be consistent with the overall depletion of recombination at TF binding sites observed in our study. However, by disaggregating different TF binding site types, we have shown that binding sites of specific TFs are systematically enriched for recombination. This constitutes a clear exception to the general rule described above. These TF binding sites include binding sites of the E2F family and regulatory subunits of MLL1/2 complexes, which play prominent regulatory roles in germ cell

development and early embryogenesis. This circumstance in humans is analogous to the case of homologous recombination at *Saccharomyces cerevisiae* promoters, where an H3K4 methyltransferase forms part of the COMPASS protein complex and links H3K4me3 marks to the formation of meiotic DSBs at promoters (Acquaviva et al., 2013). Accordingly, MLL complexes are human homologs of the COMPASS complex in yeast (Miller et al., 2001).

From an evolutionary perspective, an increased recombination rate at gene promoters regulating germ cell development and embryogenesis would render selection at these loci more effective by unlinking selective forces that act on different positions (Hill and Robertson, 2007; Iles et al., 2003). A similar argument can be made regarding the enrichments for recombination observed at repeat-derived loci matched by piRNA sequences, particularly in light of recent work uncovering the role of transposons and piRNAs in post-transcriptional regulation in the germline (Watanabe et al., 2015). The rapid genetic divergence of genes involved in meiosis suggests that these regions undergo extraordinary positive selection (Keeney, 2008; Richard et al., 2005; Schwartz et al., 2014); it stands to reason that elevated recombination at these loci would be selected for. On the other hand, DSBs targeted to regions active during meiosis may deprive the cell of necessary transcripts. Our findings highlight a potential evolutionary trade-off in regulation of recombination that merits further study.

EXPERIMENTAL PROCEDURES

Expected b_1

We performed 3.5×10^5 neutral population simulations using the program ms (Hudson, 2002) and seq-gen (Rambaut and Grassly, 1997). Simulated haplotypes were produced for the population recombination rate, ρ, the population mutation rate, θ, and the number of sampled sequences, n, taking values in the ranges 0–25,000/N_{eff} cM, 0–62.5/N_{eff} mutations per generation, and 0–160 sampled sequences, respectively. Pairwise distances were defined using Hamming distance. For each distance matrix, we built a filtration of Vietoris-Rips complexes and computed b_1 using the software Dionysus (http://www.mrzv.org/software/dionysus/index.html). For each configuration of the parameters, we estimated $E[b_1]$ based on 500 simulations (Table S1). Simulated data were empirically described by the equation,

$$E[b_1] \approx f \cdot \log\left(1 + \frac{\rho}{g + h\rho}\right), \qquad \text{(Equation 1)}$$

where the parameters f, g, and h depend on the number of sequences and segregating sites in the sample. We performed 6.4×10^4 neutral population simulations of samples with fixed number (s = 14 and 40) of segregating sites, for ρ and n taking values in the ranges 0–4,500/N_{eff} cM and 0–160 sampled sequences, respectively. Based on these simulations, we determined by least-squares fitting the following structure for the parameters in Equation 1,

$$f = f_1 \cdot \left(\frac{n}{f_2} \cdot \log n - n\right) \qquad \text{(Equation 2)}$$

$$g = g_1 \cdot n + g_2, \qquad \text{(Equation 3)}$$

where $f_1 = 0.04404$, $f_2 = 1.5734$, $g_1 = 0.50663$, $g_2 = -2.3129$, for $s = 14$, and $f_1 = -0.08225$, $f_2 = 3697124$, $g_1 = 0.11483$, $g_2 = 8.4455$, for $s = 40$. For convenience, we fixed $h = 0$ in Equation 1.

Persistent Homology Estimator of Recombination

From population genetics simulations, we observed that the first Betti number b_1 of a set of sequences sampled from a Wright-Fisher population with recombination is Poisson distributed. From this fact and Equation 1 with $h = 0$, we proved that

$$\rho_{PH} = g\left[\left(1 + \frac{1}{f}\right)^{b_1} - 1\right] \qquad \text{(Equation 4)}$$

is an estimator of the population recombination rate ρ (see Supplemental Information). Comparison of this estimator to linkage methods was performed as described in the Supplemental Information.

Genome Scan Implementation

We implemented ρ_{PH} in three sliding windows, acting on the phased SNP genotype data of 1,000 Genomes Project. Two of the sliding windows had a fixed number of segregating sites (s = 14 and s = 40) and were moved in steps of 7 and 20 segregating sites, respectively. The other sliding window had a fixed length (L = 500 bp) and was moved in steps of 250 bp. We included both chromosomes in the case of autosomal chromosomes and only one of the two X chromosomes for females. Sources used for recombination map annotation and methods for estimating relative recombination rates and similarity across human populations are described in the Supplemental Information.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.05.008.

AUTHOR CONTRIBUTIONS

P.G.C. conceived of the methodology, formulated the persistent homology estimator of recombination, and carried out the genomic analysis. D.I.S.R., K.J.E., and P.G.C. developed and carried out comparisons to linkage methods. R.R. conceived of the methodology. A.J.L. contributed to the discussion. All authors discussed results and implications and collaborated on the writing of the manuscript.

ACKNOWLEDGMENTS

We thank Debra J. Wolgemuth, Marcia Manterola, Molly Przeworski, Nicholas Parrish, and Suzanne Christen for their useful comments and Oliver Elliott for technical support. R.R. acknowledges funding from NIH (U54 CA193313-01). P.G.C. acknowledges Universidad de Barcelona for financial support and hospitality during the initial stages of the project. D.I.S.R. acknowledges funding from NIH (R01 GM117591 and T15 LM007079).

Received: September 24, 2015
Revised: March 3, 2016
Accepted: May 26, 2016
Published: June 23, 2016

REFERENCES

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1,000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Acquaviva, L., Székvölgyi, L., Dichtl, B., Dichtl, B.S., de La Roche Saint André, C., Nicolas, A., and Géli, V. (2013). The COMPASS subunit Spp1 links histone methylation to initiation of meiotic recombination. Science *339*, 215–218.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science *316*, 744–747.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science *327*, 836–840.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin

structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R.D., and Petukhova, G.V. (2012). Genetic recombination is directed away from functional genomic elements in mice. Nature *485*, 642–645.

Carlsson, G. (2009). Topology and data. Bull. Am. Math. Soc. *46*, 255–308.

Chan, J.M., Carlsson, G., and Rabadan, R. (2013). Topology of viral evolution. Proc. Natl. Acad. Sci. USA *110*, 18566–18571.

Coop, G., Wen, X., Ober, C., Pritchard, J.K., and Przeworski, M. (2008). High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science *319*, 1395–1398.

Crawford, D.C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M.J., Nickerson, D.A., and Stephens, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat. Genet. *36*, 700–706.

DeGregori, J., and Johnson, D.G. (2006). Distinct and overlapping roles for E2F family members in transcription, proliferation and apoptosis. Curr. Mol. Med. *6*, 739–748.

Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. *10*, 285–311.

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. Discrete Comput. Geom. *28*, 511–533.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al.; International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. Nature *489*, 91–100.

Ghrist, R. (2014). Elementary Applied Topology (Createspace).

Gkountela, S., Zhang, K.X., Shafiq, T.A., Liao, W.W., Hargan-Calvopiña, J., Chen, P.Y., and Clark, A.T. (2015). DNA demethylation dynamics in the human prenatal germline. Cell *161*, 1425–1436.

Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., Yong, J., Hu, Y., Wang, X., Wei, Y., et al. (2015). The transcriptome and DNA methylome landscapes of human primordial germ cells. Cell *161*, 1437–1452.

Ha, H., Song, J., Wang, S., Kapusta, A., Feschotte, C., Chen, K.C., and Xing, J. (2014). A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. BMC Genomics *15*, 545.

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., and Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature *460*, 473–478.

Hatcher, A. (2002). Algebraic Topology (Cambridge: Cambridge University Press), p. 606.

Hein, J., Schierup, M., and Wiuf, C. (2004). Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory (USA: Oxford University Press).

Hill, W.G., and Robertson, A. (2007). The effect of linkage on limits to artificial selection. Genet. Res. *89*, 311–336.

Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature *476*, 170–175.

Hudson, R.R. (2001). Two-locus sampling distributions and their application. Genetics *159*, 1805–1817.

Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics *18*, 337–338.

Hudson, R.R., and Kaplan, N.L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics *111*, 147–164.

Iles, M.M., Walters, K., and Cannings, C. (2003). Recombination can evolve in large finite populations given selection on sufficient loci. Genetics *165*, 2249–2258.

Kauppi, L., Jeffreys, A.J., and Keeney, S. (2004). Where the crossovers are: recombination distributions in mammals. Nat. Rev. Genet. *5*, 413–424.

Keeney, S. (2008). Spo11 and the formation of DNA double-strand breaks in meiosis. Genome Dyn. Stab. *2*, 81–123.

Kehoe, S.M., Oka, M., Hankowski, K.E., Reichert, N., Garcia, S., McCarrey, J.R., Gaubatz, S., and Terada, N. (2008). A conserved E2F6-binding element in murine meiosis-specific gene promoters. Biol. Reprod. *79*, 921–930.

Kirkness, E.F., Grindberg, R.V., Yee-Greenbaum, J., Marshall, C.R., Scherer, S.W., Lasken, R.S., and Venter, J.C. (2013). Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome Res. *23*, 826–832.

Koehler, K.E., Hawley, R.S., Sherman, S., and Hassold, T. (1996). Recombination and nondisjunction in humans and flies. Hum. Mol. Genet. *5*, 1495–1504.

Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.

Li, N., and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics *165*, 2213–2233.

Li, L., Cheng, W.Y., Glicksberg, B.S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E.P., and Dudley, J.T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci. Transl. Med. *7*, 311ra174.

Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A.R., Zhu, P., Hu, X., Xu, L., Yan, L., et al. (2012). Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. Science *338*, 1627–1630.

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics *160*, 1231–1241.

McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science *304*, 581–584.

Miller, T., Krogan, N.J., Dover, J., Erdjument-Bromage, H., Tempst, P., Johnston, M., Greenblatt, J.F., and Shilatifard, A. (2001). COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. Proc. Natl. Acad. Sci. USA *98*, 12902–12907.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al.; 1,000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59–65.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W.R., Hannon, G.J., and Smith, A.D. (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell *146*, 1029–1041.

Myers, S.R., and Griffiths, R.C. (2003). Bounds on the minimum number of recombination events in a sample history. Genetics *163*, 375–394.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science *310*, 321–324.

Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science *327*, 876–879.

Nicolau, M., Levine, A.J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proc. Natl. Acad. Sci. USA *108*, 7265–7270.

Pan, J., Sasaki, M., Kniewel, R., Murakami, H., Blitzblau, H.G., Tischfield, S.E., Zhu, X., Neale, M.J., Jasin, M., Socci, N.D., et al. (2011). A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. Cell *144*, 719–731.

Parvanov, E.D., Petkov, P.M., and Paigen, K. (2010). Prdm9 controls activation of mammalian recombination hotspots. Science *327*, 835.

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G.V., and Camerini-Otero, R.D. (2014). DNA recombination. Recombination initiation maps of individual human genomes. Science *346*, 1256442.

Rambaut, A., and Grassly, N.C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. *13*, 235–238.

Richard, G.F., Kerrest, A., Lafontaine, I., and Dujon, B. (2005). Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination. Mol. Biol. Evol. *22*, 1011–1023.

Sai Lakshmi, S., and Agrawal, S. (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res. *36*, D173–D177.

Schwartz, J.J., Roach, D.J., Thomas, J.H., and Shendure, J. (2014). Primate evolution of the recombination regulator PRDM9. Nat. Commun. *5*, 4370.

Smagulova, F., Gregoretti, I.V., Brick, K., Khil, P., Camerini-Otero, R.D., and Petukhova, G.V. (2011). Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature *472*, 375–378.

Smith, Z.D., Chan, M.M., Humm, K.C., Karnik, R., Mekhoubad, S., Regev, A., Eggan, K., and Meissner, A. (2014). DNA methylation dynamics of the human preimplantation embryo. Nature *511*, 611–615.

Templeton, A. (2002). Out of Africa again and again. Nature *416*, 45–51.

Velasco, G., Hubé, F., Rollin, J., Neuillet, D., Philippe, C., Bouzinba-Segard, H., Galvani, A., Viegas-Péquignot, E., and Francastel, C. (2010). Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. Proc. Natl. Acad. Sci. USA *107*, 9281–9286.

Wall, J.D. (2000). A comparison of estimators of the population recombination rate. Mol. Biol. Evol. *17*, 156–163.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. Cell *150*, 402–412.

Watanabe, T., Cheng, E.C., Zhong, M., and Lin, H. (2015). Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. Genome Res. *25*, 368–380.

Zomorodian, A., and Carlsson, G. (2005). Computing persistent homology. Discrete Comput. Geom. *33*, 249–274.

Cell Systems, Volume *3*

Supplemental Information

Topological Data Analysis Generates High-Resolution, Genome-wide Maps of Human Recombination

Pablo G. Camara, Daniel I.S. Rosenbloom, Kevin J. Emmett, Arnold J. Levine, and Raul Rabadan

Supplemental Figures

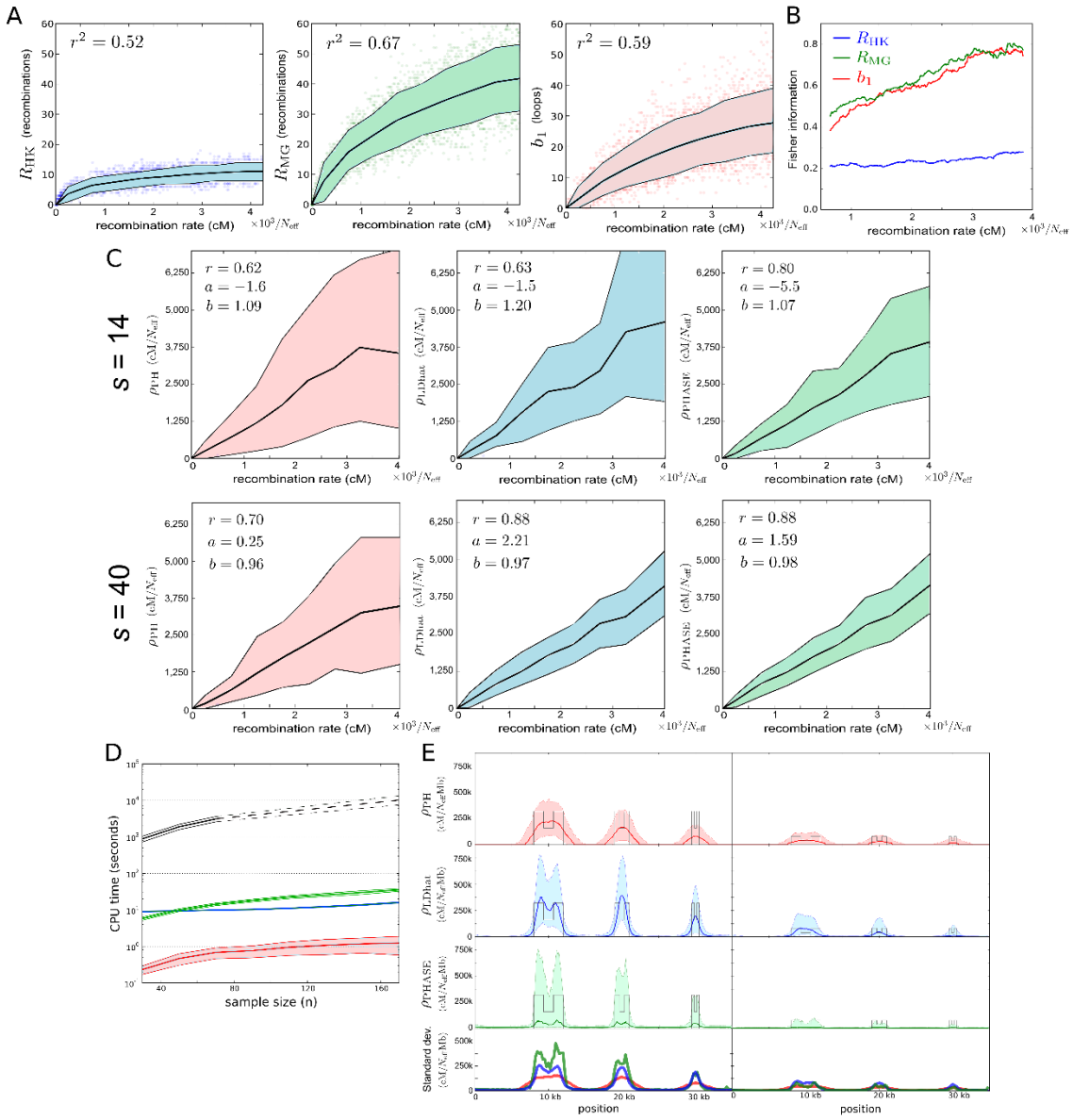


Figure S1. Persistent homology estimator of recombination (related to Figure 2).

(**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.

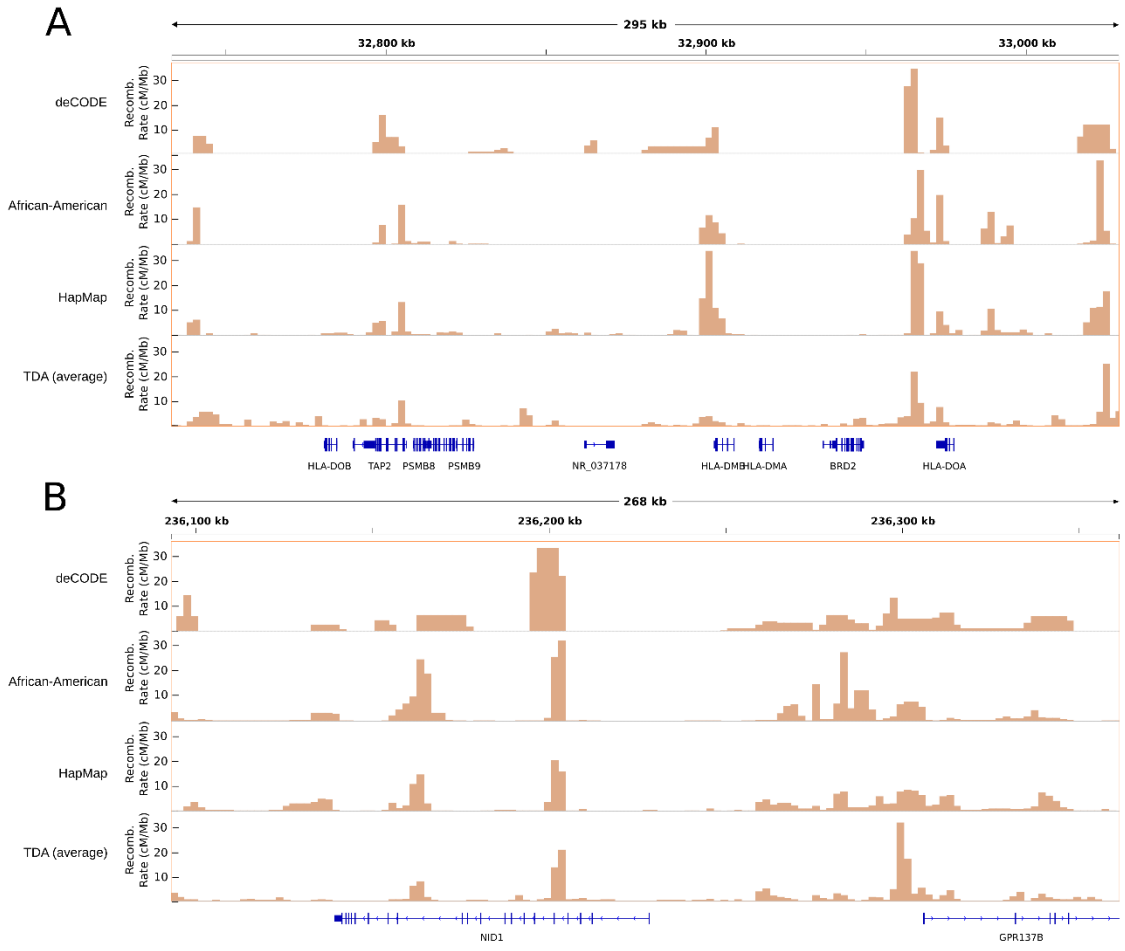
(**B**) Fisher information for each of the 3 summaries in (A) as a function of the recombination rate. Information was computed in increments of $12.5/N_{\text{eff}}$ cM. A smoothed trend is plotted by averaging windows of 101 computed values, weighted by the number of simulations.

(**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.

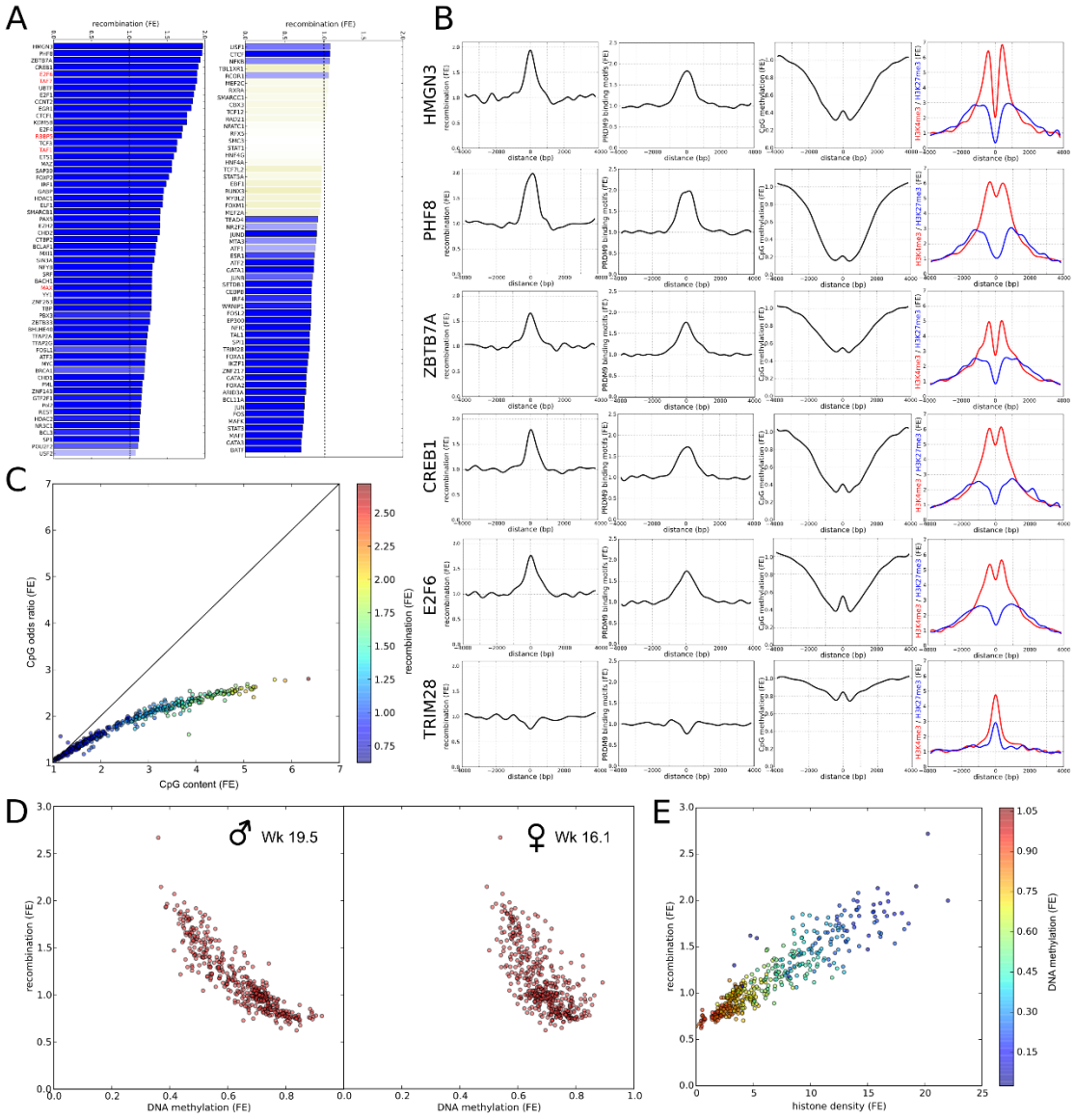


Figure S3. Recombination enrichment at TF binding sites (related to Figure 4).

(**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.

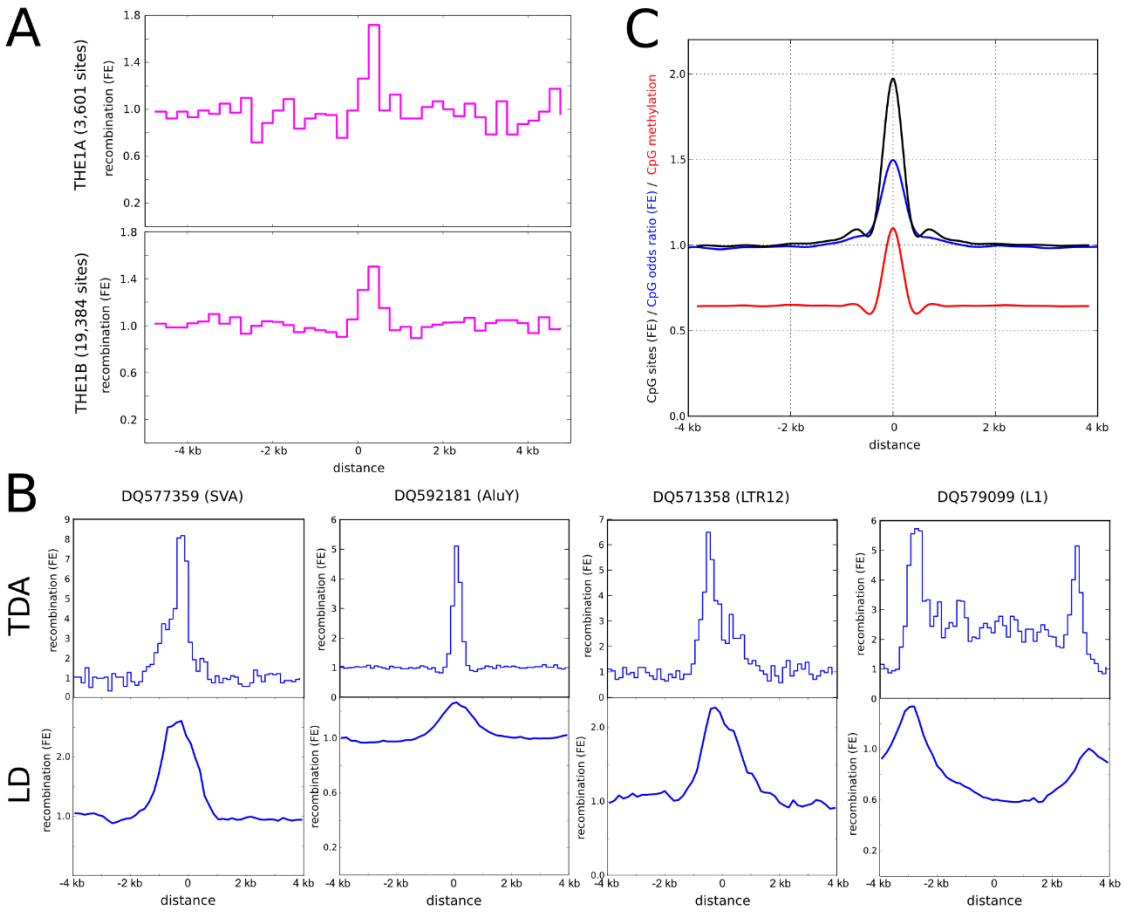


Figure S4. Recombination enrichments at repeat-derived loci matched by piRNA (related to Figure 5).

(**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.