MODELING SELECTION BIAS ON RECOMBINATION RATES INFERRED THROUGH LINKAGE DISEQUILIBRIUM

MODELING SELECTION BIAS ON RECOMBINATION RATES INFERRED THROUGH LINKAGE DISEQUILIBRIUM

By

AMY FAN

____

A Thesis Submitted to The W.A. Franke Honors College

In Partial Fulfillment of the Bachelors degree
With Honors in

Statistics and Data Science

THE UNIVERSITY OF ARIZONA

MAY 2023

Approved by:

____

Dr. Ryan Gutenkunst
Department of Mathematics

Abstract

Recombination is a key part of evolutionary theory, and understanding the ways selection can bias inferred rates in a population can help us investigate better models for inference. This experiment models the bias in recombination rate inferences on a simulated genome with selection. Using SLiM forward genetic simulation, this experiment creates two basic genomic structures, one with and one without a hotspot. Then, using Pyrho a fine-scaled linkage disequilibrium-based inference model, the experiments reveal how selection biases the linkage disequilibrium model. Notably, with increasing nonsynonymous distribution of fitness effects (DFE), the inferences worsen and show a decreasing trend. This is most notable in the hotspot region of the second genomic structure (with a hotspot). The results show that the assumption of neutral selection in popular population-based inference methods is extremely important and should be addressed. In particular, among organisms with less compact genomes, the issue of selection would become more extreme and disruptive. In future models, this could be taken into consideration to improve inference ability among a diverse set of organisms with different levels of selection in their genome. Understanding how recombination rates differ across the tree of life can also reveal interesting molecular structures which further motivates accuracy in these inference methods.

Introduction

Recombination in evolutionary biology

Recombination is a pivotal step in ensuring proper replication during sexual reproduction of eukaryotic organisms. During the first stage of meiosis, parental chromosomes will form a chiasma and crossover. This process, also known as meiotic recombination, increases genetic diversity within the population by allowing genetic information from two parents to be shuffled into a single chromosome. Meiotic recombination also ensures the fidelity of chromosome pairing during meiosis. Recombination is one of five central parameters considered by evolutionary theory along with mutation, selection, genetic drift, migration, and recombination. However, the molecular machinery of recombination has only been partially elucidated with the discovery of PRDM9 (1) and remains a mystery.

As a part of the genetic architecture, recombination cannot be disregarded as noise in the context of evolution (2) as they were in older models. The additive unconstrained gene action model, for instance, considers large, panmictic, recombining populations that allow all combinations of alleles (3). Thus, the model concludes that the evolutionary effect of an allele could be defined by taking the average of all genotype combinations. This classical model assumes that the underlying genetic interactions could be averaged and treated like statistical noise. However, empirical data show that various aspects of genetic architecture (epistasis, pleiotropy, cryptic variations, etc.) systematically influence evolutionary dynamics, and newer models seek to understand the effect of these molecular mechanisms (2).

For example, Fisher–Muller model proposes that recombination facilitates adaptation by allowing beneficial mutations to arise among different lineages and reducing clonal interference. This theory has been tested in *Escherichia coli* where a beneficial mutation was shown to fix at a faster rate when an F plasmid was inserted to mediate recombination (4). In the absence of the F plasmid, the beneficial mutation conferred a reduced competitive advantage and fixed at a slower rate, indicating interference between competing beneficial mutations. These results indicate that recombination can drive evolution by decoupling genes to allow adaptation to occur faster.

Meiotic recombination aids in speciation by evolving maladaptive gene exchange to favor reproductive isolation even when gene flow is present. Both suppression of recombination through chromosomal rearrangement meiosis in hybrids and genic modifiers can lead to the reduction of recombination rates and contribute to speciation.

Linkage disequilibrium measures genetic difference within a population by determining the non-random association of alleles at different loci. Strong linkage disequilibrium refers to one population having a large proportion of one genotype (e.g. AB) while the other population primarily retains the other genotype (e.g. ab While chromosomal rearrangements do not change linkage relationships, genic modifiers can affect linkage by hitchhiking with other genes to spread and change recombination rates across the genome. They can create recombination hotspots by binding to chromosomal regions and increasing the chance of double strand breaks (DSBs) (5).

Discussions over the paradox of sex also reveal with mathematical modeling that recombination contributes to the evolutionary advantages of sexual reproduction under certain restrictions (6). Despite the cost of sexual reproduction (two-fold cost, parasites, sexually transmitted disease), the maintenance and regularity of sexual reproduction remain. Recombination lies at the heart of sexual reproduction and evolutionary biology. Substantial genetic variation in recombination rates and patterns (among chromosomes, between sexes, among individuals, populations, and species) indicate a complex mechanism driving speciation (7).

Recombination Inference and the Linkage Disequilibrium Model

There are two major categories for estimating the recombination landscape: cytological and genomic-based methods. Cytological methods directly visualize different stages of the recombination process and count events. While cytological methods provide a coarse resolution of recombination landscape, these methods use direct visualization to help uncover mechanisms underlying recombination. For instance, comparing DSBs and resulting crossover events can suggest meiotic regulators of crossover frequencies (8,9).

Genomic-based methods, on the other hand, indirectly infer crossover events to estimate the recombination landscape. These methods fall into three broad categories of gamete-based, pedigree-based, and population-based methods (10) where each method estimates a different aspect of the recombination landscape. The gamete-based method measures the frequencies of crossovers for an individual using (often male) haploid gametes. The gamete-based method is powerful because it provides high resolution crossover frequency data from low sampling and allows comparisons between individuals. However, because this method infers on haploid gametes, the recombination maps are sex-specific, biased by individual SNP density, and temporally limited as a snapshot.

Alternatively, the classical pedigree-based method estimates a genetic linkage map from patterns of inheritance for alleles within a family tree. This method produces sex-specific recombination maps and inferences unbiased by population level processes due to the scope of the method. However, this method is more involved because of the larger sample size and complex study design to map out a pedigree. Furthermore, results are low resolution (~1 Mb, 0.5-2 cM) and temporally limited to a snapshot in time like the gamete-based method (10).

Finally, the population-based method is used to infer population recombination rate using linkage disequilibrium and is also referred to as the linkage disequilibrium model (10). Within this schema, the population is assumed to contain N diploid, randomly mating individuals, where each locus mutates at a rate μ under the infinite sites model. The infinite sites model considers mutable sites to be continuous along the genome so that there are an infinite number of sites where new mutations can occur. The distribution of linkage disequilibrium describes the probabilities of various samples of gametes (11). The population-based method infers a moderate to high resolution recombination landscape and is not restricted to a snapshot in time.

Recombination rates can vary on both a small scale and a large scale. One of the most well-known drivers of small-scale recombination rate evolution is PRDM9. PRDM9 has been highly studied as a driving factor of chromatin remodeling and recombination hotspot evolution but only explains a small portion of recombination rate evolution. For instance, while there are various PRDM9 alleles and diverse binding sites, PRDM9 activity cannot explain the fine-scale recombination rate differences among closely related populations with different demographic histories. Furthermore, the predominant PRDM9 allele in non-African populations has a weak erosion on binding sites (12). Uncovering more driving forces would help provide a better understanding of the underlying mechanisms behind recombination rate evolution. Large-scale changes in linkage can occur via chromosomal rearrangements such as translocations and inversions (5) and greatly impact the recombination landscape.

Linkage Disequilibrium Model

Various methods of the population-based approach can infer recombination events. For example, the four-gamete test is best for determining the lower bound of recombination events (13). The test infers recombination events between pairs of loci in diploids such that four possible gametic combinations exist. The R_m used to infer the number of recombination events is determined by combining the events within nonoverlapping intervals under conservative assumptions. Thus, the four-gamete method is very fast to compute but often overlooks recombination events and underestimates recombination rates (14).

More complex methods use a mathematical model known as the coalescent to recreate the underlying gene genealogies and then estimate the population recombination rate. Ancestral recombination graphs (ARGs) describe the underlying population history and contain mutation and recombination events (15).

Assumptions of the Linkage Disequilibrium Model

Bias in the estimates of population recombination rate can occur if the model assumptions are not met as linkage disequilibrium becomes distorted. The model assumptions relate to constant population size, random mating in the absence of population structure and migration, genetic drift, mutation rate, and selective neutrality (10). Gene flow, for instance, tends to lead to overestimations of population recombination rate (16).

Researchers are beginning to address some of these simplifying assumptions, and new methods are being created to ameliorate them. For example, LDpop addresses the assumption of constant population size to allow variable population sizes that are piecewise constant (17).

Similarly, Pyrho is a penalty composite likelihood (fused-LASSO approach) method that is demography-aware (12). Pyrho can account for nonequilibrium demographic histories unlike other methods such as LDhat which cannot discriminate between different populations. Pyrho inferences on 26 different human populations noted a high correlation of fine scale recombination maps for each population with their demographic history (12). This suggests that fine-scale recombination rates are highly polygenic.

By comparing across the tree of life in a systematic manner, recombination rates differences can reflect underlying mechanisms that drive the evolution of recombination rates and speciation. However, one of the important assumptions that could be violated during this process is the assumption of selective neutrality. Although selection in human genomes is largely neutral (coding sequences make up ~1.1% of the genome) (18), other organisms such as *Mus musculus* and *Drosophila melanogaster* have more compact genomes with higher proportions of coding regions (19). To investigate and compare different organisms in the tree of life, the assumption of selective neutrality must be investigated further.

Methods

SLiM 3 – Forward Genetic Simulation (20)

SLiM 3 is a powerful evolutionary simulation package that is based on the Wright-Fisher model of evolution. In this experiment, we use the default Wright-Fisher model to generate VCF files from a population of 10,000 diploid individuals. The genome length for the population is 2Mb, and the VCF files are sampled from 25 individuals in the population after 200,000 generations. The mutation rate is set to 1.133e-8 (21) as inferenced from an Icelandic population, and the baseline recombination rate (not hotspots) is 1e-8 (22). Each experiment is repeated 10 times (n = 10) to determine trends.

Pyrho - Fine-Scale Recombination Inference (12)

Pyrho is a fast demography-aware inference method for fine scale recombination rates which uses penalized composite likelihood inference (fused-LASSO). Pyrho creates a lookup table from the population size, sample size, and the mutation rate to increase the efficiency of the optimization step. The hyperparameters best for this experiment were a window size of 50 and a block size of 50. Pyrho uses this information along with the VCF file from SLiM's output to infer the recombination maps for each VCF file. The output is an RMAP file which maps out the recombination rates across the genome.

Experimental Constructs (SLiM)

To systematically investigate the effect of selection on inferred recombination rates, the first experiment uses a simple uniform model containing a single genomic element containing neither intron-exon structures nor hotspots. The proportion of selection in the genome in this first experiment are 0% (neutral), 20%, 40%, 60%, 80%, and 100%. The proportion of selection in this first construct is represented by the proportion of nonsynonymous mutations in the genome and are randomly assigned across the genome with a uniform distribution. The parameters for the nonsynonymous distribution of fitness effects (DFE) come from a population of Yoruban Nigerians (Table 1) ([23]).

The second construct builds upon the first construct by adding hotspots. Since Pyrho infers fine-scale recombination maps, this construct contains a 100kb recombination hotspot at between base pairs 1,900,000 the final 2,000,00,000 base pair. The proportion of selection in the genome in this first experiment are 0% (neutral), 20%, 40%, 60%, 80%, and 100%. The proportion of selection in the second construct is also represented by the proportion of nonsynonymous mutations in the genome and are randomly assigned across the genome with a uniform distribution.

Data Processing – RMAP Visualization

To visualize the RMAP files, the raw data was processed and displayed on a plot using matplotlib in Python 3.10.10. The true recombination landscape is displayed in red (linewidth = 3) alongside the experimental recombination landscape inferences.

Data Processing – Bias and Variation

To model the accuracy of recombination inferences, the average bias across the genome is measured for each RMAP. Boxplots are used to observe changes in recombination inference accuracy over a range of selection levels (0-100%). For the genome structures with hotspots, the bias is calculated for each segment with a single recombination rate and then summed over all segments.

The variation of recombination rates across the genome is quantified using the standard deviation:

$$\sigma = \sqrt{\int_{-\infty}^{+\infty} (x-\mu)^2 \, f(x)dx}$$

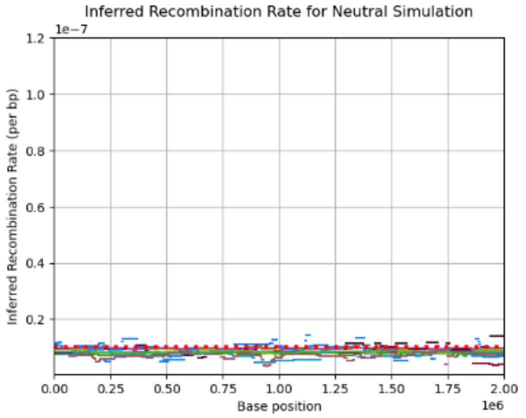
where x is the inferred recombination rate, μ is the mean inferred recombination rate, and f(x) is the proportion of the inferred genome represented by that recombination rate. All plots for bias and variation metrics are plotted in Rstudio using ggplot2.

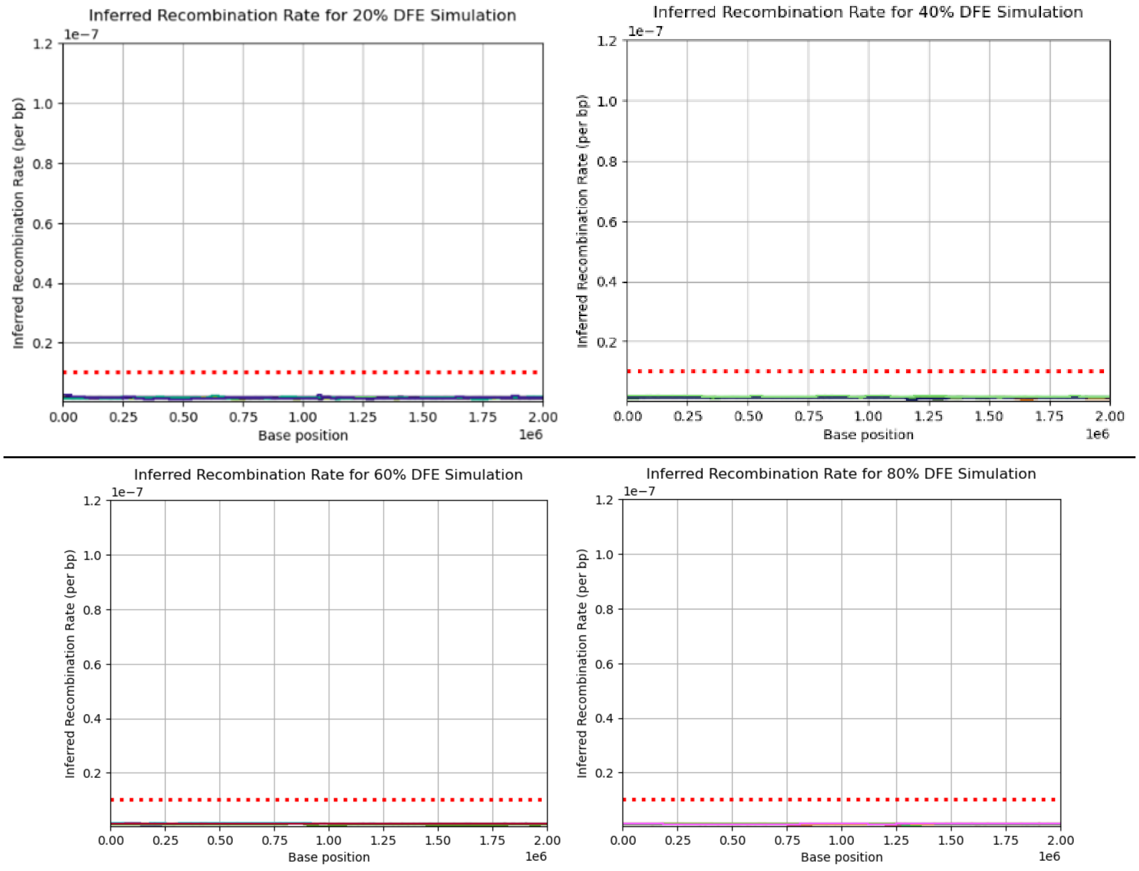
Other Resources

All scripts are run on the High Performance Computing Systems at the University of Arizona. Batch scripts are run using slurm workload manager. SLiM and Pyrho are installed and compiled using Miniconda3.

Results

Recombination Landscape of Uniform Genome with No Hotspots





Inferred Recombination Rate for 20% DFE Simulation

Inferred Recombination Rate for 40% DFE Simulation

Inferred Recombination Rate for 60% DFE Simulation

Inferred Recombination Rate for 80% DFE Simulation

Figure 1: Visualizing the recombination landscape for uniform genome structures with no hotspots. As the portion of nonsynonymous distribution of fitness effects (DFEs) increases, the inference becomes poor.

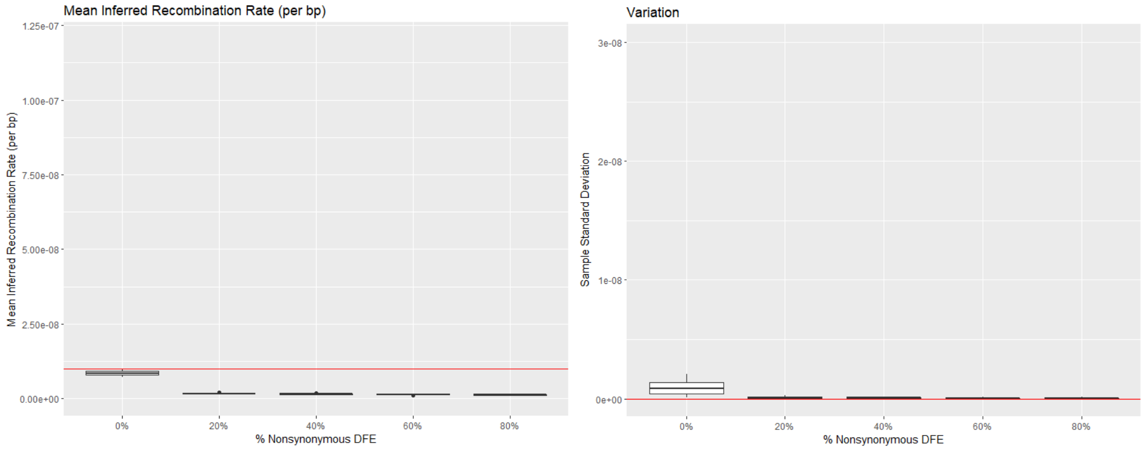
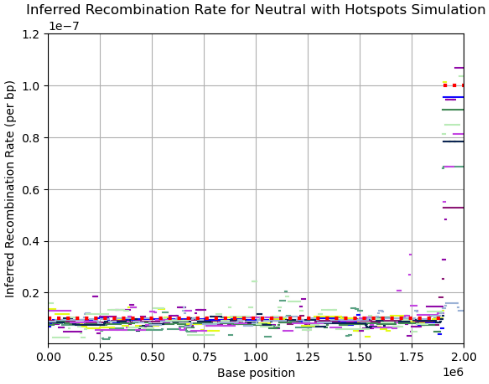


Figure 2: Summarizing the n = 10 recombination landscapes for each percentage nonsynonymous DFE into boxplots. Left: mean inferred recombination rate per base pair. Right: sample standard deviation.

As the percentage of nonsynonymous distribution of fitness effects increases in the genome, the recombination rate inferences noticeably worsened. Fig. 1 reveals poor inferences as soon as the percentage goes up to 20%, and there appears to be a decrease in the inferred value as the percentages increase (Fig. 2).

Recombination Landscape of Uniform Genome with Hotspots



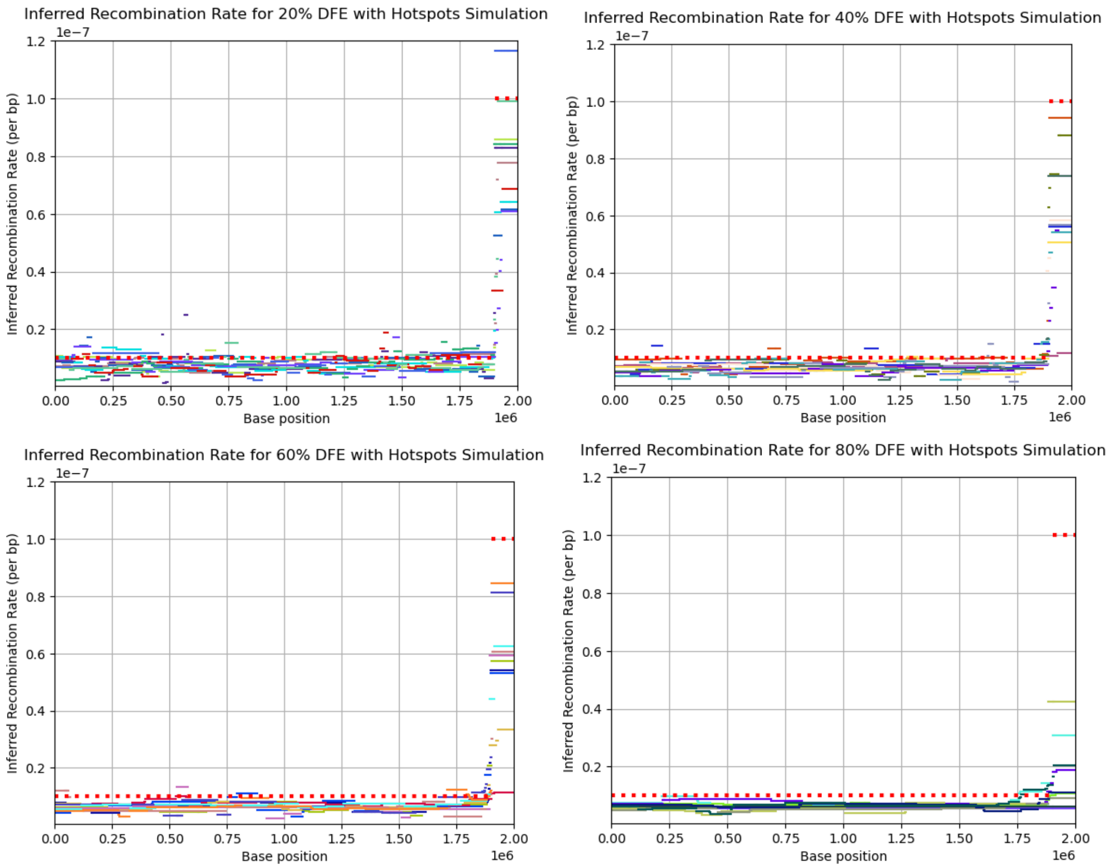


Figure 3: Visualizing the recombination landscape for uniform genome structures with hotspots.

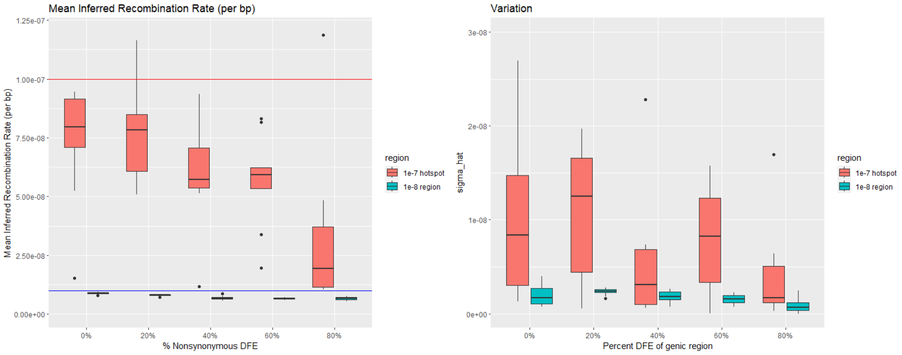


Figure 4: Summarizing the n = 10 recombination landscapes for each percentage nonsynonymous DFE into boxplots. Left: mean inferred recombination rate per base pair.

Right: sample standard deviation. The genomic region with recombination rate 1e-8 and the hotspot region with recombination 1e-7 are considered separately.

In this structure, a 100kb hotspot is present at the end of the 2Mb genome. A similar trend is seen in the previous experiment where increasing percent nonsynonymous DFE produces poor estimates (Fig. 3). The estimated values seem to show a decreasing trend as the percentage of nonsynonymous DFE increases. There is also a higher variation and a more extreme decreasing trend in estimated values within the hotspot region (Fig. 4).

Conclusion

When selection becomes a stronger influence in the evolutionary dynamic, recombination rate inferences become heavily affected. Modern methods of population-based inference assume constant population size, random mating in the absence of population structure and migration, genetic drift, mutation rate, and selective neutrality. Strides have been made to address these assumptions such as with LDpop which allows for various population sizes. By addressing the bias resulting from the model, researchers can further discover new genetic mechanisms of recombination rate and develop a better understanding of evolution.

References

1. McVean, G., Myers, S. PRDM9 marks the spot. Nat Genet 42, 821–822 (2010). https://doi.org/10.1038/ng1010-821
2. Hansen, T. F. The evolution of genetic architecture. Annu. Rev. Ecol. Evol. Syst. 37, 123–157 (2006).
3. Punnett, R. The Genetical Theory of Natural Selection . Nature 126, 595–597 (1930). https://doi.org/10.1038/126595a0
4. Cooper, T. F. Recombination speeds adaptation by reducing competition between beneficial mutations in populations of Escherichia coli. PLoS Biol. 5, e225 (2007).
5. Ortiz- Barrientos, D., Engelstädter, J. & Rieseberg, L. H. Recombination rate evolution and the origin of species. Trends Ecol. Evol. 31, 226–236 (2016).
6. Otto, S. P. & Lenormand, T. Evolution of sex: resolving the paradox of sex and recombination. Nat. Rev. Genet. 3, 252 (2002).
7. Butlin, R. K. Recombination and speciation. Mol. Ecol. 14, 2621–2635 (2005). This influential perspective article discusses the variation in recombination, theoretical expectations and its importance for speciation.
8. Zickler, D., Moreau, P. J., Huynh, A. D. & Slezec, A. M. Correlation between pairing initiation sites, recombination nodules and meiotic recombination in Sordaria macrospora. Genetics 132, 135-148 (1992).

9. Gruhn, J. R., Rubio, C., Broman, K. W., Hunt, P. A. & Hassold, T. Cytological studies of human meiosis: sex- specific differences in recombination originate at, or prior to, establishment of double- strand breaks. PLoS One 8, e85075 (2013).

10. Peñalba, J.V., Wolf, J.B.W. From molecules to populations: appreciating and estimating recombination rate variation. Nat Rev Genet 21, 476–492 (2020). https://doi.org/10.1038/s41576-020-0240-1

11. Golding, G. B. The sampling distribution of linkage disequilibrium. Genetics 108, 257–274 (1984).

12. Spence JP, Song YS. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. Sci Adv. 2019 Oct 23;5(10):eaaw9206. doi: 10.1126/sciadv.aaw9206. PMID: 31681842; PMCID: PMC6810367.

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

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

15. Arenas, M. The importance and application of the ancestral recombination graph. Front. Genet. 4, 206 (2013).

16. Samuk, K. Noor, MAF. Gene flow biases population genetic inference of recombination rate. G3 (Bethesda). 2022 Nov 4;12(11):jkac236. doi: 10.1093/g3journal/jkac236. PMID: 36103705; PMCID: PMC9635666.

17. Kamm, JA. Spence, JP. Chan, J. Song, YS. Two-Locus Likelihoods Under Variable Population Size and Fine-Scale Recombination Rate Estimation, Genetics, Volume 203, Issue 3, 1 July 2016, Pages 1381–1399, https://doi.org/10.1534/genetics.115.184820

18. Kirschner, M. Human genome and human gene statistics.

19. Pray, L. (2008) Eukaryotic genome complexity. Nature Education 1(1):96

20. Haller BC, Messer PW, SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model, *Molecular Biology and Evolution*, Volume 36, Issue 3, March 2019, Pages 632–637, https://doi.org/10.1093/molbev/msy228

21. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, Ward LD, Arnadottir GA, Helgason EA, Helgason H, Gylfason A, Jonasdottir A, Jonasdottir A, Rafnar T, Frigge M, Stacey SN, Th Magnusson O, Thorsteinsdottir U, Masson G, Kong A, Halldorsson BV, Helgason A, Gudbjartsson DF, Stefansson K. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017

Sep 28;549(7673):519-522. doi: 10.1038/nature24018. Epub 2017 Sep 20. PMID: 28959963.

22. Dumont BL, Payseur BA. Evolution of the genomic rate of recombination in mammals. Evolution. 2008 Feb;62(2):276-94. doi: 10.1111/j.1558-5646.2007.00278.x. Epub 2007 Dec 6. PMID: 18067567.
23. Castellano D, Macià MC, Tataru P, Bataillon T, Munch K, Comparison of the Full Distribution of Fitness Effects of New Amino Acid Mutations Across Great Apes, *Genetics*, Volume 213, Issue 3, 1 November 2019, Pages 953–966, https://doi.org/10.1534/genetics.119.302494