Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing

Uffe Hellsten[1], Kevin M. Wright [2], Jerry Jenkins [1,3], Shenqiang Shu[1], Yao-Wu Yuan[4], Susan R. Wessler[5], Jeremy Schmutz[1,3], John H. Willis[6], and Daniel S. Rokhsar[1,7]

[1] Berkeley National Laboratory/DOE Joint Genome Institute, Walnut Creek, California 94598, USA

Department of Organismic and Evolutionary Biology - Harvard University, Cambridge, Massachusetts, USA

[2] Hudson Alpha Institute of Biotechnology, Huntsville, Alabama 35806, USA

[3] Department of Ecology and Evolutionary Biology - University of Connecticut, Storrs, Connecticut 06269, USA

[4] Botany and Plant Sciences, Genomics 4107A, University of California - Riverside, Riverside, California 92521, USA

[5] Department of Biology - Duke University, Durham, NC 27708, USA

[6] Department of Molecular and Cell Biology, University of California – Berkeley, Berkeley, California 94720, USA

**To whom correspondence should be addressed*: Uffe Hellsten (uhellsten@lbl.gov)

November 1, 2013

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

DISCLAIMER:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of

Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing

their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing

Uffe Hellsten[1], Kevin M. Wright[2], Jerry Jenkins[1,3], Shengqiang Shu[1], Yao-Wu Yuan[4], Susan R. Wessler[5], Jeremy Schmutz[1,3], John H. Willis[6], and Daniel S. Rokhsar[1,7]

1. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA
2. Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA.
3. HudsonAlpha Institute of Biotechnology, Huntsville, Alabama 35806, USA
4. Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA
5. Botany and Plant Sciences, Genomics 4107A, University of California, Riverside, Riverside, CA 92521, USA
6. Department of Biology, Duke University, Durham, NC 27708, USA
7. Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA

Corresponding Author:

Uffe Hellsten
DOE Joint Genome Institute
2800 Mitchell Drive
Walnut Creek , CA 94598

Ph: (925) 890-3008
Email: uhellsten@lbl.gov

Classification: BIOLOGICAL SCIENCES; Genetics

Keywords: Recombination, Plant biology, Population genetics

Abstract

Meiotic recombination rates can vary widely across genomes, with hotspots of intense activity interspersed among cold regions. In yeast, hotspots tend to occur in promoter regions of genes, whereas in humans and mice hotspots are largely defined by binding sites of the *PRDM9* protein. To investigate the detailed recombination pattern in a flowering plant we use shotgun resequencing of a wild population of the monkeyflower *Mimulus guttatus* to precisely locate over 400,000 boundaries of historic crossovers or gene conversion tracts. Their distribution defines some 13,000 hotspots of varying strengths, interspersed with cold regions of undetectably low recombination. Average recombination rates peak near starts of genes and fall off sharply, exhibiting polarity. Within genes, recombination tracts are more likely to terminate in exons than in introns. The general pattern is similar to that observed in yeast, as well as in *PRDM9*-knockout mice, suggesting that recombination initiation described here in *Mimulus* may reflect ancient and conserved eukaryotic mechanisms.

Significance statement:

This work characterizes variation in recombination across the genome of a flowering plant in unprecedented detail using novel population genomic and computational approaches. The resulting recombination map approaches nucleotide-level resolution and advances our understanding of basic properties of recombination, notably the findings of enhanced recombination near starts of genes, varying degrees of intensities of "hot spots", higher activity in exons than introns, and that a large fraction of the genome appears devoid of any recombination activity.

/body

Introduction

Meiotic recombination is a highly regulated process that enables pairing of homologous chromosomes and, by the formation of crossovers, ensures proper segregation[1]. Along with mutation, drift, and selection, recombination is a critical factor in shaping genome-wide sequence variation. Recombination rates vary substantially across eukaryote genomes[2] in a manner that we are only beginning to understand. In human and mice, the location of regions of strong recombination ("hot spots") are largely determined by PRDM9 binding sites[3], whereas in yeast such regions are associated with nucleosome-depleted open chromatin often associated with gene promoters[4]. When PRDM9 is disabled in mouse, hot spots tend to re-localize to promoter regions[5]. In flowering plants, at least one example of a promoter-associated hot spot has been reported[6], but it remains an open question whether this is a general tendency in plants.

The positions of crossovers and the boundaries of gene conversion tracts resulting from meiotic recombination are often imprecisely known, since they can only be identified

based on the location of nearby segregating markers. Within a species, genome-wide variation in recombination rates can be determined by following the inheritance of such genetic markers in crosses or pedigrees[7-10], or by examining patterns of linkage disequilibrium within a population[11-15]. Population-based approaches have the advantage that in diverse populations hundreds of thousands of historical recombination events can be sampled, compared with only hundreds in the largest pedigrees.

The monkeyflower *Mimulus guttatus* has an exceptionally high nucleotide diversity, which makes it a particularly appealing system for characterizing the boundaries of recombination events. We observed an average pairwise nucleotide difference of π = 2.9% in a sample of 98 wild plants (196 haploid genomes) from four locations within a 16-km radius in the Sierra Nevada foothills in Northern California (SI). At such high diversity, pairs of adjacent SNPs defining local haplotypes are often found on the same Illumina sequencing read (*e.g.*, within 50 bases). Thus short-range haplotypes can be determined cost-effectively by shotgun sequencing pooled samples rather than by sequencing each plant individually.

For a pair of nearby segregating biallelic SNPs we expect to observe only three of the four possible haplotypes unless recombination and/or parallel mutation has occurred since the originating mutations. This is the essence of Hudson's four-gamete test[16] and allows us to identify putative boundaries of historical crossovers and gene conversion tracts (Fig. 1) to within a fraction of a read length. If this information is combined with population genetic models, local recombination rates can then be inferred[17].

Results

With this goal in mind, we sequenced pooled population samples of *M. guttatus* to an average of 255X genomic coverage using Illumina. In parallel, an independent reference genome sequence for *M. guttatus* was assembled for the IM62 line using conventional Sanger whole genome shotgun methods. This 322 Mb annotated reference genome is available at Phytozome[18] and described in more detail in the SI. About 111.3 million bases of the Sanger-derived reference genome were outside of repetitive regions and covered by between 58 and 450 Q30+ bases from the population samples, which corresponds, on average, to sampling 50 distinct haploid genomes (SI). We identified 9.43 million of these positions (8.5%) as "common" SNPs with minor allele frequency (MAF) of at least 5%. The folded allele frequency spectrum is shown in Supplemental Figure S5 and is well modeled by a coalescent process with an exponentially increasing effective population size.

To develop a collection of nearby segregating variants suitable for our analysis we first identified all pairs of common, silent SNPs separated by 50 bp or less. To simplify comparison to coalescent models, we normalized the analysis to a sample size of 50 haplotypes per locus. Of these, 11.5 million pairs of SNPs had sufficient power for four-gamete testing in a sample of 50 (*i.e.,* the rarest haplotype had expected frequency > 1/50 in linkage equilibrium (SI)). We define F4 as the fraction of such SNP-pairs that have all four possible haplotypes represented in the dataset. As in Hudson's four-gamete test[16],

nonzero F4 indicates the occurrence of either historical recombination between the SNPs, or parallel mutation at one or both SNPs. Fig. 2 shows average F4 values as a function of distance to the nearest annotated gene coding sequence (CDS). This value peaks immediately after the CDS starts, suggesting that the first coding exon of genes have on average nearly twice the amount of the recombination activity seen elsewhere.

We used coalescent simulations to convert observed F4 values into an effective recombination rate ρ, defined as the number of recombination boundaries per base per generation per haploid genome. At fixed sample size, F4 depends on (**1**) the minor allele frequency of the rarer SNP of the two sites, since rarer alleles typically arose more recently and therefore chromosomes bearing these alleles have had less time to participate in recombination events (SI Fig. S6); (**2**) the distance between the two SNPs; and (**3**) the local recombination rate ρ. Since in our study (**1**) and (**2**) are known, we can fit the local recombination rate ρ to F4 using a lookup table based on coalescent simulations that account for the observed allele frequency spectrum.

To build intuition, it is useful to consider first how F4 is expected to vary in a simplified model for recombination. For this simplified initial analysis we use SNP pairs within 500 bp of the 5' end of a CDS start. The reason for initially limiting ourselves to these regions will soon become clear, as recombination is found to vary systematically relative to the positions of genes. The dependence of average F4 on allele frequency and inter-SNP separation in these regions is shown in Fig. 3. If we make the (overly simple) assumption that the 500 bp around a CDS start has a constant recombination rate $\rho/\mu = 3.5$ for half of the predicted genes, and $\rho/\mu = 0$ for the other half, then (using our lookup table) we find reasonable agreement (solid line) with our data (points). This rudimentary model captures the short-range linear variation and eventual saturation in F4 vs. interSNP separation, as well as the dependence of F4 on the lower frequency minor allele of the SNP-pair. Note that the observed F4 is small but non-zero in the limit of zero inter-SNP separation, which is consistent with a small fraction of fourth haplotypes being due to parallel mutations.

To examine the variation of ρ in and around genes we binned all nearby SNP pairs according to their position relative to individual exons and introns in annotated genes. Figure 4 shows the results for genes with 5 or more exons. As also suggested by Figure 2, average recombination rates are highest around the start of genes and decay with distance from the gene start in both coding and non-coding sequence. This observation is known as "polarity" and has previously been reported in studies of specific gene conversion hotspots near the promoters of genes in yeast[4]. The correlation of recombination with CDS starts in *Mimulus* is an average effect, since a substantial amount of recombination occurs at other genomic locations. Conversely, only a fraction of gene starts would need to overlap with hotspots for this effect to be visible, since the recombination activity in many hotspots is much higher than average. The second striking observation in Figure 4 is that average recombination rates are higher in coding exons than in surrounding introns or UTRs. That is, crossover boundaries and/or the ends of gene conversion tracts tend to occur within exons rather than introns. While the 3' ends of genes also exhibit some of

the same features as gene promoters, we found no excess of recombination over and above what could be accounted for by the nearby transcription starts of adjacent genes.

Local recombination rates vary dramatically across the genome. We analyzed sliding windows of 15 non-overlapping pairs of SNPs (which typically span 300-350 bases) and inferred the local recombination rate ρ as that which best matches the observed F4 according to the lookup table based on coalescent simulations. Figure 5 shows typical examples of the recombination landscape at two genomic scales. Genome-wide, 11.7% of non-overlapping SNP-pairs (414,734 of 3,557,964) pass the four-gamete test and so provide evidence for historical recombination. To reproduce this number with a constant recombination rate requires a per-base recombination rate slightly less than, but comparable to the mutation rate ($\rho/\mu = 0.8$, yellow line in Figure 5). The local recombination rate per base, however, is highly variable, with peaks in intensity ("hot spots") interspersed by stretches of zero detectable historic recombination activity ("cold spots"). Such cold spots account for more than 25% of the sampled genome. The remaining ~75% exhibits varying degree of recombination; we find no sharp distinction between "hot" *vs.* "tepid" regions.

Recombination rates in hot spots can be several hundred times the mutation rate, although lower intensity hot spots are much more common. As "tepid" regions with ρ/μ barely larger than the genome-wide average are so much more prevalent than hot spots, any given crossover is more likely to happen in regions of modest ($\rho/\mu <\sim 1.8$) but non-zero recombination, rather than in relatively sparse hot spots with $\rho/\mu \sim 5$ to 100s. The observed sizes of the hot and cold features range from a few hundred (the typical resolution limit of our analysis) to a few thousand bases. In the lower panel of Figure 5 hotspots with $\rho/\mu >=5$ are depicted along with the location of annotated genes in a 60 kb region. Again it can be seen that regions of high recombination activity tend to be located near the start of genes, whereas cold spots are frequently found within genes and often in close vicinity (within ~1kb) of hot spots. About 54% of all annotated genes have $\rho/\mu >= 2$ at the start of the first exon, but only 22% have $\rho/\mu >= 5$. Figure 6 shows the genomic distribution of observed recombination rates along with data based on Monte Carlo simulations using the average value of $\rho/\mu = 0.8$ (SI).

If we define hot spots as local peaks with values $\rho/\mu >= 15$, we detect 3,235 highly reliable hotspots (FP rate = 3.5%). Relaxing the condition to $\rho/\mu >= 5$, we observe 21,501 putative hotspots, but with a high estimated false-positive rate that suggests that only 13,000 of these are *bona fide*. We also catalogued 44,674 cold spots, defined as regions with $\rho/\mu = 0$, including flanking regions with $\rho/\mu <= 0.4$, and requiring lengths of at least 200 bases. The CpG to GpC dinucleotide ratio in all hot and cold spot associated sequences is 1.05 and 0.84, respectively. CpG deficiency is typically caused by higher mutability of the C in methylated CpG[19], suggesting that hot spots are associated with un-methylated CpG islands.

Discussion

We have developed a novel method for finding localized genomic signatures of historic recombination within a highly diverse natural population. The essence of our approach is applying the four-gamete test to pairs of nearby SNPs that are spanned by multiple short sequence reads, each read sampling a short-range haplotype from the population. Importantly, plants are not sequenced or otherwise genotyped individually, but rather are sequenced in unlabeled pools.

Since our analyses rely on alignments of short reads to a somewhat divergent reference sequence, concerns naturally arise about possible mismapping-related artifacts. Our analyses are robust to such possibilities. First, we restrict analysis to reads with high mapping qualities (*i.e.*, unambiguous alignment), require both reads of every read-pair to map properly, and discard the first and last 5 bases of all alignments for SNP-detection purposes. In addition, we exclude sites with coverage outside a specified range. This is discussed in detail in the SI. More subtle effects such as misalignments related to segmental duplications are possible, but do not have significant impact on our key findings. This is evident by the observed dependence in Figure 3 of F4 with distance (and minimum allele frequency). At the zero-distance limit, F4 is as low as 2-4%, which is quantitatively explained by parallel mutations. If a significant fraction of the underlying four-gamete test passing rates had been due to artifacts (e.g., reads mis-mapping between copies of an unmasked repeat, in such a fashion as to simulate the presence of four distinct two-SNP haplotypes), we would expect to see no dependence on distance in any reasonable scenario.

Finally, we demonstrate in Figures S2-S4 in the SI, and related text, that our findings are not due to artifacts related to the faster mutation rates at CpG sites, systematic variation of allele frequencies in the genome, or miscounting of gene conversion tract boundaries due to nested SNP-pairs.

The unprecedented resolution of our analysis allows us to demonstrate that recombination events exhibit (**1**) polarity, (**2**) a preference for the 5' ends of genes, (**3**) , the intron-exon dependence of recombination and (**4**) association of recombination hotspots with CpG islands. In yeast recombination-initiating double strand breaks (DSBs) tend to occur in nucleosome-depleted open chromatin that is often located in gene promoters[20-22]. The boundaries of gene conversion tracts tend to be asymmetrically located around DSBs in yeast[23], suggesting that the density of recombination boundaries should decay with distance from DSB sites. Together, these two features of recombination observed in yeast would generate the same pattern of recombination hotspots and polarity that we observe near CDS starts in *Mimulus*. Promoters and other *cis*-regulatory regions are also characterized by nucleosome-depleted open chromatin in other eukaryotes, including *Arabidopsis* and rice[24-26], which suggests that DSBs in these organisms may show preference for the 5' ends of genes. While the detailed nucleosome organization in *Mimulus* remains to be resolved, it is notable that we observed that hotspots are associated with CpG islands, which in *Arabidopsis* are also negatively correlated with nucleosome density[27].

While our method sheds light on the workings of recombination at its finest scales, it does not provide us with sufficient information at large scales to readily reconcile our results with a standard genetic map, since two-thirds of the assembled genome is unascertainable for SNP pairs probes (SI). For such regions we do not have a direct measure of the recombination rate.

Our finding of recombination variation within genes suggests that the intermediate steps in the process of recombination, such as strand displacement, DNA synthesis, and branch migration, are more likely to stall or terminate in exons than introns. It is notable that high resolution maps of nucleosome positioning in fungi, invertebrates, mammals, and plants[27-30] consistently reveal that exons and their boundaries are enriched for well-positioned nucleosomes and RNA polymerase II, compared to introns, suggesting that well-placed nucleosomes stall RNA synthesis during transcription. Perhaps a similar mechanism tends to stall recombination intermediates in exons, resulting in the exon-intron dependence we observe here. We note that average GC content in *Mimulus* is 42% in exons but only 29% in introns and UTR regions, so there is also a free energy barrier for expansion of strand displacement loops through the more GC-rich regions.

As population surveys of other species with at least 1% nucleotide diversity become available it will be interesting to see if the pattern of recombination initiation occurring in nucleosome-depleted regions is confirmed in other eukaryotes. Although DNA binding of the PRDM9 protein initiates recombination in humans and mice, this appears to be a derived mechanism in mammals that overrides a PRDM9-independent ancestral eukaryotic recombination initiation process[4] which we speculate was predominantly based on the accessibility of the recombination machinery to DNA.

Methods and Materials

Details on this analysis are described in the Supplemental Information

Acknowledgments

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Keeney S, Neale MJ (2006) Initiation of meiotic recombination by formation of DNA double-strand breaks: mechanism and regulation. *Biochemical Society Transactions*, *34*(4), 523-525.

2. Petes TD (2001) Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics*, *2*(5), 360-369.

3. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, *327*(5967), 836-840.

4. Lichten, M, Goldman, AS (1995) Meiotic recombination hotspots. *Annual Reviews of Genetics*, *29*(1), 423-444.

5. Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV (2012) Genetic recombination is directed away from functional genomic elements in mice. *Nature*, *485*(7400), 642-645.

6. Xu X, Hsia AP, Zhang L, Nikolau BJ, Schnable PS (1995) Meiotic recombination break points resolve at high rates at the 5'end of a maize coding sequence. *The Plant Cell Online*, *7*(12), 2151-2161.

7. Miller DE, Takeo S, Nandanan K, Paulson A, Gogol MM et al. (2012) A whole-chromosome analysis of meiotic recombination in Drosophila melanogaster. *G3: Genes| Genomes, and Genetics*, *2*(2), 249-260.

8. Giraut L, Falque M, Drouaud J, Pereira L, Martin OC et al. (2011) Genome-wide crossover distribution in Arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. *PLoS Genetics*, *7*(11), e1002354.

9. Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, *454*(7203), 479-485.

10. Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in Drosophila melanogaster. *PLoS Genetics*, *8*(10), e1002905.

11. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, *160*(3), 1231-1241.

12. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S et al. (2012) Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, *44*(2), 212-216.

13. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science*, *310*(5746), 321-324.

14. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM et al. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, *39*(9), 1151-1155.

15. Paape T, Zhou P, Branca A, Briskine R, Young N, et al. (2012) Fine-scale population recombination rates, hotspots, and correlates of recombination in the Medicago truncatula genome. *Genome Biology and Evolution*, *4*(5), 726-737.

16. Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, *111*(1), 147-164.

17. Fearnhead P, Harding RM, Schneider JA, Myers S, Donnelly P (2004) Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics*, *167*(4), 2067-2081.

18. Goodstein, DM, Shu S, Howson R, Neupane R, Hayes RD et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178-D1186.

19. Scarano E, Iaccarino M, Grippo P, Parisi E (1967) The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. *Proceedings of the National Academy of Sciences of the United States of America*, *57*(5), 1394.

20. Tischfield SE, Keeney S (2012) Scale matters: The spatial correlation of yeast meiotic DNA breaks with histone H3 trimethylation is driven largely by independent colocalization at promoters. *Cell Cycle*, *11*(8), 1496-1503.

21. Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG et al. (2011). A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, *144*(5), 719-731.

22. Berchowitz, LE, Hanlon SE, Lieb JD, Copenhaver GP (2009) A positive but complex association between meiotic double-strand break hotspots and open chromatin in Saccharomyces cerevisiae. *Genome Research*, *19*(12), 2245-2257.

23. Jessop L, Allers T, Lichten M (2005) Infrequent co-conversion of markers flanking a meiotic recombination initiation site in *Saccharomyces cerevisiae*. *Genetics*, *169*(3), 1353-1367.

24. Zhang W, Zhang T, Wu Y, Jiang J (2012) Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *The Plant Cell Online*, *24*(7), 2719-2731.

25 . Arabidopsis, GI (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, *408*(6814), 796.

26. Yu J, Hu S, Wang J, Wong GK, Li S et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, *296*(5565), 79-92.

27. Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H et al. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature*, *466*(7304), 388-392.

28. Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J (2009) Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Research*, *19*(10), 1732-1741.

29. Brogaard K, Xi L, Wang JP, Widom J (2012) A map of nucleosome positions in yeast at base-pair resolution. *Nature*, *486*(7404), 496-501.

30. Jiang C, Pugh BF (2009) A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. *Genome Biol*, *10*(10), R109.

Figure Legends:

Figure 1: Appearance of four haplotypes at a pair of SNP loci by recombination. From a single ancestral sequence (top) a single mutation produces a second haplotype. A second mutation at a nearby site (middle) generates a third haplotype In the population. Finally, a recombination boundary between the two SNP loci (bottom) generates a fourth haplotype. Note that the recombination boundary can be due to a crossover event or a gene conversion tract. A fourth haplotype can also appear due to a parallel mutation (not shown) but this scenario can be distinguished from recombination since parallel mutation at a site should not depend on the distance to the nearest SNP.

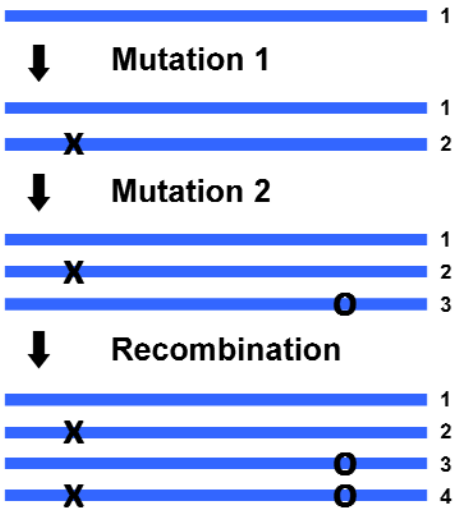

Figure 2: Fraction F4 of non-redundant SNP-pairs passing the four-gamete test, as a function of strand-dependent distance to the closest annotated CDS start.

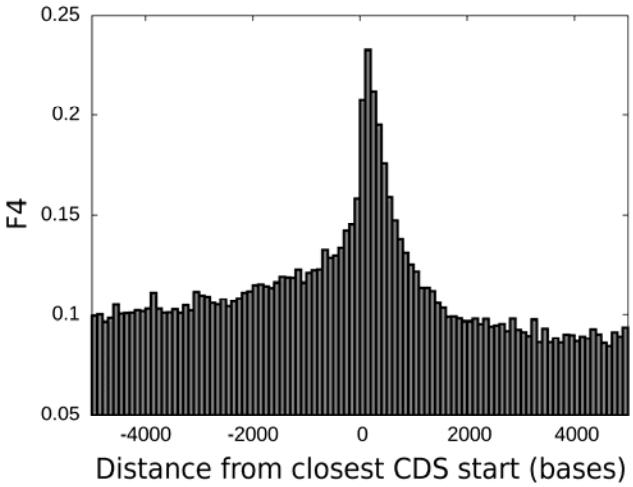

Figure 3: F4 vs. SNP separation for SNP pairs within 500 bp of CDS start. As predicted by a simplified model models (solid lines, see main text), F4 increases with distance between SNPs, d_H, and with frequency of the least common allele. The y-intercept is non-zero due to a modest contribution to F4 from parallel mutations.

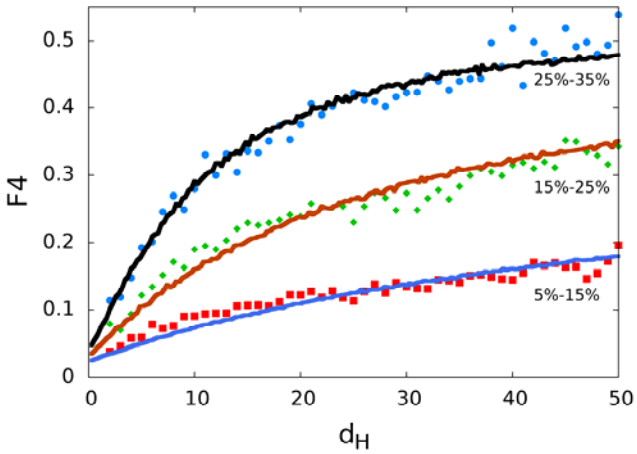

Figure 4: Average recombination rates per base relative to genes with 5 or more exons. Error bars show 95% confidence intervals. A gradient (polarity) in recombination is evident. Also, exons show systematically higher recombination activity than introns and 5' non-coding sequence. Solid line shows average recombination rate within transposable elements and other complex repeats, which are mainly intergenic.

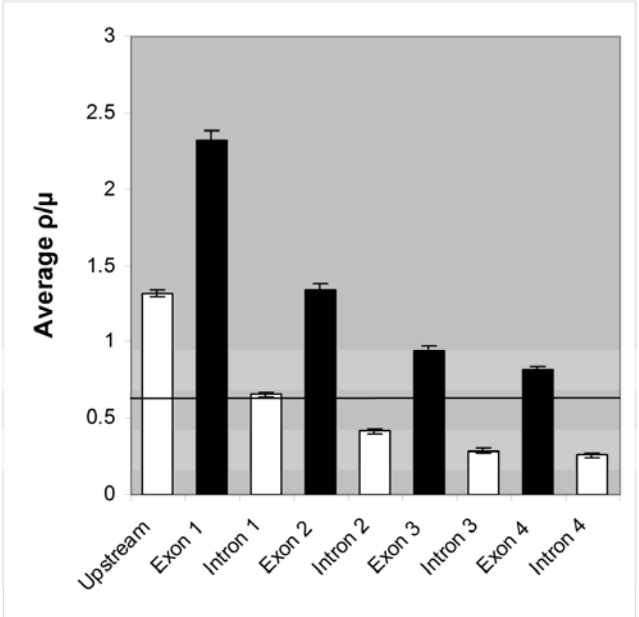

Figure 5: Inferred per-base recombination rate across a 1 Mbp region of the *Mimulus* genome (upper panel) using sliding windows of 15 adjacent non-overlapping SNP pairs. Yellow line shows genome-wide average ($\rho/\mu \sim 0.8$) and dashed line indicates an arbitrary cutoff at $\rho/\mu = 5$ for hotspot detection. Lower panel focuses on a 60 kbp-region, showing association of hotspots with 5' ends of protein-coding genes.

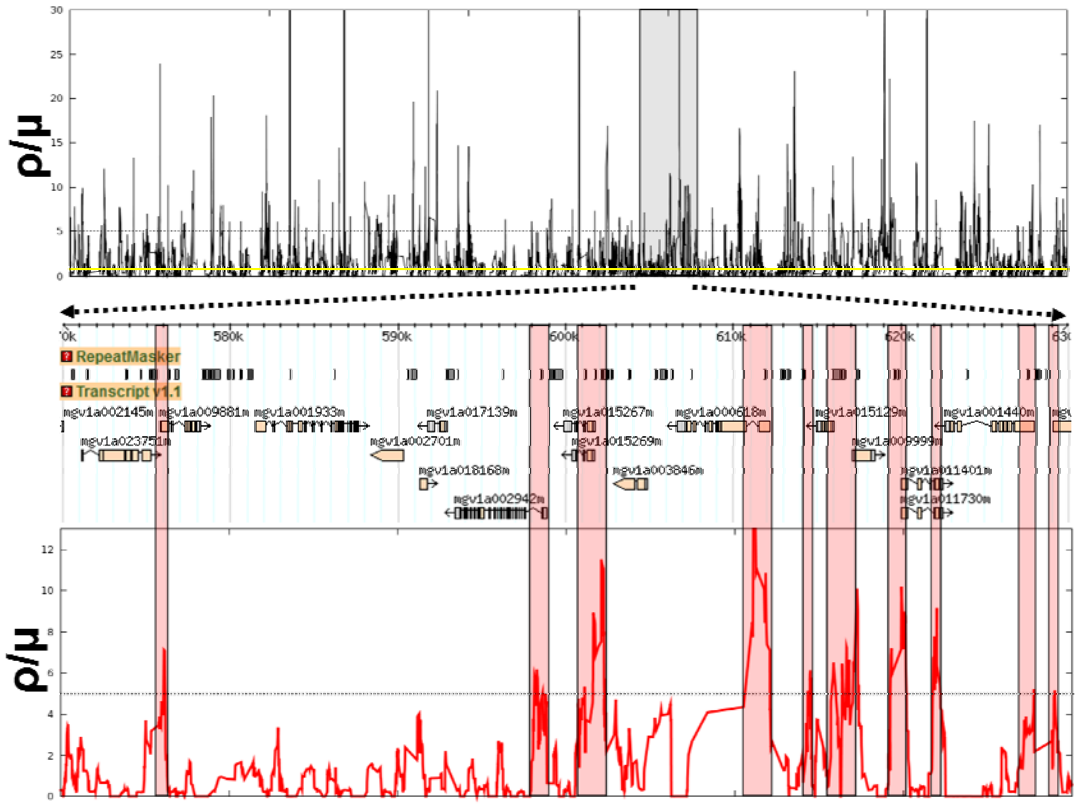
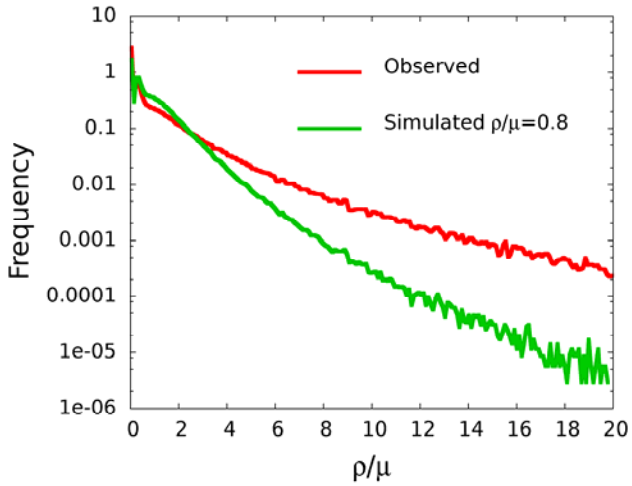
Figure 6: Genome-wide distribution of inferred recombination rates (red) compared to expected distribution under a constant recombination rate equal to the genomic average. Specificity increases with increasing ρ/μ at the expense of sensitivity.