Supporting Information

Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*

Andrew H. Chan[1,*], Paul A. Jenkins[1,*], Yun S. Song[1,2,**]

[1] Computer Science Division, University of California, Berkeley, CA, USA
[2] Department of Statistics, University of California, Berkeley, CA, USA

[*] These authors contributed equally to this work
[**] Corresponding author e-mail: yss@cs.berkeley.edu

Text S1

Two-locus recursion relation

Suppose we sample n haplotypes, observing their alleles at each of two loci and obtaining configuration $\boldsymbol{n} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. Here $\boldsymbol{c} = (c_{ij})$ is a matrix of the counts of haplotypes for which both alleles were observed; c_{ij} is the number of haplotypes with allele i at the first locus and allele j at the second locus. We also allow for the possibility that a haplotype had data missing at one locus: $\boldsymbol{a} = (a_i)_{i=1\ldots,K}$ is the vector of counts of haplotypes with allele i observed at the first locus and missing data at the second locus, and $\boldsymbol{b} = (b_j)_{j=1,\ldots,L}$ is the vector of counts of haplotypes with allele j observed at the second locus and missing data at the first locus. Further, let:

$$a = \sum_{i=1}^{K} a_i, \quad c_{i\cdot} = \sum_{j=1}^{L} c_{ij}, \quad c = \sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij},$$
$$b = \sum_{j=1}^{L} b_j, \quad c_{\cdot j} = \sum_{i=1}^{K} c_{ij}, \quad n = a + b + c.$$

The probability that, when we sample n haplotypes in some fixed order, we obtain a set consistent with configuration $\boldsymbol{n}$, is denoted by $q(\boldsymbol{n}; \theta_A, \theta_B, \rho)$. This probability is a function of θ_A, θ_B, and ρ: the mutation rates at the two loci, and the recombination rate between them. The respective mutation transition matrices at the two loci, which we denote $\boldsymbol{P}^A$ and $\boldsymbol{P}^B$, are fixed. A system of equations for $q(\boldsymbol{n}; \theta_A, \theta_B, \rho)$ is given in [1]. We denote by $q(\boldsymbol{n}, s_1, s_2; \theta_A, \theta_B, \rho)$ the joint probability of obtaining $\boldsymbol{n}$ with the events that there were precisely s_1 mutations in the history of the sample at the first locus and s_2 mutations in the history of the sample at the second locus. The corresponding system of equations for $q(\boldsymbol{n}, s_1, s_2; \theta_A, \theta_B, \rho)$ is:

$$\begin{aligned}
&[n(n-1) + \theta_A(a+c) + \theta_B(b+c) + \rho c]\, q((\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}), s_1, s_2; \theta_A, \theta_B, \rho) = \\
&\sum_{i=1}^{K} a_i(a_i - 1 + 2c_{i\cdot})\, q((\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b}, \boldsymbol{c}), s_1, s_2; \theta_A, \theta_B, \rho) + \sum_{j=1}^{L} b_j(b_j - 1 + 2c_{\cdot j})\, q((\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c}), s_1, s_2; \theta_A, \theta_B, \rho) \\
&+ \sum_{i=1}^{K}\sum_{j=1}^{L} [c_{ij}(c_{ij} - 1)\, q((\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij}), s_1, s_2; \theta_A, \theta_B, \rho) + 2a_i b_j\, q((\boldsymbol{a} - \boldsymbol{e}_i, \boldsymbol{b} - \boldsymbol{e}_j, \boldsymbol{c} + \boldsymbol{e}_{ij}), s_1, s_2; \theta_A, \theta_B, \rho)] \\
&+ \theta_A \sum_{i=1}^{K} \left[\sum_{j=1}^{L} c_{ij} \sum_{t=1}^{K} P^A_{ti}\, q((\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij} + \boldsymbol{e}_{tj}), s_1 - 1, s_2; \theta_A, \theta_B, \rho) \right.
\end{aligned}$$

$$
\begin{aligned}
&\quad + a_i \sum_{t=1}^{K} P^A_{ti} q((\boldsymbol{a} - \boldsymbol{e}_i + \boldsymbol{e}_t, \boldsymbol{b}, \boldsymbol{c}), s_1 - 1, s_2; \theta_A, \theta_B, \rho) \Bigg] \\
&+ \theta_B \sum_{j=1}^{L} \Bigg[\sum_{i=1}^{K} c_{ij} \sum_{t=1}^{L} P^B_{tj} q((\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} - \boldsymbol{e}_{ij} + \boldsymbol{e}_{it}), s_1, s_2 - 1; \theta_A, \theta_B, \rho) \\
&\quad + b_j \sum_{t=1}^{L} P^B_{tj} q((\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j + \boldsymbol{e}_t, \boldsymbol{c}), s_1, s_2 - 1; \theta_A, \theta_B, \rho) \Bigg] \\
&+ \rho \sum_{i=1}^{K} \sum_{j=1}^{L} c_{ij} q((\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}), s_1, s_2; \theta_A, \theta_B, \rho), \qquad (1)
\end{aligned}
$$

where $\boldsymbol{e}_{ij}$ is a unit matrix whose (i, j)th entry is one and the rest are zero. As before, we suppose that we know the identity of the ancestral alllele at each locus, say λ_A and λ_B at locus A and B, respectively. Then we replace the relevant instances of (1) with the following:

$$
\begin{aligned}
q((\boldsymbol{0}, \boldsymbol{b}, \boldsymbol{e}_{ij}), s_1, s_2; \theta_A, \theta_B, \rho) &= \begin{cases} q((\boldsymbol{0}, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{0}), 0, s_2; \theta_A, \theta_B, \rho) & \text{if } i = \lambda_A \text{ and } s_1 = 0, \\ 0 & \text{otherwise,} \end{cases} \\
q((\boldsymbol{a}, \boldsymbol{0}, \boldsymbol{e}_{ij}), s_1, s_2; \theta_A, \theta_B, \rho) &= \begin{cases} q((\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}), s_1, 0; \theta_A, \theta_B, \rho) & \text{if } j = \lambda_B \text{ and } s_2 = 0, \\ 0 & \text{otherwise,} \end{cases} \\
q((\boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}), s_1, s_2; \theta_A, \theta_B, \rho) &= \begin{cases} 1 & \text{if } i = \lambda_A \text{ and } s_1 = s_2 = 0, \\ 0 & \text{otherwise,} \end{cases} \\
q((\boldsymbol{0}, \boldsymbol{e}_j, \boldsymbol{0}), s_1, s_2; \theta_A, \theta_B, \rho) &= \begin{cases} 1 & \text{if } j = \lambda_B \text{ and } s_1 = s_2 = 0, \\ 0 & \text{otherwise.} \end{cases} \qquad (2)
\end{aligned}
$$

Padé summation

Modifications to the approach described in [2] are made, following from the boundary conditions given above. These can be converted into modifications of entries of the dynamic programming tables given in [2]. For example, using (2) we have that

$$
\begin{aligned}
q((\boldsymbol{a}, \boldsymbol{0}, \boldsymbol{e}_{i\lambda_B}), 1, 0; \theta_A, \theta_B, \rho) &= q((\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}), 1, 0; \theta_A, \theta_B, \rho) \\
&= q(\boldsymbol{a} + \boldsymbol{e}_i, 1; \theta_A) + \frac{0}{\rho} + \frac{0}{\rho^2} + \ldots,
\end{aligned}
$$

where $q(\boldsymbol{a} + \boldsymbol{e}_i, 1; \theta_A)$ is the one-locus solution given by equation (3) in the main text. Notice that this expansion is in fact independent of ρ, from which it follows (by comparison with eq. (3.7) of [2]) that a number of entries in the dynamic programming tables are modified. For example, the second row in the dynamic programming table for the configuration $(\boldsymbol{a}, \boldsymbol{0}, \boldsymbol{e}_{i\lambda_B})$ is set to zero. Other boundary conditions may be interpreted in a similar fashion.

Ancestral allele estimation

Suppose we have one genomic sequence of *D. simulans* and n sequences of *D. melanogaster*. Let S represent the sequence of *D. simulans* and $M^{(k)}$ represent the sequence of the kth *D. melanogaster*, where S_l denotes the lth base of the sequence, and $S_{\hat{l}}$ represents the sequence with the exclusion of the lth base. Given $(S, M^{(k)})$, let $T_l^{(k)}$ be the time to the most recent common ancestor (TMRCA) at locus l; $f_l^{(k)}(t \mid M_{\hat{l}}, S_{\hat{l}})$ be the density of the TMRCA conditioned on both their sequences but *excluding* the lth locus; and $A_l^{(k)}$ be the ancestral allele at the lth locus, i.e., the allele of the most recent common ancestor (MRCA).

To compute the distribution on the ancestral allele at the lth locus conditioned on $M^{(k)}$ and S, we use

Bayes' theorem to obtain

$$\begin{aligned}
&\mathbb{P}(A_l^{(k)} = i \mid M^{(k)}, S) \\
&= \frac{\int_0^\infty p(A_l^{(k)} = i, M^{(k)}, S, T_l^{(k)} = t)dt}{\mathbb{P}(M^{(k)}, S)} \\
&= \frac{\int_0^\infty \mathbb{P}(M_l^{(k)}, S_l^{(k)} \mid A_l^{(k)} = i, T_l^{(k)})p(A_l^{(k)} = i, T_l^{(k)} = t)dt}{\mathbb{P}(M^{(k)}, S)} \\
&= \frac{\int_0^\infty \mathbb{P}(M_l^{(k)} \mid A_l^{(k)} = i, T_l^{(k)} = t)\mathbb{P}(S_l \mid A_l^{(k)} = i, T_l^{(k)} = t)\mathbb{P}(A_l^{(k)} = i) f_l^{(k)}(t \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})dt}{\sum_j \int_0^\infty \mathbb{P}(M_l^{(k)} \mid A_l^{(k)} = j, T_l^{(k)} = t)\mathbb{P}(S_l \mid A_l^{(k)} = j, T_l^{(k)} = t)\mathbb{P}(A_l^{(k)} = j) f_l^{(k)}(t \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})dt}.
\end{aligned} \tag{3}$$

In equation (3), the prior on the ancestral allele at locus l, $\mathbb{P}(A_l^{(k)} = i)$, is given by the stationary distribution of the allele frequencies from the mutation matrix $\boldsymbol{P}$. (In the above, p denotes a joint probability of discrete events together with the density for $T_l^{(k)}$.) The density on the TMRCA, $f_l^{(k)}(t \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})$, is estimated using Li & Durbin's `psmc` [3]. In practice, we use `psmc` to compute $f_l^{(k)}(t \mid M^{(k)}, S)$ and assume $f_l^{(k)}(t \mid M^{(k)}, S) \approx f_l^{(k)}(t \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})$.

The remaining two probabilities, $\mathbb{P}(M_l^{(k)} \mid A_l^{(k)} = i, T_l^{(k)} = t)$ and $\mathbb{P}(S_l \mid A_l^{(k)} = i, T_l^{(k)} = t)$, are computed as follows. For the computation of $\mathbb{P}(M_l^{(k)} \mid A_l^{(k)} = i, T_l^{(k)} = t)$, let $\boldsymbol{P} = (P_{ij})$ denote the mutation matrix, and let $r_l^{(k)}$ specify the number of mutations that have occurred at the lth locus of the kth *D. melanogaster* sequence during time $T_l^{(k)}$. Then we have

$$\begin{aligned}
\mathbb{P}(M_l^{(k)} = j \mid A_l^{(k)} = i, T_l^{(k)} = t) &= \sum_{s=0}^{\infty} \mathbb{P}(r_l^{(k)} = s \mid T_l^{(k)} = t)(\boldsymbol{P}^s)_{ij} \\
&= \sum_{s=0}^{\infty} \left(\frac{\theta t}{2}\right)^s \frac{e^{-\theta t/2}}{s!}(\boldsymbol{P}^s)_{ij} \\
&= \sum_{s=0}^{\infty} \left[\left(\frac{\theta t}{2}\boldsymbol{P}\right)^s\right]_{ij} \frac{e^{-\theta t/2}}{s!} \\
&= \left[e^{\frac{\theta t}{2}(\boldsymbol{P}-\mathbf{I})}\right]_{ij},
\end{aligned}$$

where $\mathbf{I}$ is the identity matrix with the same dimensions as $\boldsymbol{P}$. The computation for $\boldsymbol{P}(S_l \mid A_l^{(k)} = j, T_l^{(k)} = t)$ is analogous.

After computing $\mathbb{P}(A_l^{(k)} = i \mid M^{(k)}, S)$ for every k and given l, we heuristically aggregate these pairwise probabilities to estimate $\mathbb{P}(A_l^{(k)} = i \mid M^{(1)}, \ldots, M^{(n)}, S)$ as follows. Let $\bar{t}_l^{(k)}$ be the posterior mean of $f_l^{(k)}(t \mid M^{(k)}, S)$, i.e.:

$$\bar{t}_l^{(k)} = \int_0^\infty t f_l^{(k)}(t \mid M^{(k)}, S)dt,$$

and define $\tau_l = \max_k \bar{t}_l^{(k)}$. We approximate $\mathbb{P}(A_l^{(k)} = i \mid M^{(1)}, \ldots, M^{(n)}, S)$ as

$$\mathbb{P}(A_l^{(k)} = i \mid M^{(1)}, \ldots, M^{(n)}, S) \approx \frac{\sum_{k=1}^n \mathbb{P}(A_l^{(k)} = i \mid M^{(k)}, S) f_l^{(k)}(\tau_l \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})}{\sum_j \sum_{k=1}^n \mathbb{P}(A_l^{(k)} = j \mid M^{(k)}, S) f_l^{(k)}(\tau_l \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})},$$

which is a weighted average of $\mathbb{P}(A_l^{(k)} = i \mid M^{(k)}, S)$ over k, weighted by the density of the TMRCA evaluated at τ_l for each k. This averaging ought to mitigate effects such as genotyping errors and incomplete lineage sorting in individual *D. melanogaster* genomes.

References

1. Jenkins PA, Song YS (2009) Closed-form two-locus sampling distributions: accuracy and universality. Genetics 183: 1087–1103.

2. Jenkins PA, Song YS (2012) Padé approximants and exact two-locus sampling distributions. Annals of Applied Probability 22: 576–607.

3. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

Supporting Figures

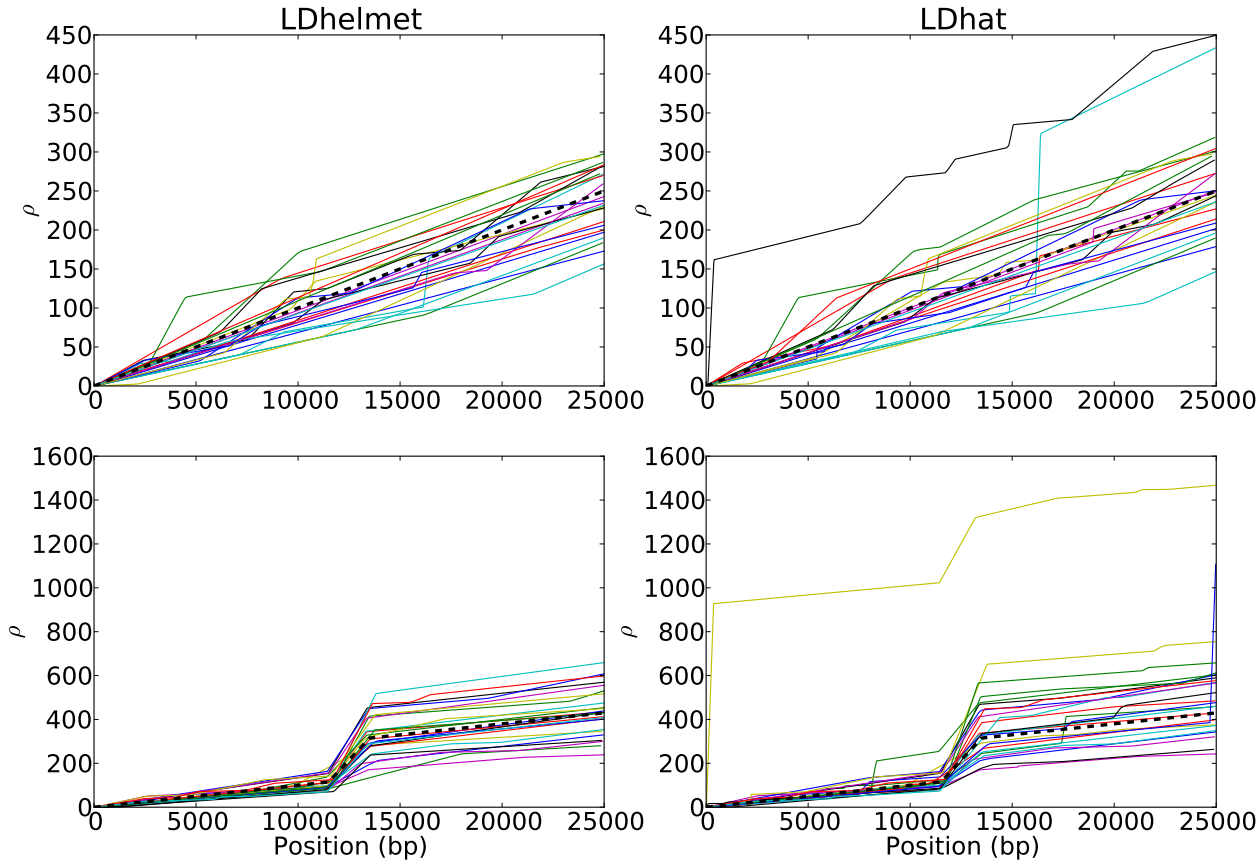

Figure S1: **Comparison of the cumulative recombination maps of `LDhelmet` and `LDhat` for 25 datasets simulated under neutrality** In each plot, different colors represent the cumulative recombination maps for different datasets. The datasets in these plots correspond to the same datasets used in Figure 1. The thick dashed line indicates the true cumulative recombination map for the given recombination landscape. The left and right columns show the estimated recombination maps of `LDhelmet` and `LDhat`, respectively, using the same block penalty of 50. (First Row) Each dataset was simulated with a constant recombination rate of 0.01 per bp. (Second Row) Each dataset was simulated with a hotspot of width 2 kb starting at location 11 kb. The background recombination rate was 0.01 per bp, while the hotspot intensity was 10× the background rate, i.e., 0.1 per bp. The cumulative maps are shown in their entirety, including potential edge effects.

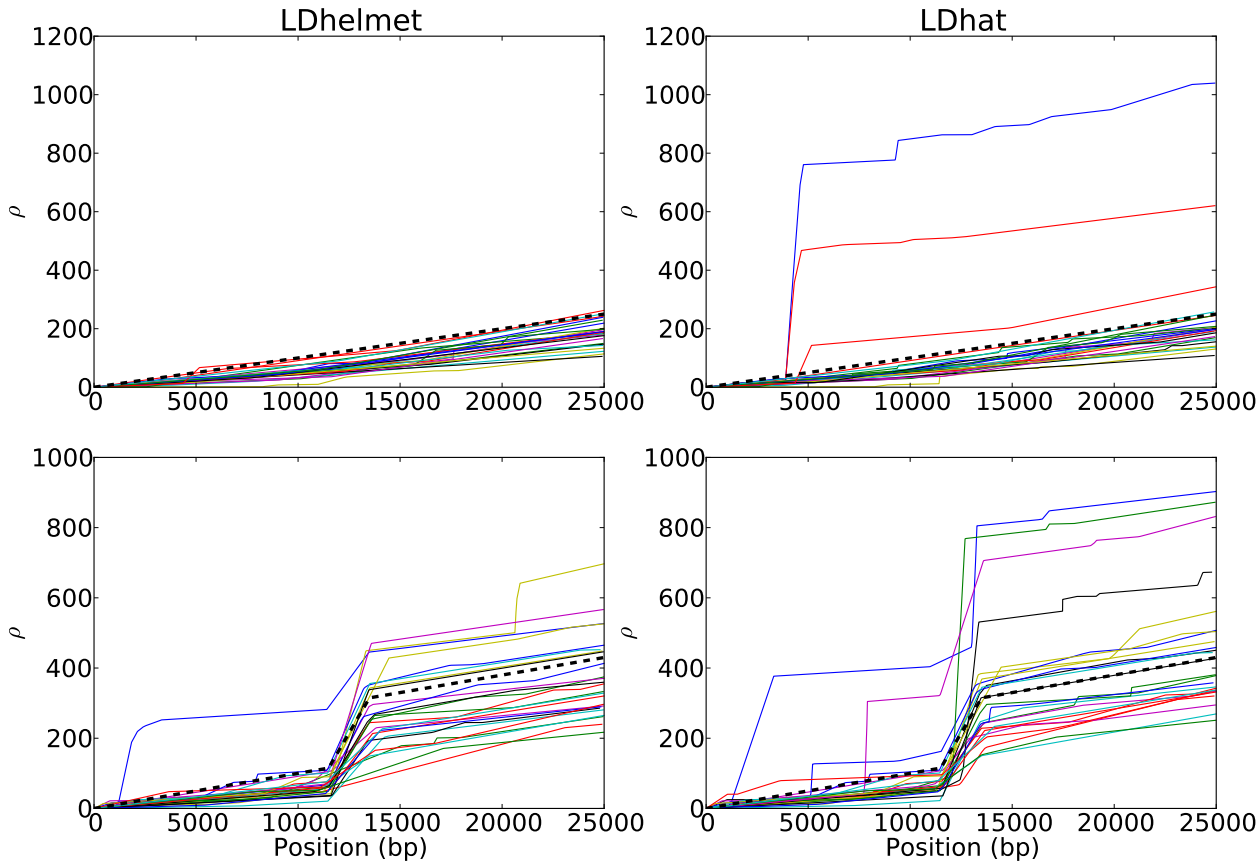

Figure S2: **Comparison of the cumulative recombination maps of `LDhelmet` and `LDhat` for 25 datasets simulated under strong positive selection.** In each plot, different colors represent the results for different datasets. The datasets in these plots correspond to the same datasets used in Figure 3. The thick dashed line indicates the true cumulative recombination map for the given recombination landscape. The left and right columns show the estimated recombination maps of `LDhelmet` and `LDhat`, respectively, using the same block penalty of 50. In each simulation, the selected site was placed at position 5 kb and the population-scaled selection coefficient was set to 1000. The fixation time of the selected site was 0.01 coalescent units in the past. The same scenarios of recombination patterns as in Figure 1 were considered: (First Row) with a constant recombination rate of 0.01 per bp, and (Second Row) with a hotspot of width 2 kb starting at location 11.5 kb. The background recombination rate was 0.01 per bp, while the hotspot intensity was 10× the background rate, i.e., 0.1 per bp. The cumulative maps are shown in their entirety, including potential edge effects.

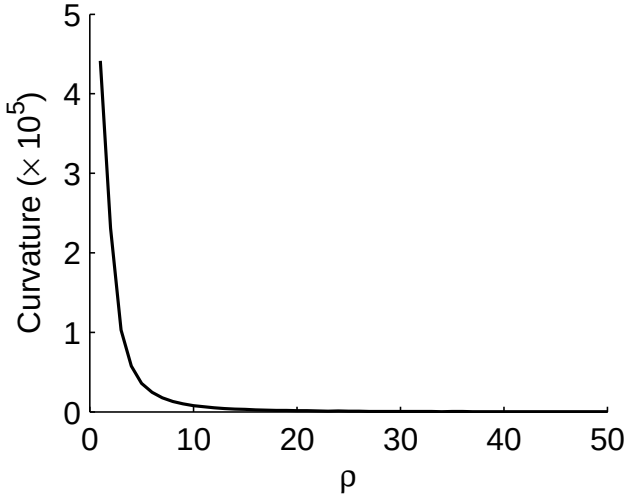

Figure S3: Fisher's information for two-locus samples of size $n = 37$ using lookup tables for $\theta = 0.006$ and under the infinite-sites assumption. The ancestral haplotype is assumed to be known.

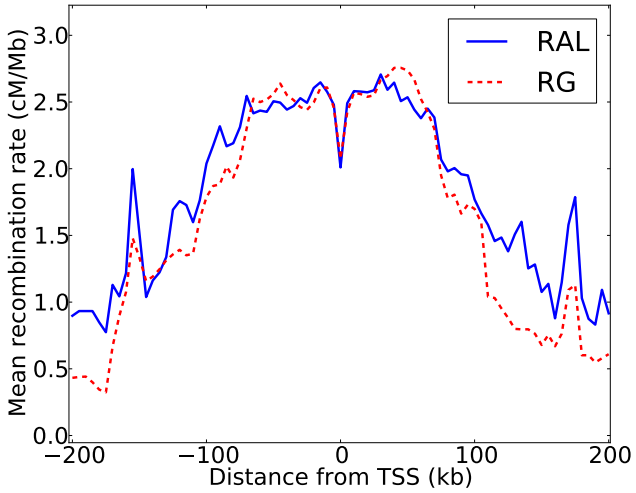

Figure S4: **Distribution of recombination rates relative to transcription start sites.** Plots for RAL (solid) and RG (dashed) of the average estimated recombination rate as a function of distance from the midpoint of the nearest transcription start site (TSS) to the left (negative x-axis) and to the right (positive x-axis) of every base. A 5-kb averaging window was used to smooth the estimates.

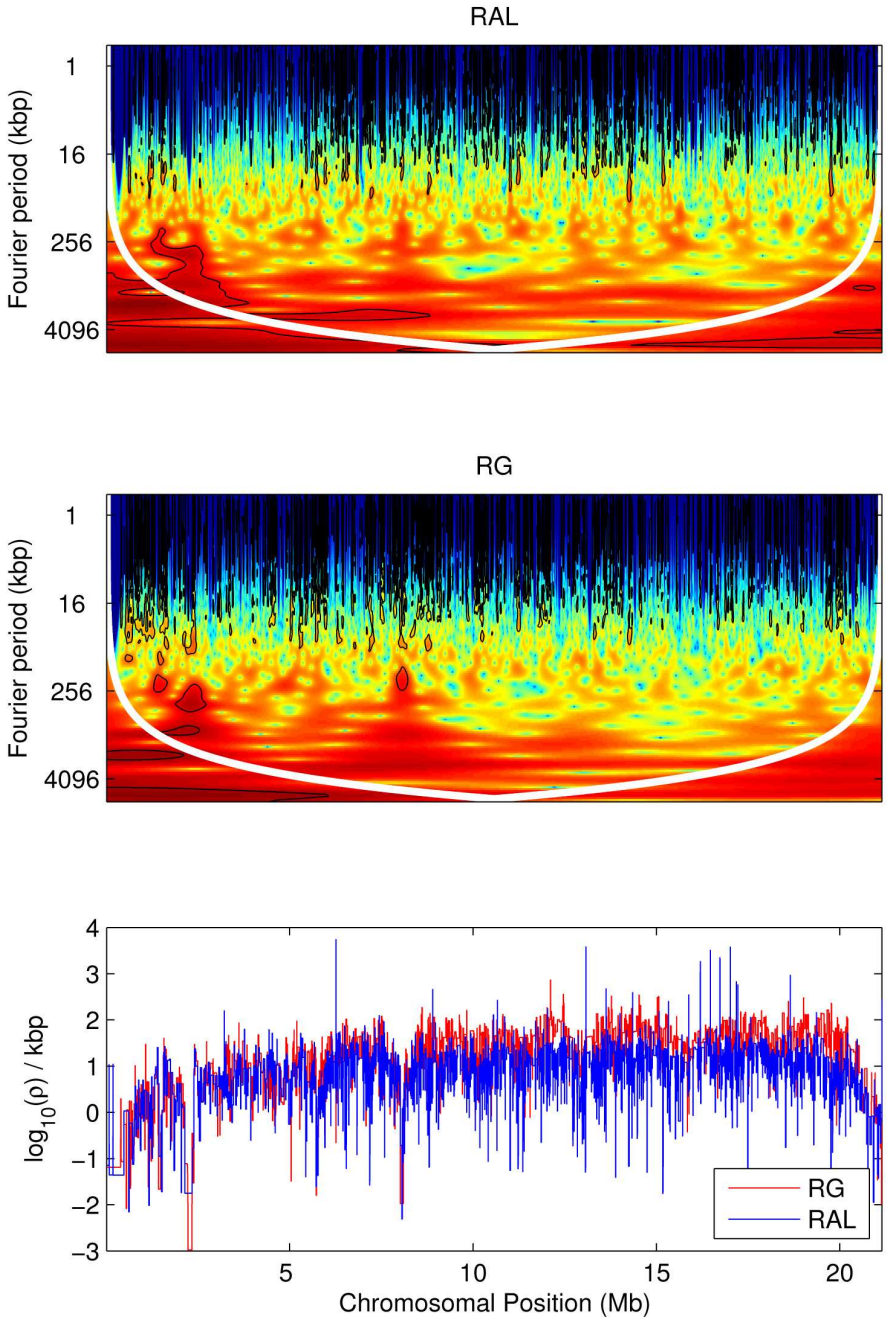

Figure S5: **Local wavelet power spectrum of recombination rate variation in chromosome arm 2R.** A power spectrum is shown for RAL and RG. Black contours denote regions of significant power at the 5% level, and the white contour denotes the cone of influence. Color scale is relative to a white-noise process with the same variance. The lower panels shows estimates of the corresponding genetic maps.

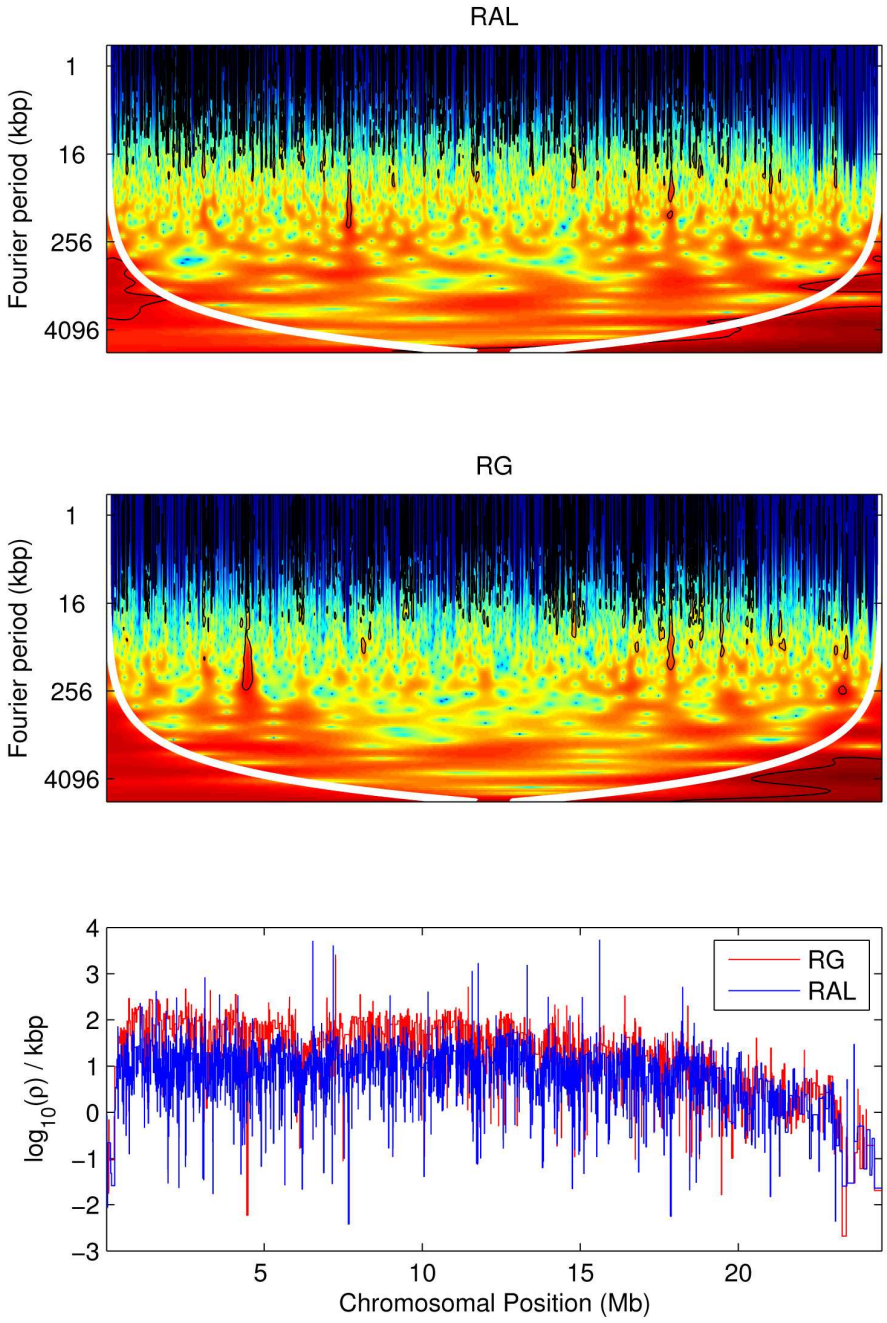

Figure S6: **Local wavelet power spectrum of recombination rate variation in chromosome arm 3L.** A power spectrum is shown for RAL and RG. Black contours denote regions of significant power at the 5% level, and the white contour denotes the cone of influence. Color scale is relative to a white-noise process with the same variance. The lower panels shows estimates of the corresponding genetic maps.

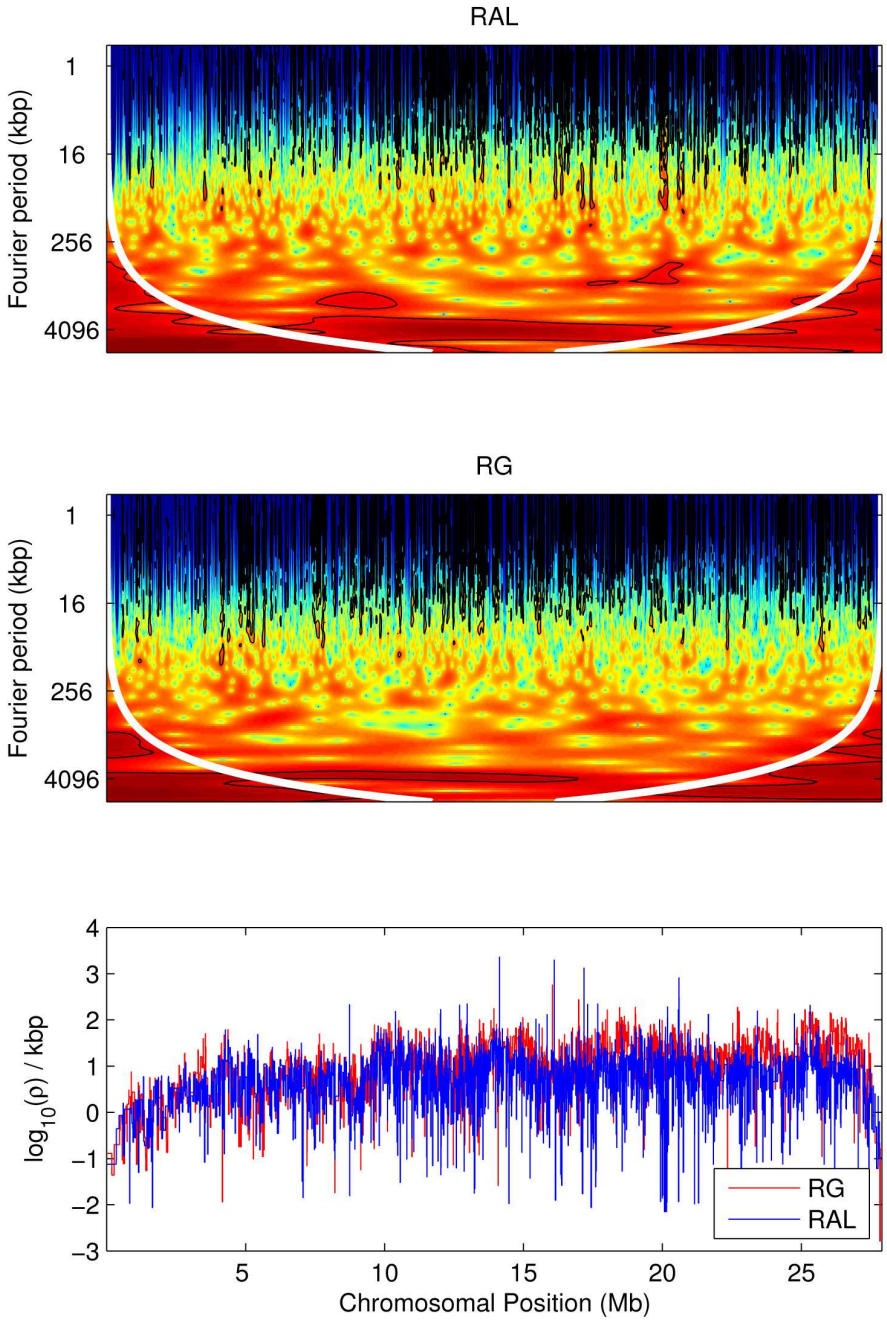

Figure S7: **Local wavelet power spectrum of recombination rate variation in chromosome arm 3R.** A power spectrum is shown for RAL and RG. Black contours denote regions of significant power at the 5% level, and the white contour denotes the cone of influence. Color scale is relative to a white-noise process with the same variance. The lower panels shows estimates of the corresponding genetic maps.

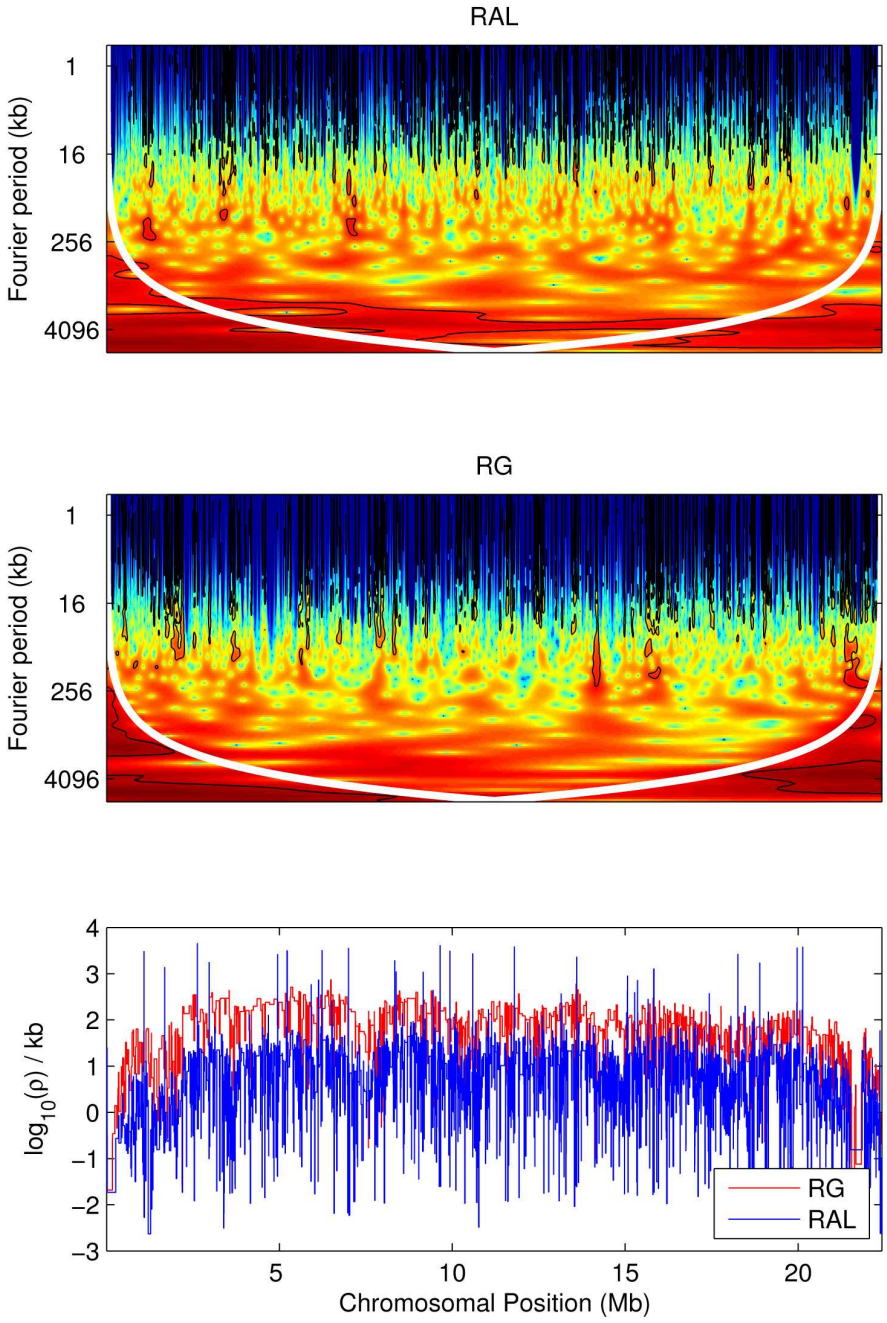

Figure S8: **Local wavelet power spectrum of recombination rate variation in chromosome X.** A power spectrum is shown for RAL and RG. Black contours denote regions of significant power at the 5% level, and the white contour denotes the cone of influence. Color scale is relative to a white-noise process with the same variance. The lower panels shows estimates of the corresponding genetic maps.

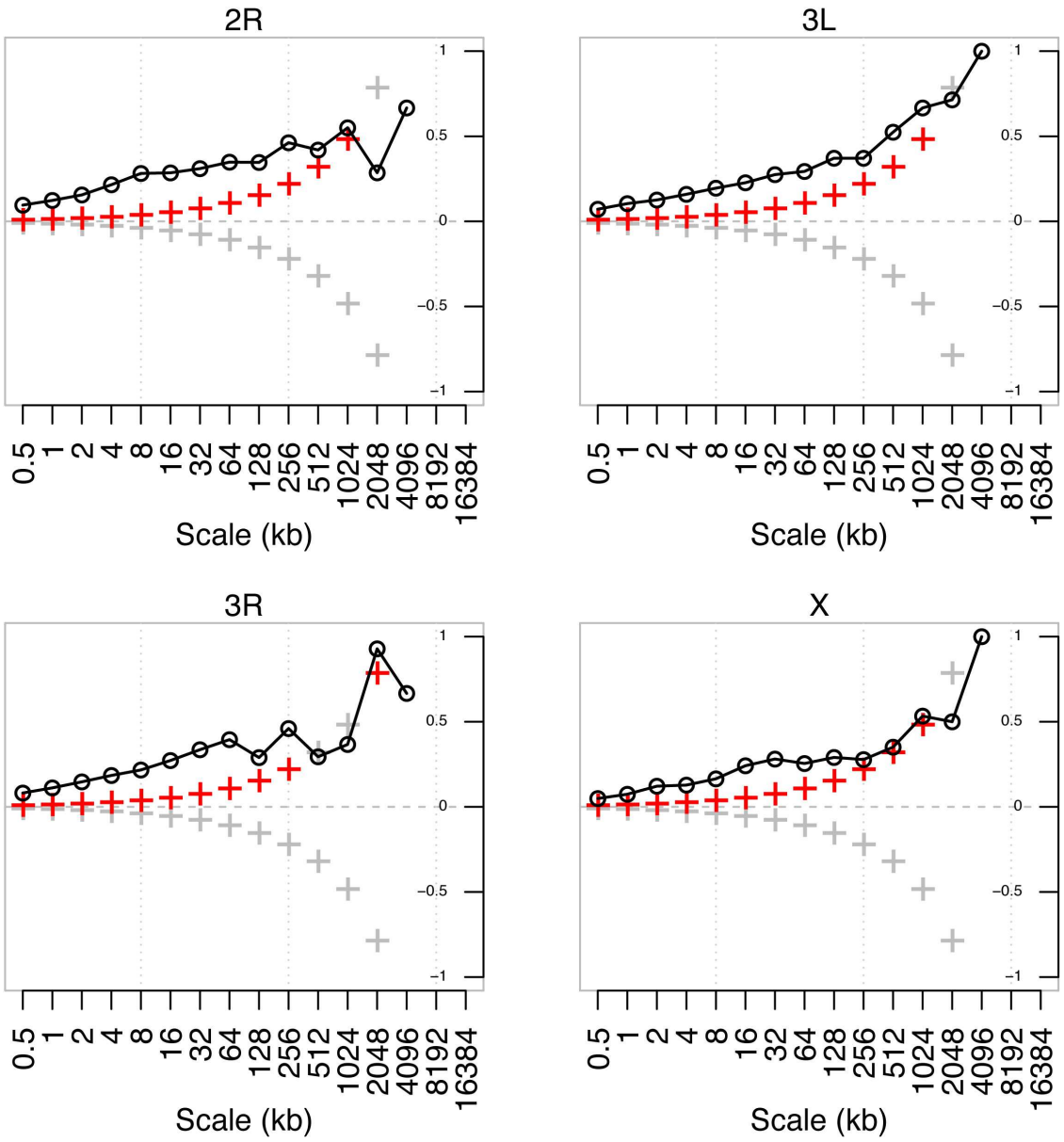

Figure S9: **Pairwise correlation of detail wavelet coefficients of RAL and RG recombination maps for chromosome arms 2R, 3L, 3R, and X.** Black circles denote Kendall's rank correlation between pairs of detail coefficients at each scale. Crosses denote the correlation that would be required for significance at the 1% level in a two-tailed test; red crosses are those scales at which the correlation is in fact significant.

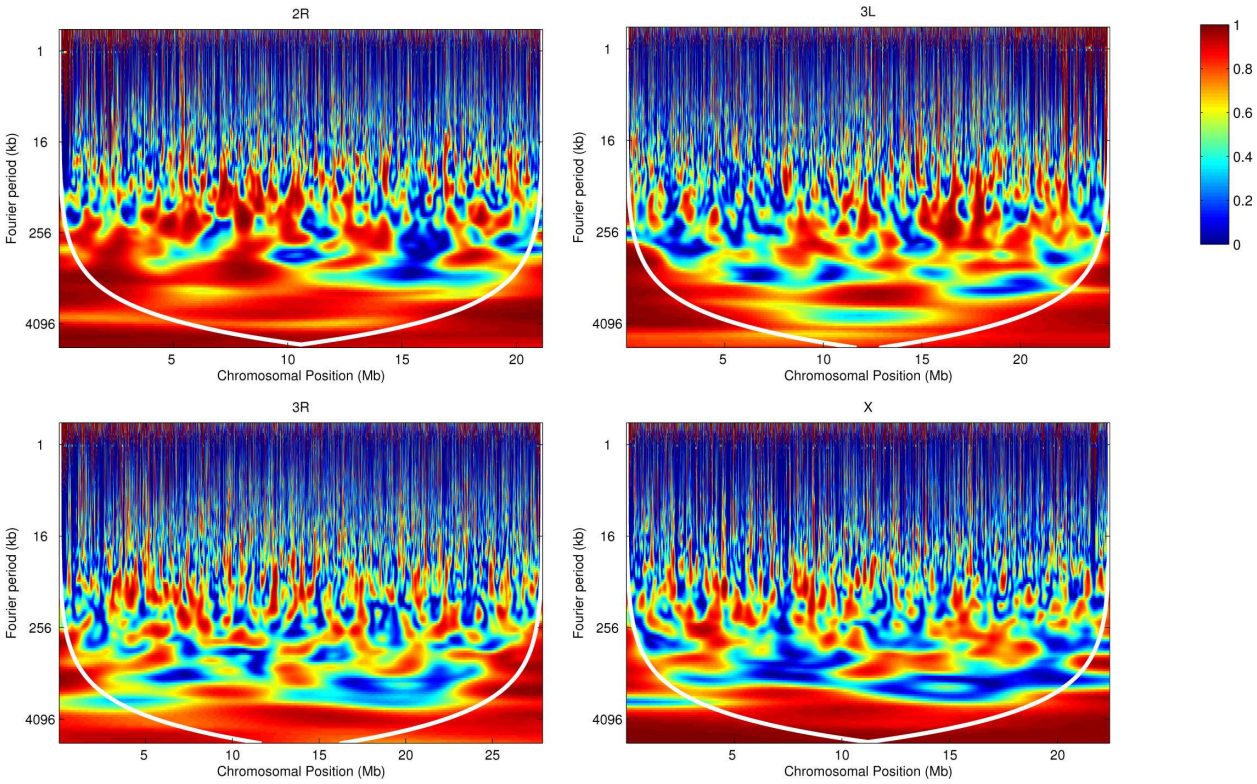

Figure S10: **Wavelet coherence analysis comparing RAL against RG for chromosome arms 2R, 3L, 3R, X.** The cone of influence is shown in white.

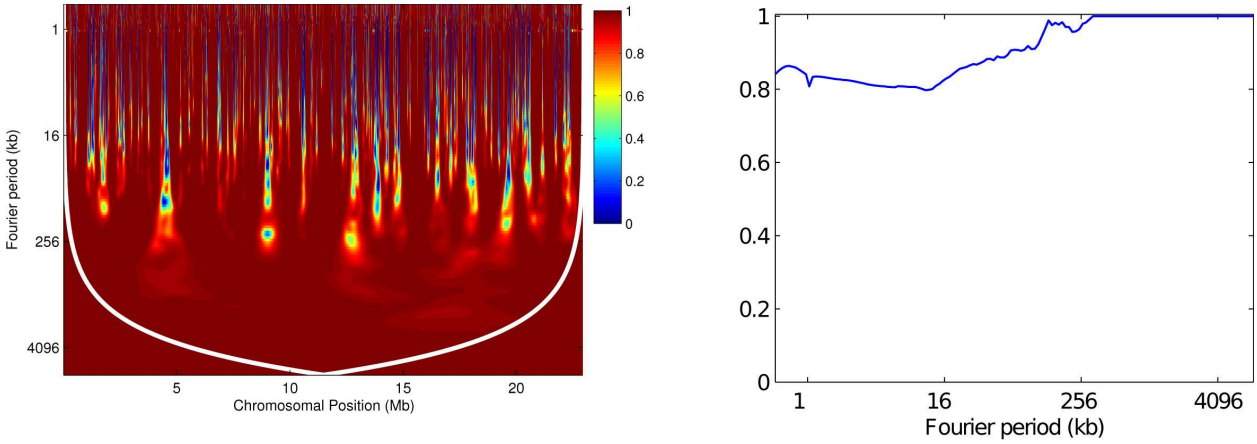

Figure S11: **Positive control for wavelet coherence analysis.** (Left): Coherence plot for two independent estimates of the recombination map across chromosome arm 2L using the same (RG) dataset. (Right): The fraction of chromosome arm 2L with significantly high coherence at the 5% level, at each scale.

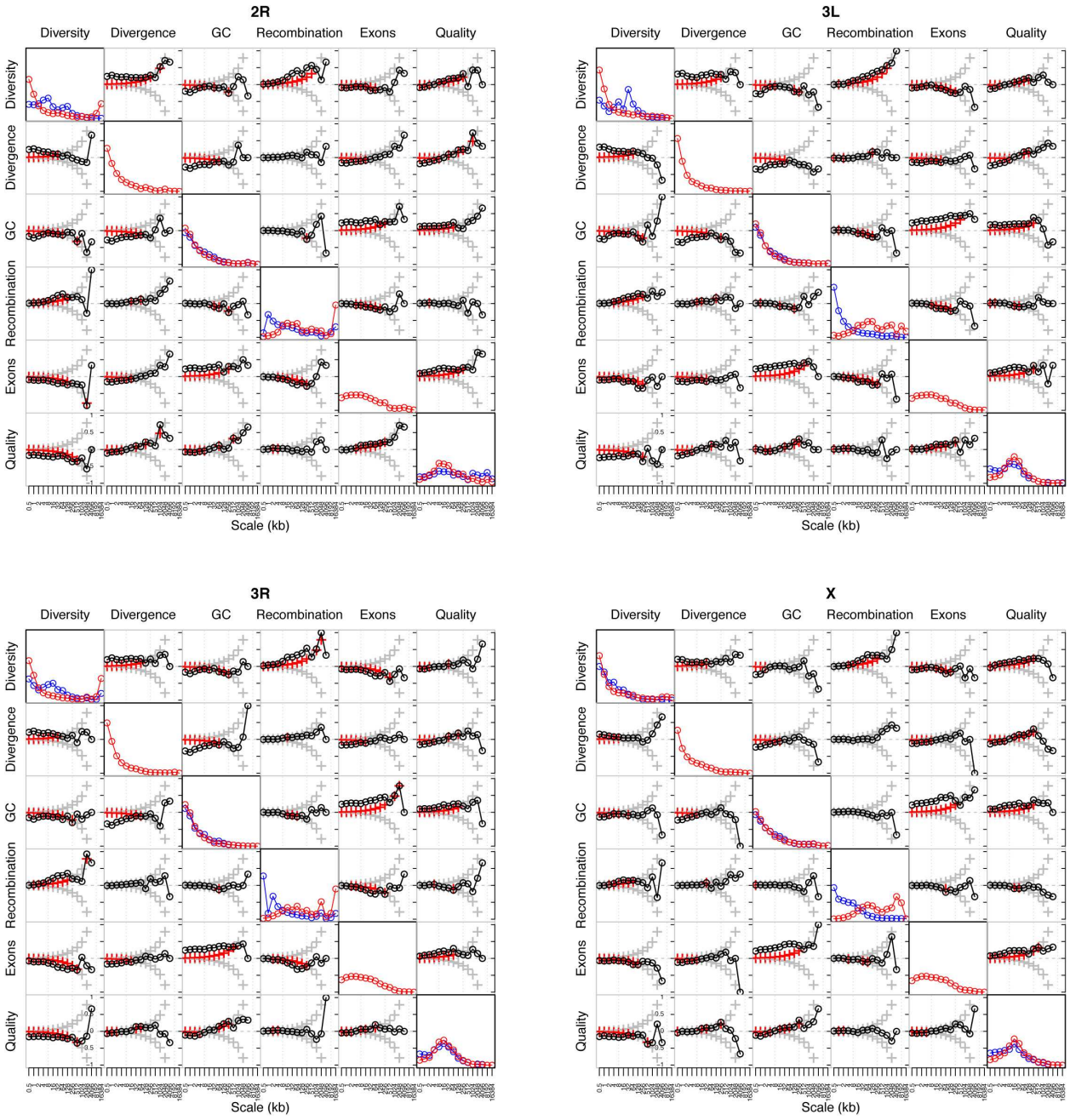

Figure S12: **Global wavelet power spectrum and pairwise correlations of detail wavelet coefficients of RAL and RG data for chromosome arms 2R, 3L, 3R, and X.** Diagonal plots show the global wavelet power spectrum of each feature of the RAL (blue) and RG (red) data. Off-diagonal plots show Kendall's rank correlation between pairs of detail coefficients at each scale, with respect to the wavelet decomposition of the two indicated features. Crosses denote the correlation that would be required for significance at the 1% level in a two-tailed test; red crosses are those scales at which the correlation is in fact significant. The lower left triangle and upper right triangle of plots correspond to RAL and RG, respectively.

A

RAL

Quality	6.58	15.49	16.81	8.92	14.46	8.80	7.43	3.41	4.87	8.08	1.87	1.83
Exons	0.85	0.81	1.86	2.04	6.98	8.46	6.06	4.21	4.16	3.86	1.17	0.38
GC	4.04	1.76	1.80	1.45	3.00	2.28	0.22	0.60	0.35	0.71	0.08	0.21
Divergence	0.17	0.04	0.07	1.91	1.11	0.53	0.54	0.69	0.16	0.31	0.14	0.03
Diversity	4.60	2.95	2.73	5.35	8.12	6.47	12.05	6.43	2.13	1.87	1.48	2.09
Adjusted r^2	0.00	0.00	0.01	0.02	0.07	0.11	0.24	0.27	0.31	0.58	0.37	0.68
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

RG

Quality	1.27	4.36	1.33	0.32	0.11	0.24	1.53	0.32	0.23	2.46	0.01	0.20
Exons	0.14	2.28	2.66	1.39	0.93	4.25	0.94	0.89	1.03	1.64	0.07	0.19
GC	2.00	3.36	3.43	0.31	0.60	0.40	0.01	1.77	0.07	0.01	0.94	0.00
Divergence	0.42	1.29	0.20	1.14	0.01	0.13	0.15	3.18	0.02	0.32	0.17	0.30
Diversity	4.66	7.62	12.99	16.49	18.25	16.75	25.88	16.68	11.90	2.83	6.13	1.44
Adjusted r^2	0.00	0.00	0.01	0.02	0.05	0.12	0.28	0.38	0.44	0.56	0.67	0.51
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

B

RAL

RG map	26.17	35.32	47.13	56.43	56.31	43.22	24.41	13.19	19.89	5.76	5.53	3.08
Quality	6.41	13.74	15.17	6.56	8.49	3.74	4.17	2.18	0.66	3.51	0.28	1.30
Exons	0.87	0.61	1.34	1.43	5.53	4.38	3.56	1.79	0.77	2.25	1.12	0.10
GC	3.79	1.38	1.17	1.35	2.09	2.58	0.70	0.75	0.47	0.64	0.32	0.53
Divergence	0.14	0.01	0.05	2.45	1.08	0.77	0.48	1.46	0.75	0.10	0.12	0.22
Diversity	4.36	2.45	1.60	3.12	4.43	2.63	3.43	2.35	1.44	1.37	0.23	1.41
Adjusted r^2	0.00	0.01	0.04	0.08	0.17	0.26	0.39	0.41	0.66	0.71	0.72	0.89
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

RG

RAL map	25.95	35.44	46.67	55.63	56.75	40.87	21.47	7.92	16.56	4.34	3.31	3.12
Quality	0.96	3.25	0.46	0.29	0.38	0.72	1.77	0.11	0.26	0.42	0.27	0.71
Exons	0.19	2.08	2.19	0.75	0.00	1.56	0.14	0.26	0.39	0.07	0.43	0.04
GC	1.78	3.14	3.22	0.07	0.24	0.57	0.48	1.69	0.33	0.14	0.60	0.11
Divergence	0.41	1.26	0.20	1.62	0.25	0.28	0.04	4.04	1.11	0.47	0.13	0.83
Diversity	4.32	7.03	11.83	13.79	12.66	9.62	13.96	7.99	4.18	1.48	2.74	0.14
Adjusted r^2	0.00	0.01	0.03	0.08	0.16	0.26	0.40	0.45	0.69	0.66	0.79	0.84
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

Figure S13: **Linear model for wavelet transform of recombination map of chromosome arm 2R.** (A) In a linear model for the detail coefficients of the wavelet transform of the recombination map of chromosome arm 2R, covariates are the detail coefficients of wavelet transforms of data quality, gene content, GC content, divergence, and diversity. Shown is the $-\log_{10} p$-value of the regression coefficient at the given scale, as determined by a t-test. Colored boxes indicate significant relationships, with red positive and blue negative. Also shown in the adjusted r^2. (B) As above, but with the recombination map of the other population as an additional covariate.

A

RAL

Quality	5.37	12.20	14.07	20.63	2.92	12.72	6.05	7.51	0.83	1.99	1.25	0.38
Exons	0.43	0.78	2.09	1.53	2.01	7.33	6.95	2.57	3.70	2.20	0.66	0.99
GC	5.97	6.33	5.22	1.35	0.08	0.68	0.19	0.07	0.11	1.21	0.78	1.19
Divergence	0.54	2.58	0.35	0.32	2.31	0.17	0.29	1.00	1.15	0.44	0.16	0.72
Diversity	4.43	4.38	4.75	7.53	9.94	11.01	10.18	9.91	5.50	4.15	4.29	5.65
Adjusted r^2	0.00	0.00	0.01	0.03	0.04	0.15	0.20	0.25	0.43	0.51	0.67	0.88
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

RG

Quality	5.10	14.22	4.30	2.43	0.04	1.09	6.09	2.31	3.42	0.64	0.61	1.39
Exons	0.68	1.03	0.66	1.09	1.96	5.03	4.20	1.71	0.95	0.21	0.08	1.44
GC	2.55	6.96	1.92	2.04	0.35	0.17	0.51	0.06	0.63	0.48	1.17	2.18
Divergence	0.02	1.23	0.51	1.68	0.38	1.21	0.02	1.04	0.17	0.08	1.38	0.22
Diversity	1.32	2.15	0.98	7.62	11.60	16.57	21.79	17.20	15.68	6.32	9.02	7.14
Adjusted r^2	0.00	0.01	0.00	0.01	0.03	0.11	0.23	0.32	0.51	0.41	0.79	0.92
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

B

RAL

RG map	18.62	48.20	38.39	50.24	44.70	18.86	10.64	11.14	2.40	1.23	3.20	0.96
Quality	4.78	9.81	12.15	15.93	1.19	9.82	5.68	5.86	0.63	1.57	0.83	0.22
Exons	0.49	0.69	1.96	1.19	1.26	5.20	4.60	1.39	3.29	1.78	0.67	0.31
GC	5.57	4.87	4.41	0.65	0.03	0.37	0.08	0.35	0.07	1.12	0.71	0.58
Divergence	0.54	2.53	0.31	0.09	3.10	0.03	0.15	0.68	0.98	0.49	0.59	0.29
Diversity	4.04	4.04	4.19	5.13	5.46	5.89	6.51	4.43	3.62	3.16	1.01	1.69
Adjusted r^2	0.00	0.02	0.03	0.08	0.13	0.21	0.26	0.37	0.46	0.53	0.78	0.90
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

RG

RAL map	18.81	47.84	38.36	52.65	45.56	22.02	10.01	11.77	1.12	0.91	2.29	0.68
Quality	4.64	11.65	2.28	0.51	0.12	3.09	7.53	3.66	3.53	0.72	1.06	0.68
Exons	0.74	0.86	0.36	0.66	1.16	2.21	1.98	0.44	0.41	0.01	0.13	1.24
GC	2.28	6.23	1.53	2.27	0.43	0.55	0.57	0.34	0.65	0.73	0.70	1.33
Divergence	0.04	0.84	0.34	1.40	0.95	0.97	0.03	0.65	0.12	0.01	1.58	0.15
Diversity	1.22	1.84	0.79	6.78	8.08	14.22	17.11	12.63	12.85	4.94	4.50	2.29
Adjusted r^2	0.00	0.02	0.02	0.07	0.12	0.19	0.29	0.44	0.52	0.42	0.84	0.93
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

Figure S14: **Linear model for wavelet transform of recombination map of chromosome arm 3L.** (A) In a linear model for the detail coefficients of the wavelet transform of the recombination map of chromosome arm 3L, covariates are the detail coefficients of wavelet transforms of data quality, gene content, GC content, divergence, and diversity. Shown is the $-\log_{10} p$-value of the regression coefficient at the given scale, as determined by a t-test. Colored boxes indicate significant relationships, with red positive and blue negative. Also shown in the adjusted r^2. (B) As above, but with the recombination map of the other population as an additional covariate.

A

RAL

Quality	9.15	14.12	11.51	12.13	10.37	6.99	8.05	4.26	2.38	0.94	0.05	1.62
Exons	0.43	0.27	1.90	2.09	1.55	5.62	2.18	7.94	3.02	2.05	0.77	1.07
GC	2.98	2.98	1.60	0.70	0.03	0.96	0.77	0.15	0.25	0.27	0.22	1.69
Divergence	0.20	1.30	0.19	0.12	0.10	0.21	0.31	0.50	1.58	0.05	0.03	0.90
Diversity	0.80	8.81	10.73	18.82	27.24	22.71	14.83	5.81	5.20	2.82	0.97	1.43
Adjusted r^2	0.00	0.00	0.01	0.03	0.09	0.16	0.21	0.30	0.41	0.40	0.28	0.30
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

RG

Quality	9.14	13.46	15.50	4.82	2.54	1.17	0.31	0.69	0.63	0.97	0.24	1.03
Exons	0.91	1.49	2.89	1.74	2.21	4.51	4.51	6.63	1.60	2.60	0.33	0.01
GC	0.83	1.20	0.97	0.78	0.40	0.45	0.29	0.56	0.88	0.76	0.04	1.95
Divergence	0.23	0.10	0.36	0.38	0.25	0.37	0.41	0.22	0.49	0.03	0.33	0.45
Diversity	8.78	9.10	13.83	17.23	28.09	20.90	18.50	13.66	8.64	7.76	2.90	3.86
Adjusted r^2	0.00	0.01	0.02	0.04	0.10	0.17	0.26	0.43	0.37	0.58	0.33	0.74
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

B

RAL

RG map	29.79	58.35	45.48	34.45	46.73	19.13	20.15	7.80	3.20	3.26	1.88	2.45
Quality	8.38	11.65	7.59	9.29	6.03	3.49	5.18	3.95	1.89	0.57	0.19	1.89
Exons	0.39	0.17	1.31	1.59	0.70	2.93	0.10	2.93	2.16	1.34	0.94	0.82
GC	2.82	2.56	1.21	0.51	0.07	0.93	1.49	0.35	0.21	0.01	0.32	1.32
Divergence	0.20	1.50	0.16	0.33	0.24	0.18	0.27	1.04	1.42	0.18	0.16	1.84
Diversity	0.62	7.46	9.44	14.87	17.74	17.88	11.07	2.81	3.70	1.04	0.30	0.07
Adjusted r^2	0.01	0.02	0.04	0.07	0.18	0.23	0.33	0.39	0.46	0.50	0.41	0.68
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

RG

RAL map	29.31	57.37	44.51	33.51	43.57	17.01	14.23	6.82	1.69	1.97	1.82	0.56
Quality	8.19	10.13	12.58	3.36	1.65	0.79	0.51	1.11	0.60	0.61	0.36	0.38
Exons	0.85	1.37	2.28	1.27	1.36	2.58	3.07	3.26	0.49	1.17	0.91	0.03
GC	0.70	1.14	0.82	0.85	0.43	0.58	0.26	0.36	0.39	0.45	0.17	0.61
Divergence	0.20	0.25	0.39	0.45	0.18	0.34	0.27	0.40	0.29	0.06	0.27	0.62
Diversity	8.47	8.34	12.01	13.28	16.81	14.23	9.86	10.33	5.87	4.80	2.44	2.29
Adjusted r^2	0.01	0.02	0.04	0.07	0.19	0.23	0.34	0.49	0.39	0.62	0.45	0.75
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

Figure S15: **Linear model for wavelet transform of recombination map of chromosome arm 3R.** (A) In a linear model for the detail coefficients of the wavelet transform of the recombination map of chromosome arm 3R, covariates are the detail coefficients of wavelet transforms of data quality, gene content, GC content, divergence, and diversity. Shown is the $-\log_{10} p$-value of the regression coefficient at the given scale, as determined by a t-test. Colored boxes indicate significant relationships, with red positive and blue negative. Also shown in the adjusted r^2. (B) As above, but with the recombination map of the other population as an additional covariate.

A

RAL

Quality	2.46	4.71	3.64	2.48	0.82	0.62	0.36	0.11	2.11	0.54	0.35	1.57
Exons	1.43	2.11	0.03	1.18	0.31	0.10	2.18	1.95	0.79	2.11	0.53	1.51
GC	1.54	2.02	2.11	0.91	0.71	0.01	1.61	0.68	0.01	0.01	0.47	1.19
Divergence	0.11	0.03	0.05	0.13	0.02	0.01	0.65	0.02	0.26	0.45	0.08	1.01
Diversity	0.55	0.59	2.26	5.30	11.04	13.94	19.30	11.85	6.16	1.93	0.24	3.21
Adjusted r^2	0.00	0.00	0.00	0.01	0.02	0.05	0.18	0.23	0.25	0.28	0.03	0.57
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

RG

Quality	1.68	0.81	0.36	1.41	4.81	8.64	10.49	3.12	1.48	0.12	0.15	1.15
Exons	0.48	0.90	0.06	0.53	0.16	0.11	0.11	2.35	0.62	2.73	0.67	0.01
GC	0.32	2.13	1.30	1.18	1.67	0.65	0.12	0.36	0.11	0.63	0.95	1.70
Divergence	0.04	0.14	0.61	0.97	1.24	0.08	0.02	0.31	0.94	0.45	0.13	0.23
Diversity	3.33	5.88	3.24	5.29	15.37	26.97	24.56	17.46	12.69	4.25	2.75	3.89
Adjusted r^2	0.00	0.00	0.00	0.01	0.03	0.11	0.20	0.37	0.43	0.43	0.51	0.75
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

B

RAL

RG map	1.75	3.74	11.81	13.72	25.39	17.48	13.42	8.96	6.38	2.70	2.45	0.82
Quality	2.39	4.36	3.13	2.23	0.81	0.36	0.20	0.25	1.02	0.22	0.24	0.40
Exons	1.44	2.17	0.02	1.32	0.24	0.07	1.49	0.53	0.38	1.28	0.01	1.26
GC	1.53	1.93	1.90	0.77	0.34	0.26	1.58	0.70	0.31	0.34	0.21	0.84
Divergence	0.11	0.03	0.02	0.06	0.16	0.02	0.81	0.00	0.41	0.73	0.64	0.56
Diversity	0.52	0.53	1.98	4.71	8.50	8.98	12.52	4.94	3.23	0.91	0.07	1.44
Adjusted r^2	0.00	0.00	0.01	0.02	0.07	0.12	0.27	0.33	0.39	0.38	0.24	0.62
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

RG

RAL map	1.79	4.04	12.42	14.08	25.45	17.84	14.48	10.32	5.67	3.11	3.02	0.49
Quality	1.64	0.74	0.22	1.71	5.11	8.45	10.35	2.62	1.60	0.03	0.01	0.66
Exons	0.50	0.96	0.04	0.74	0.02	0.17	0.24	1.58	0.37	1.24	0.36	0.18
GC	0.30	2.05	1.12	1.03	1.33	0.58	0.31	0.13	0.07	0.38	0.78	1.44
Divergence	0.04	0.14	0.61	0.97	1.22	0.12	0.11	0.30	1.10	0.80	0.36	0.24
Diversity	3.33	5.91	3.39	4.85	13.05	22.64	18.24	12.28	9.12	3.38	3.33	1.89
Adjusted r^2	0.00	0.00	0.01	0.02	0.09	0.18	0.29	0.47	0.52	0.53	0.67	0.75
Scale (kb)	0.5	1	2	4	8	16	32	64	128	256	512	1024

Figure S16: **Linear model for wavelet transform of recombination map of chromosome X.** (A) In a linear model for the detail coefficients of the wavelet transform of the recombination map of chromosome arm X, covariates are the detail coefficients of wavelet transforms of data quality, gene content, GC content, divergence, and diversity. Shown is the $-\log_{10} p$-value of the regression coefficient at the given scale, as determined by a t-test. Colored boxes indicate significant relationships, with red positive and blue negative. Also shown in the adjusted r^2. (B) As above, but with the recombination map of the other population as an additional covariate.

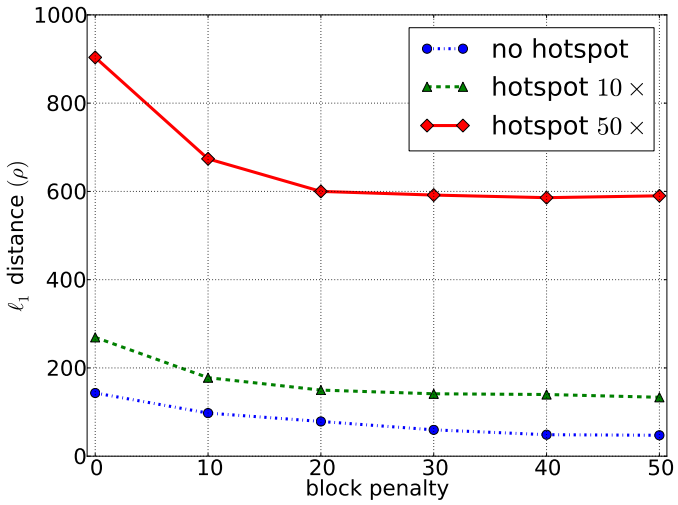

Figure S17: **Plot of the average ℓ_1-distance between the true and estimated recombination maps.** Each plot shows the results averaged over 100 simulated datasets per block penalty for a given recombination landscape. In each simulation, we considered a 25 kb region with the background recombination rate of $\rho = 10$/kb. "no hotspot": The true recombination map is constant. "hotspot 10×": In the middle of the 25 kb region, the true recombination map has a hotspot of width 2 kb and intensity 10× the background rate. "hotspot 50×": In the middle of the 25 kb region, the true recombination map has a hotspot of width 2 kb and intensity 50× the background rate.

Supporting Tables

Table S1: **Summary of comparison between `LDhelmet` and `LDhat` in the neutral case.** Based on 100 simulated datasets for a 25 kb region. "No Hotspot" corresponds to the case of a constant recombination map, whereas "Hotspot 10×" corresponds to the case with a 2 kb wide hotspot situated at the center of the region. The first row shows the regional average of ρ obtained by `LDhelmet` and `LDhat`, averaged over the 100 datasets. The second row shows the total rate in the hotspot region, averaged over the datasets. The third row shows the percentage of datasets for which the estimate had at least one false peak with height ≥ 5 times the background rate. The fourth row shows the percentage of datasets for which the estimate had at least one false peak with height ≥ 10 times the background rate. The fifth row shows the percentage absolute error of the estimated ρ average outside the hotspot region from the true ρ average outside the hotspot region. The true ρ average outside the hotspot region is $\rho = 0.01$/bp. To account for edge effects, 2.5 kb from each end of the map were removed prior to computing the statistics.

	No Hotspot			Hotspot 10×		
Measure of Accuracy	True Value	LDhelmet	LDhat	True Value	LDhelmet	LDhat
ρ average (per bp)	0.01	0.0097	0.0109	0.0172	0.0184	0.0203
Total hotspot area	20.0	19.0	20.3	200.0	195.2	210.0
% with false peak $\geq 5\times$		5%	30%		4%	30%
% with false peak $\geq 10\times$		2%	21%		4%	21%
% abs. error outside hotspot region		14%	23%		15%	20%

Table S2: **SNP densities (per kb) of neutral and single-sweep simulations.** The mean, minimum, maximum and standard deviation of the SNP density for the datasets used in Tables S1 and S3. The simulations assumed a finite-sites, quadra-allelic mutation model, with mutation matrix $\boldsymbol{P}_{\mathrm{RAL}}$ and $\theta = 0.008$, which is the effective population-scaled mutation rate adjusted for $\boldsymbol{P}_{\mathrm{RAL}}$ (see Estimation of mutation transition matrices).

	Neutral		Single-Sweep Model	
	No Hotspot	Hotspot 10×	No Hotspot	Hotspot 10×
Mean	21.82	21.68	18.15	18.38
Min	18.32	17.40	14.84	14.68
Max	26.12	25.72	24.08	22.76
Std dev	1.71	1.38	1.64	0.61

Table S3: **Summary of comparison between `LDhelmet` and `LDhat` in the case of single selective sweep.** Based on 100 simulated datasets for a 25 kb region. For each dataset, a selected site was placed at position 5 kb and the population-scaled selection coefficient was set to 1000. The fixation time of the selected site was 0.01 coalescent units in the past. The column and the row labels are the same as in Table S1. As for Table S1, 2.5 kb from each end of the map were removed prior to computing the statistics to account for edge effects.

	No Hotspot			Hotspot 10×		
Measure of Accuracy	True Value	LDhelmet	LDhat	True Value	LDhelmet	LDhat
ρ average (per bp)	0.01	0.0079	0.0108	0.0172	0.0162	0.0220
Total hotspot area	20.0	14.7	15.4	200.0	169.8	224.6
% with false peak $\geq 5\times$		10%	42%		8%	34%
% with false peak $\geq 10\times$		6%	39%		5%	24%
% abs. error outside hotspot region		39%	58%		30%	56%

Table S4: **SNP densities (per kb) of recurrent-sweep and demography simulations.** The statistics for each selection or demography scenario are merged over the three recombination landscapes (i.e., no hotspot, hotspot 10× and hotspot 50×). The simulations use θ_{RAL} and $\boldsymbol{P}_{\mathrm{RAL}}$ as parameters. The third column shows the SNP density per kb across the hundred datasets, and the fourth column shows the standard deviation. For the definitions of the scenario names, refer to Simulation study on the impact of natural selection and Simulation study on the impact of demographic history of the main text. "Control" refers to a control dataset with constant population size and no selection.

Simulation Type	Model	Mean	Std dev
Recurrent Sweeps	RS1	18.22	1.66
	RS2	4.10	1.05
	RS3	2.71	1.24
Demography	G1	12.86	1.07
	G2	15.85	1.24
	B1	13.84	2.78
	B2	5.53	2.14
Neutral	Control	22.51	1.49

Table S5: **SNP densities (per kb) of the real *Drosophila* data.**

Chromosome Arm	RAL	RG
2L	24.54	25.49
2R	22.56	24.21
3L	22.29	25.20
3R	19.77	20.79
X	14.92	28.15

Table S6: **Subsampling of real data.** To assess the effect of subsampling individuals, we subsampled a 2 Mb excerpt from chromosome arm 2L for both the RAL and RG datasets. We performed subsampling four times, and each row is the average of the four subsampled datasets. The column labeled n is the number of individuals in each subsample. The percentiles are given in the three rightmost columns. The results show that sample size has a slight positive bias, but does not impact estimates greatly.

		Percentile (ρ per kb)		
	n	2.5%	50%	97.5%
RAL	17	6.1	6.2	6.5
	27	7.2	7.3	7.4
	37	7.8	7.8	7.9
RG	12	8.1	8.4	9.2
	17	9.0	9.0	9.2
	22	9.2	9.3	9.4

Table S7: **Thinned SNPs on RG dataset.** To assess the effect of SNP density on the recombination rate inference, we thinned the SNPs on chromosome arm 2L and chromosome X of RG to the SNP density of RAL. The 2.5%, 50% and 97.5% percentiles are shown for estimates. The number of SNPs in the original dataset and in the thinned dataset are shown in the fourth column. For chromosome arm 2L, the change in SNP density is negligible. For chromosome X, the difference in SNP density is significant. The results show that SNP density impacts the estimate, but not to the extent of the difference observed between RAL and RG on chromosome X.

				Percentile (ρ per kb)		
Dataset	Arm	Type	# SNPs	2.5%	50%	97.5%
RG	2L	Original	586476	33.0	35.9	39.4
		Thinned	564673	32.5	35.5	38.9
	X	Original	631205	110.0	121.4	134.1
		Thinned	334647	97.5	106.8	117.4

Table S8: **Exclusion of individuals with inversions.** To assess the effect of inversions on the recombination rate estimate, we excluded individuals known to carry the given inversion, and performed inference on the remaining sample. 0.5 Mb was added to both ends of the region to eliminate possible edge effects.The ρ average is over the inversion region only. The column labeled **Original** gives the estimate using the entire sample. The column labeled **Excluded** gives the estimate excluding the individuals with the given inversion. The inversion region length and the number of individuals with the inversion are provided in the rightmost two columns.

			Original	**Excluded**	Inversion	# with
Dataset	Arm	Inversion	ρ per kb	ρ per kb	length (Mb)	inversion
RAL	2L	2Lt	16.97	16.45	10.9	3
	2R	2RNS	17.34	16.66	4.9	2
	3R	3RK	11.80	11.39	14.4	1
	3R	3RMO	12.51	14.56	14.6	7
	3R	3RP	12.49	11.35	8.3	1
RG	2L	2Lt	54.44	50.80	10.9	2
	2R	2RNS	53.93	50.81	4.9	1
	3R	3RP	22.44	17.24	8.3	4
	X	1Be	106.26	103.21	1.8	3

Table S9: **Running times (in seconds) for solving recursions and computing Padé coefficients.** The second column is the time to solve the two-locus recursion described in Text S1 to compute the likelihood of a *single* value of ρ for all sample configurations of size n. The third column is the time to compute 11 Padé coefficients for all sample configurations of size n. Recall that the recursion must be solved afresh for every value of ρ in the lookup table. On the other hand, the Padé coefficients are used to construct a rational function of ρ that approximates the likelihood; once the Padé coefficients are determined, evaluating the likelihood is instantaneous. A single 2.5 Ghz core was used in this benchmarking to provide representative estimates of the running time. However, note that both the recursion and Padé coefficient computations are highly parallelizable, which we exploit in the implementation of `LDhelmet`. Also note that the presence of missing data does not increase the running time for either computation.

Sample size n	Two-locus recursion (seconds)	Padé coefficients (seconds)
10	0.1	5
20	11	429
30	189	5271
40	1523	26405
50	7755	75704