Quantitative modeling of fine-scale variations in the *Arabidopsis thaliana* crossover landscape

Yu-Ming Hsu, Matthieu Falque and Olivier C. Martin

Supplementary Material

epigenomic feature	Sample accession or series accession number	tissue	reference
H3K4me1	GSM3674621	leaves	Lu et al., 2019; Crisp et al., 2020
	GSM4668649	seedlings	Niu et al., 2021
	GSM4609829	root non- hair cells	missing
	GSM4785549	inflorescence	Liu et al., 2021
	E-MTAB-7370	unopened flower buds	Lambing et al., 2020
H3K4me3	GSM3674620	leaves	Lu et al., 2019; Crisp et al., 2020
	GSM4154769	seedlings	Liu et al., 2020
	GSM2210857	roots	Yen et al., 2017
	GSM4785552	inflorescence	Liu et al., 2021
	GSE120664	sperm nuclei	Borg et al., 2020
H3K9me2	GSM4734580	leaves	Wang et al., 2021
	GSM3040062	10-day seedlings	Ma et al., 2018
	GSM4422529	mature embryos	Parent et al., 2021
	GSM4818168	flowers	Feng et al., 2020

	E-MTAB-7370	unopened flower buds	Lambing et al., 2020
H3K27me3	GSM3674617	leaves	Lu et al., 2019; Crisp et al., 2020
	GSM3617717	seedlings	Shu et al., 2021
	GSM2210865	roots	Yen et al., 2017
	GSM4785573	inflorescences	Liu et al., 2021
	GSE120664	sperm nuclei	Borg et al., 2020
ATAC	GSM3674715	leaves	Lu et al., 2019; Crisp et al., 2020
	GSM2719200	stem cells	Sijacic et al., 2018
	GSM2719204	mesophyll cells	
	GSM3498708	flowers	Potok et al., 2019
	GSE155344	microspores	Borg et al., 2021
DNase	GSM1289358	seedlings	Sullivan et al., 2014; Sullivan et al., 2019
	GSM1289374	whole roots	
	GSM1289378	seed coats	
	GSM1289380	open flowers	
	GSM1289381	unopened flower	

Supplementary Table S1. Origin and description of datasets for the 6 epigenomic features used in this study.

	intercept (a_0)	gene (a_1)	TE (a_2)	TSS (a_3)	H3K4me1 (a_4)	H3K4me3 (a_5)	H3K9me2 (a_6)	H3K27me3 (a_7)	ATAC (a_8)	DNase (a_9)	R^2
50kb	1.56**	-3.6***	-1.67**	0.17	-0.04	0.05	-0.004***	0.11***	0.65***	0.006*	0.28
100kb	1.00	-5.02***	-1.14	0.26	-0.07	0.16*	-0.01***	0.14**	0.71***	-0.005	0.36
200kb	0.06	-4.44*	0.23	0.3	-0.09	0.16	-0.01**	0.14	0.75***	-0.0007	0.42
500kb	-1.01	-5.82	1.08	0.27	-0.08	0.24	-0.01	0.16	0.83***	0.003	0.50

Supplementary Table S2. Adjusted parameters and R^2 values for the additive model when using different bin sizes. The 9 successive features are those in Fig. 1 (ordered left to right and top to bottom). Parameter values were obtained using the lm() function in R. *, ** and *** correspond to parameters having *p*-values less than 0.05, 0.01 and 0.001 respectively for the hypothesis that the true value of the parameter vanishes. The first column gives the bin size used for each fit. Note that the statistical noise intrinsic to CO formation inevitably drives R^2 (last column, *cf*. Eq. 2 in Main) downward as bin size decreases.

bin size (kb)	AIC	BIC	R^2	Model considered
50	247310	247367.8	0.33	10 states
50	247214.7	247289.8	0.34	10 states + IR
50	246531.8	246618.4	0.39	10_states + IR + SNP
50	246459.3	246545.9	0.4	10_states + IR + SNP + rescaling
100	224047.6	224098.4	0.41	10 states
100	223974	224040	0.43	10 states + IR
100	223515.7	223592	0.48	10_states + IR + SNP
100	223444.3	223520.6	0.49	10_states + IR + SNP + rescaling
200	201007.7	201051.7	0.49	10 states
200	200953	201010.2	0.5	10 states + IR
200	200670.5	200736.4	0.54	10_states + IR + SNP
200	200590.1	200656	0.56	10_states + IR + SNP + rescaling
500	170023	170057.8	0.58	10 states
500	170017.7	170062.9	0.59	10 states + IR
500	169754	169806.2	0.64	10_states + IR + SNP
500	169681	169733.2	0.66	10_states + IR + SNP + rescaling

Supplementary Table S3. Model selection *via* AIC and BIC values. For each of the different bin sizes, we consider the sequence of models of increasing complexity, starting with the 10 parameters for the 10 states, adding to that the 3 parameters for the IR size effect, adding to that the 2 parameters for the SNP effect, and finally adding the rescaling (no additional parameters). The AIC and BIC approaches penalize the goodness of fit measure by an amount that depends on the number of parameters. Using a more complex model (with more parameters) is only justified if the associated criterion (AIC or BIC) is lower. The table shows that the data drives one to use the full model having 15 parameters and scaling.

name	50kb	100kb	200kb	500kb
r_state1	1.367	1.199	1.663	0.984
r_state2	1.908	1.998	2.457	1.965
r_state3	5.43E-09	5.95E-09	5.52E-09	4.95E-09
r_state4	1.822	1.832	2.54	1.926
r_state5	0.713	0.804	1.397	0.809
r_state6	0.328	5.95E-09	5.52E-09	4.95E-09
r_state7	5.43E-09	5.95E-09	5.52E-09	4.95E-09
r_state8	1.325	1.538	2.481	1.782
r_state9	0.007	0.002	5.52E-09	4.95E-09
r_SV	0.009	0.008	0.007	0.001
α_1	1.087	0.948	0.774	1.008
α_2	0.087	0.085	0.087	0.082
β_1	0.513	0.452	0.542	0.487
β_2	7.218	7.63	12.743	2.708
β_3	2.998	3.068	2.245	1.554
R^2	0.403	0.488	0.563	0.657

Supplementary Table S4. Parameter values after calibration of the quantitative model having 15 parameters when using bin sizes from 50 to 500 kb. In the column “name”, r_state1 to r_SV refer to the “base recombination rate” for each of the 10 chromatin states, α_1 and α_2 (respectively β_1, β_2. β_3) refer to the parameters in the SNP (respectively intergenic-region size) modulation effect, and finally R^2 refers to the fraction of the variance explained by the model (*cf*. Eq. 2 in Main).

	Chr1 (fit)	Chr2 (fit)	Chr3 (fit)	Chr4 (fit)	Chr5 (fit)
Chr1 (predict)	0.463	0.299	0.347	0.297	0.438
Chr2 (predict)	0.403	0.502	0.448	0.48	0.434
Chr3 (predict)	0.523	0.556	0.607	0.534	0.56
Chr4 (predict)	0.426	0.472	0.473	0.54	0.466
Chr5 (predict)	0.453	0.376	0.41	0.374	0.473

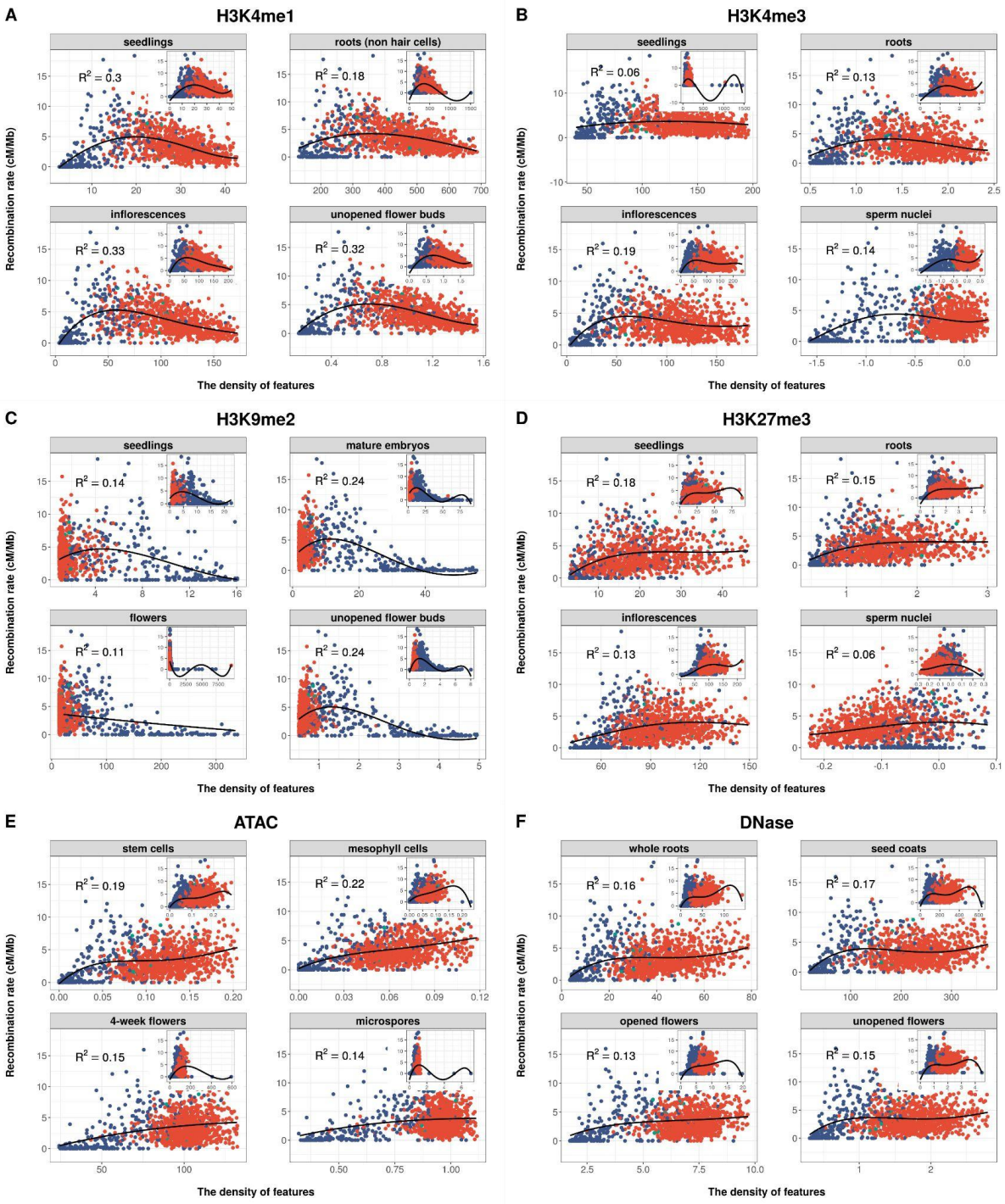
Supplementary Table S5. Predictive power of the model with 15 parameters. We provide the R^2 values when using one chromosome (that labeled by the considered column) to fit the 15 parameters and then apply that calibrated model to predict recombination landscapes of all 5 chromosomes. The genome has been segmented into bins of size 100 kb. Note that in each row the largest R^2 value must occur for the chromosome that has been used to do the fitting of parameters. Omitting the R^2 values produced by the calibrations (on the diagonal), the average R^2 of the predictions (remaining 20 values) is 0.427.

	Chr1 (fit)	Chr2 (fit)	Chr3 (fit)	Chr4 (fit)	Chr5 (fit)
Chr1 (predict)	0.348	0.222	0.22	0.171	0.292
Chr2 (predict)	0.211	0.409	0.263	0.344	0.339
Chr3 (predict)	0.138	0.35	0.455	0.353	0.383
Chr4 (predict)	0.218	0.347	0.34	0.383	0.328
Chr5 (predict)	0.281	0.218	0.274	0.215	0.346

Supplementary Table S6. Predictive power of the additive model (Eq. 1) with 10 parameters exploiting the genomic and epigenomic features of Fig. 1. We provide the R^2 values when using one chromosome (that labeled by the considered column) to fit the 10 parameters and then apply that calibrated model to predict recombination landscapes of all 5 chromosomes (same procedure as in Supplementary Table S5, again with bins of size 100 kb). Omitting the R^2 values produced by the calibrations (on the diagonal), the average R^2 of the predictions (remaining 20 values) is 0.275.

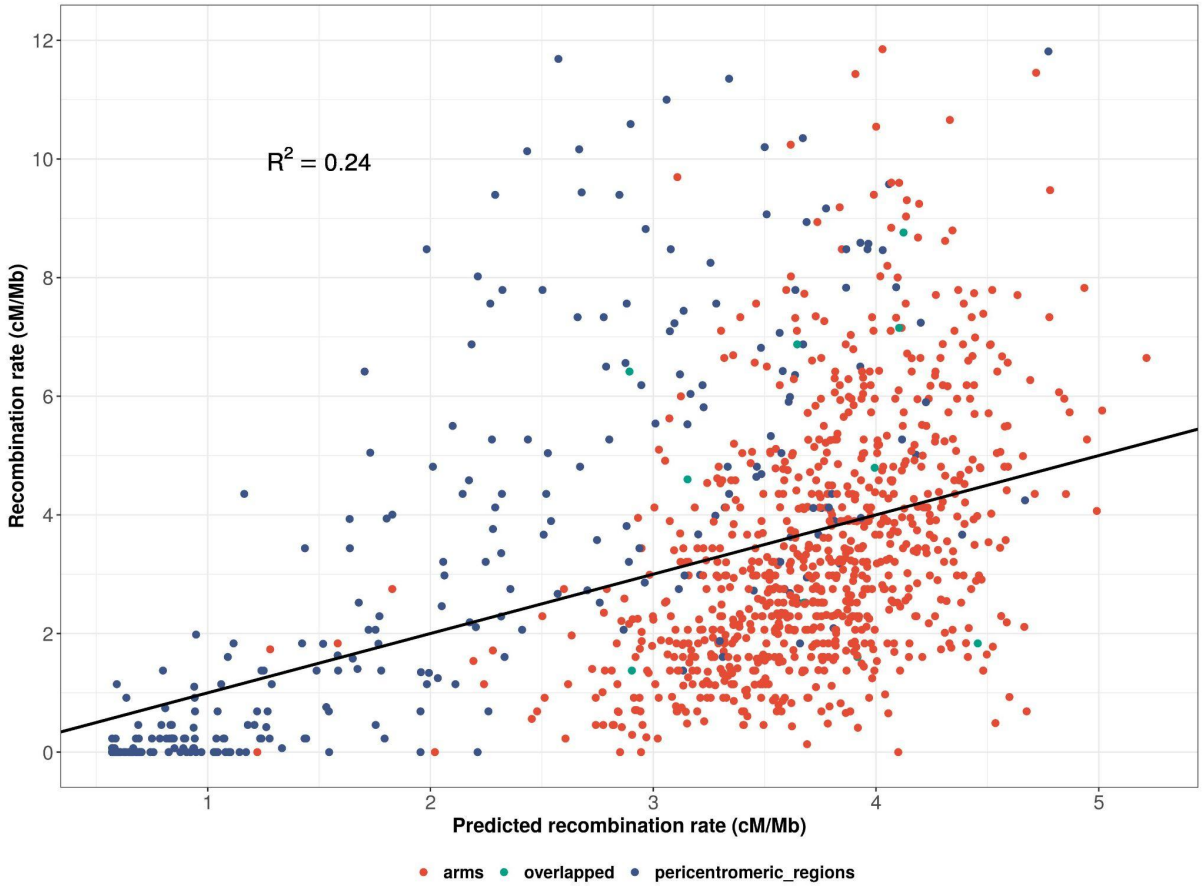
	Chr1_fit	Chr2_fit	Chr3_fit	Chr4_fit	Chr5_fit
Chr1_predict	0.447	-1.364	-0.493	-0.407	0.176
Chr2_predict	-0.299	0.579	-0.614	-39.286	-8.073
Chr3_predict	-0.307	-78.667	0.568	-39.829	-2.22
Chr4_predict	0.074	-17.86	0.001	0.545	-0.349
Chr5_predict	-0.3	-27.968	-1.393	-2.783	0.501

Supplementary Table S7. Predictive power of the model with interactions (Eq. 3) with 46 parameters exploiting the genomic and epigenomic features of Fig. 1. We provide the R^2 values when using one chromosome (that labeled by the considered column) to fit the 46 parameters and then apply that calibrated model to predict recombination landscapes of all 5 chromosomes (same procedure as in Supplementary Table S5, again with bins of size 100 kb). Note that the R^2 of most of the predictions are negative, showing that this model with interactions has no predictive power, presumably because it strongly overfits the data during calibration.

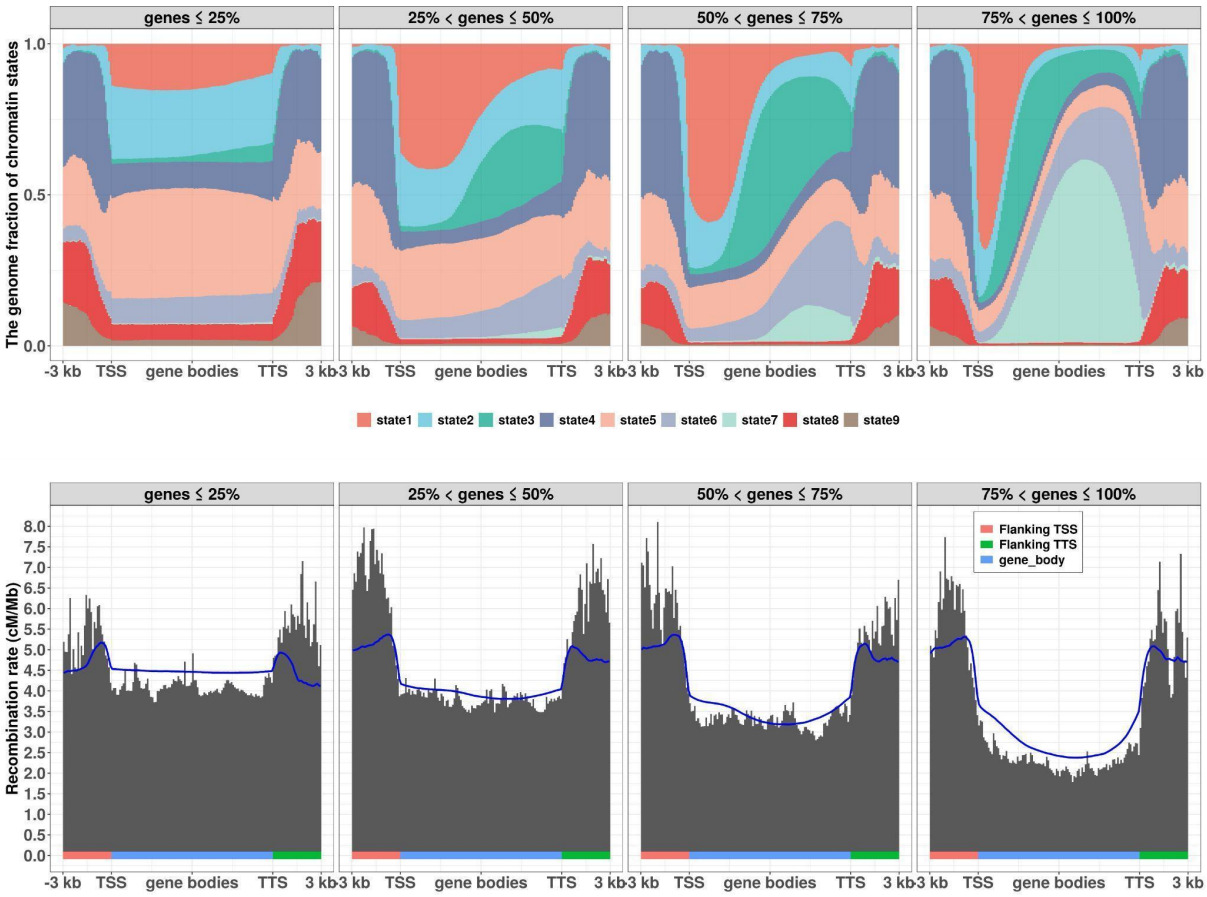


Supplementary Figure S1. The correlations between recombination rate and six epigenomic features when measured in somatic vs. germinal tissues. From (A) to (F), each sub figure combines four plots using data from two somatic and two germinal tissues for the same epigenomic feature. The subtitle on each plot indicates the corresponding tissue. Each dot represents the values for a 100-kb bin. The x-axis values correspond to the density of peaks or reads of each feature according to the format of raw data downloaded from NCBI or ArrayExpress databases. The y-axis gives the associated recombination rate based on a total of 17,077 crossovers from the Col-0-Ler F_2 population. As in Fig. 1 of Main, curves show the fits using a

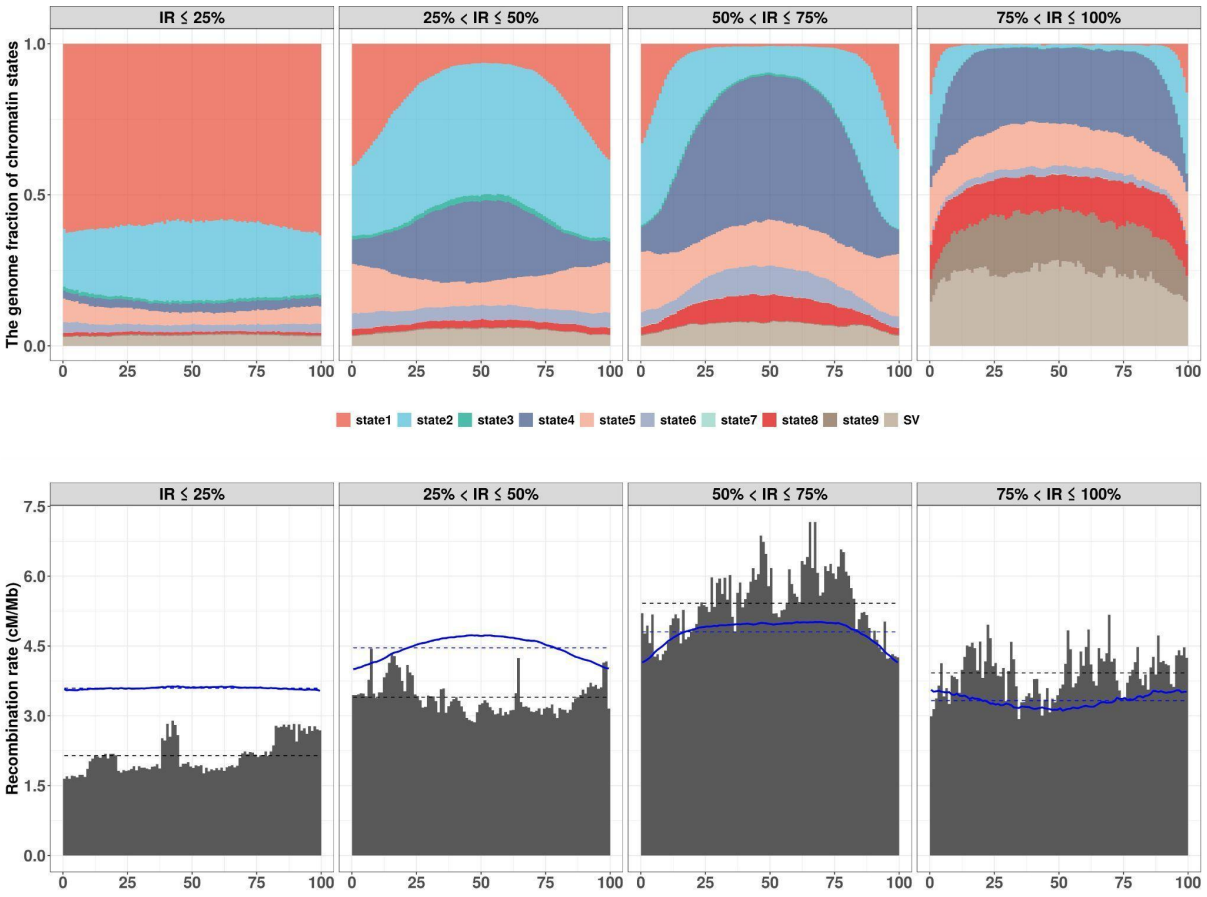
polynomial of degree 4 over the full data range from which the R^2 values are calculated. The main part of each panel corresponds to a zoom of the inset to show greater detail in the main part of the scatter plot.



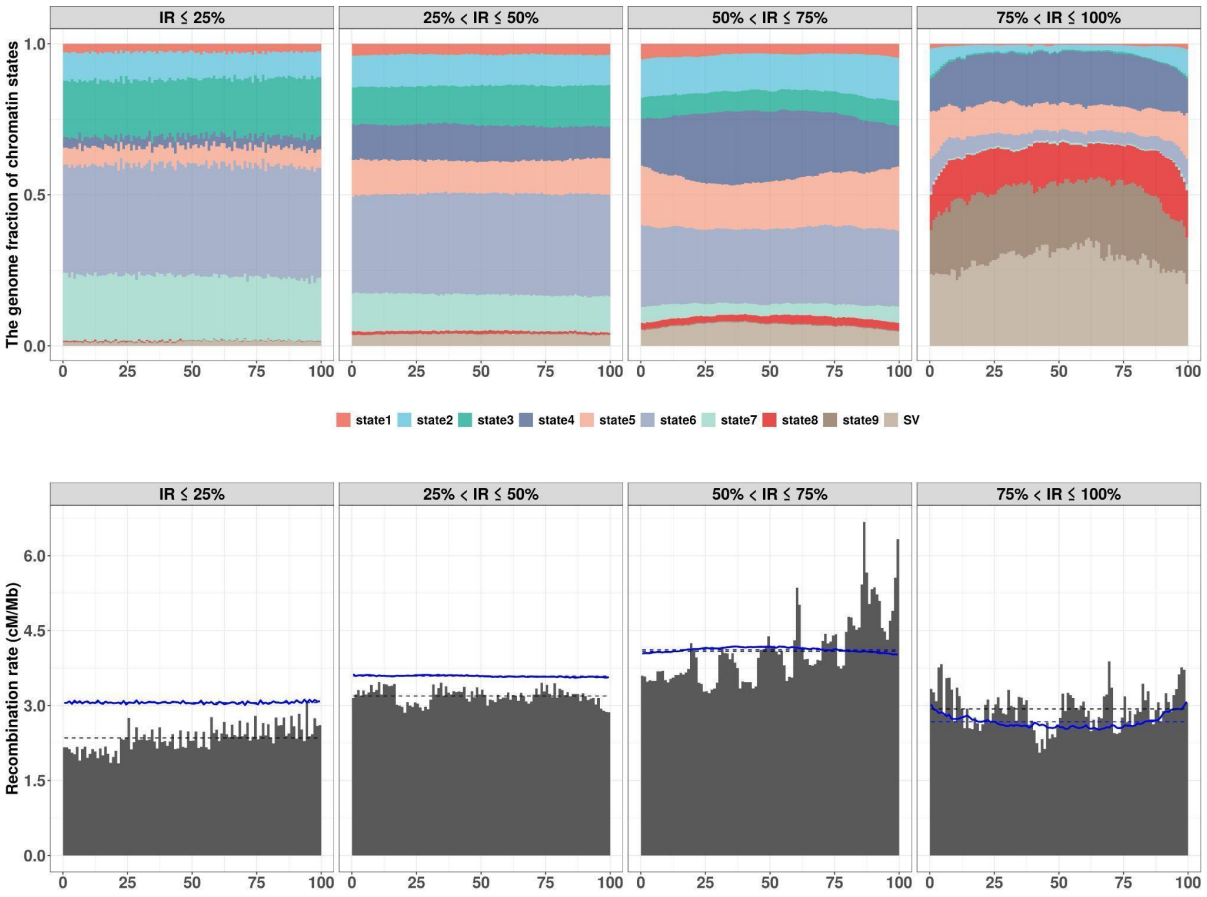
Supplementary Figure S2. Comparison of experimental and predicted recombination rates. Here the predictions are those of the 10 chromatin states model using the experimentally measured state-specific recombination rates (no adjustable parameters). Each data point is associated with a bin of 100 kb along the genome. The fraction of variance explained by the model (computed using the deviations from the predicted recombination rates) is $R^2 = 0.24$.



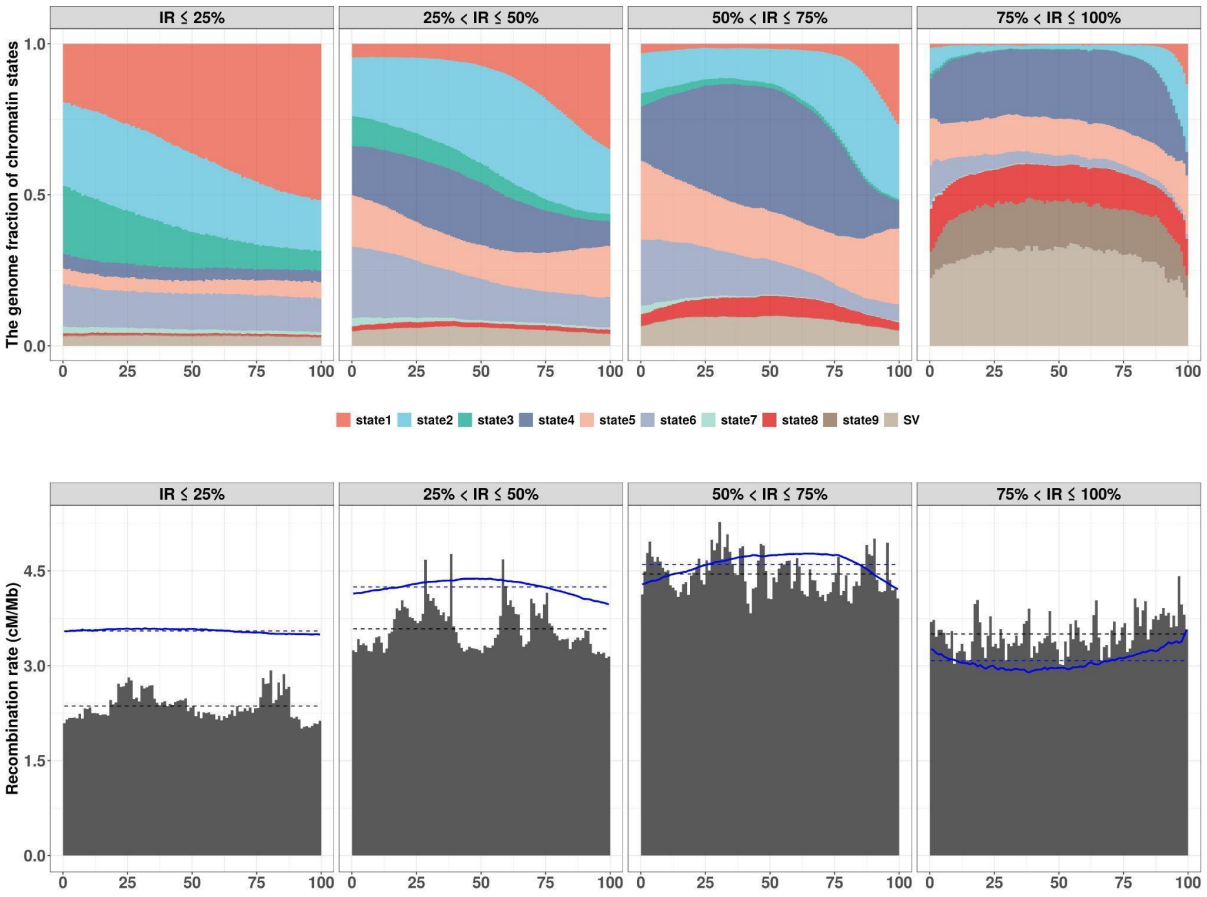
Supplementary Figure S3. Dependence of recombination patterns on gene body size. The profiles of chromatin states and the recombination rate patterns are determined separately in the four quantiles of gene body size. The procedures are the same as in Fig 2B, and the blue curve shows the prediction of the model with 10 chromatin states when using the experimentally measured state-specific recombination rates (no adjustable parameters). The predictions of the model follow the experimental values rather well.



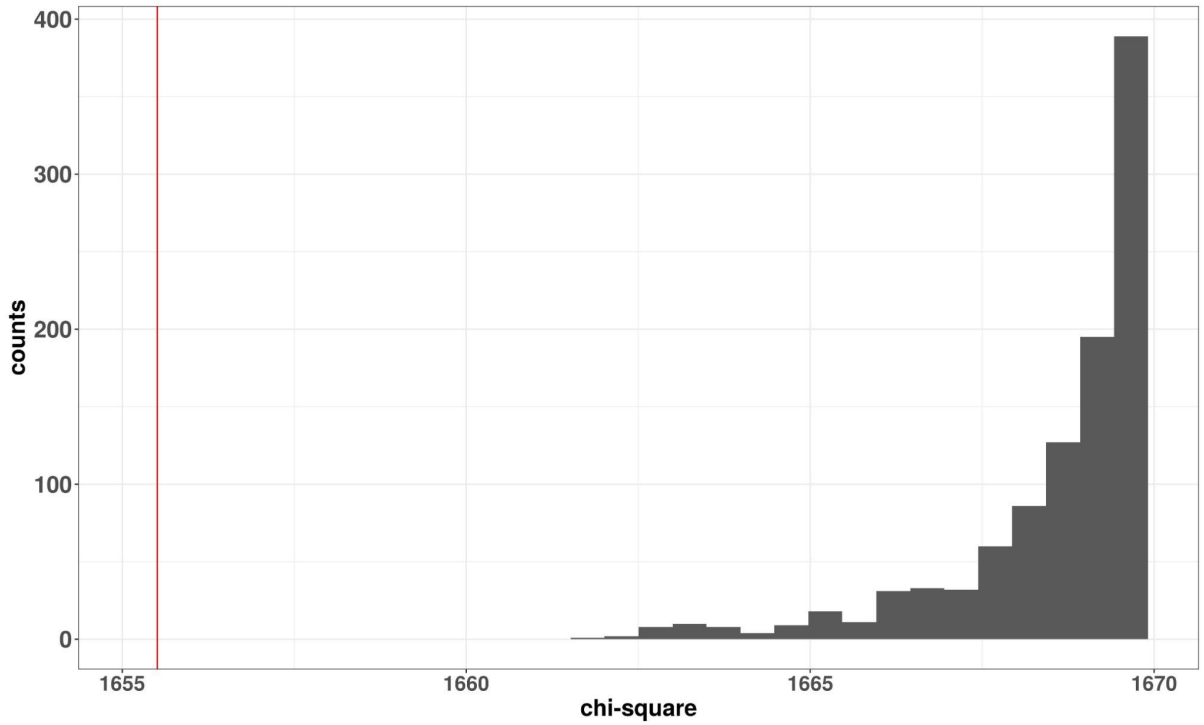
Supplementary Figure S4. The profiles of chromatin states and recombination rate in intergenic regions between genes of divergent orientation. All “divergent” intergenic regions larger than 100 base pairs are divided into 4 groups depending on their size, and each group has one quantile (25 %) of intergenic-region events. In each group, we segmented every intergenic region into 100 bins, then pooled all data of each bin, and calculated the fraction of 9 chromatin states and SVs and the recombination rate of each bin. In the top of this figure we show the fraction of states on the y-axis while the x-axis gives the relative position using 100 bins. At the bottom of this figure, the y-axis corresponds to the recombination rate, while the x-axis is as above. The bottom histograms show the experimental recombination rate in the 100 bins, the black dashed line giving the corresponding average. The procedures are the same as in Fig 2B. The continuous blue curve shows the prediction of the model with 10 chromatin states when using the experimentally measured state-specific recombination rates (no adjustable parameters). The blue dashed line is the corresponding average. The predictions of the model systematically overestimate recombination rates in the small intergenic regions.



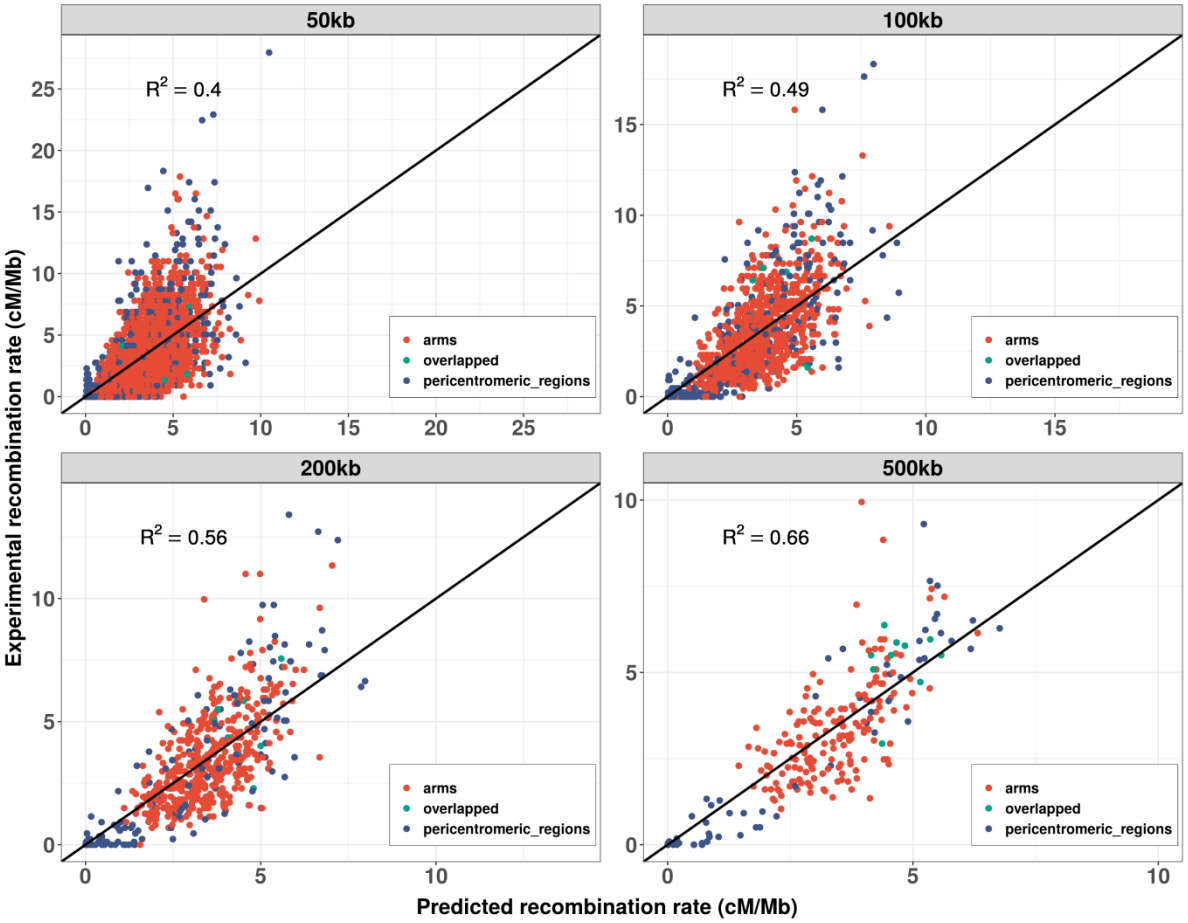
Supplementary Figure S5. The profiles of chromatin states and patterns of recombination rate in intergenic regions between genes of convergent orientation. The procedures and quantities displayed are as in Supplementary Figure S4. The predictions of the model systematically overestimate recombination rates in the small intergenic regions.



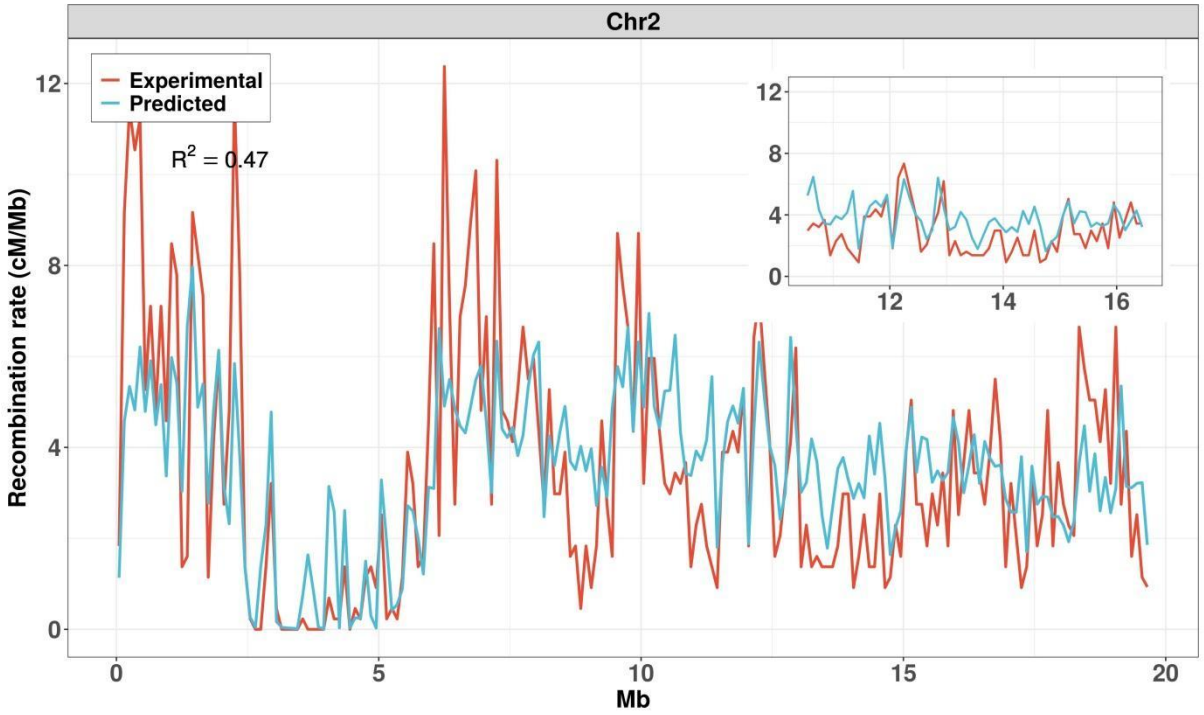
Supplementary Figure S6. The profiles of chromatin states and recombination rate in intergenic regions between genes of parallel orientation. The procedures and quantities displayed are as in Supplementary Figure S4. The predictions of the model systematically overestimate recombination rates in the small intergenic regions.

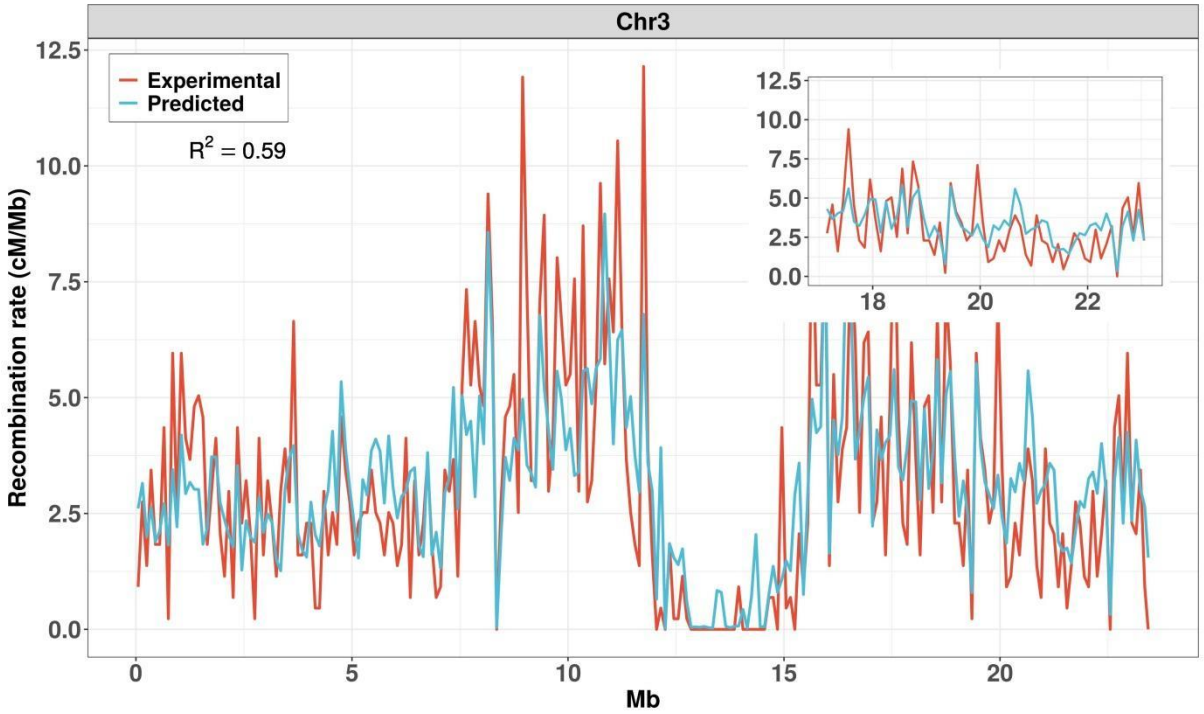


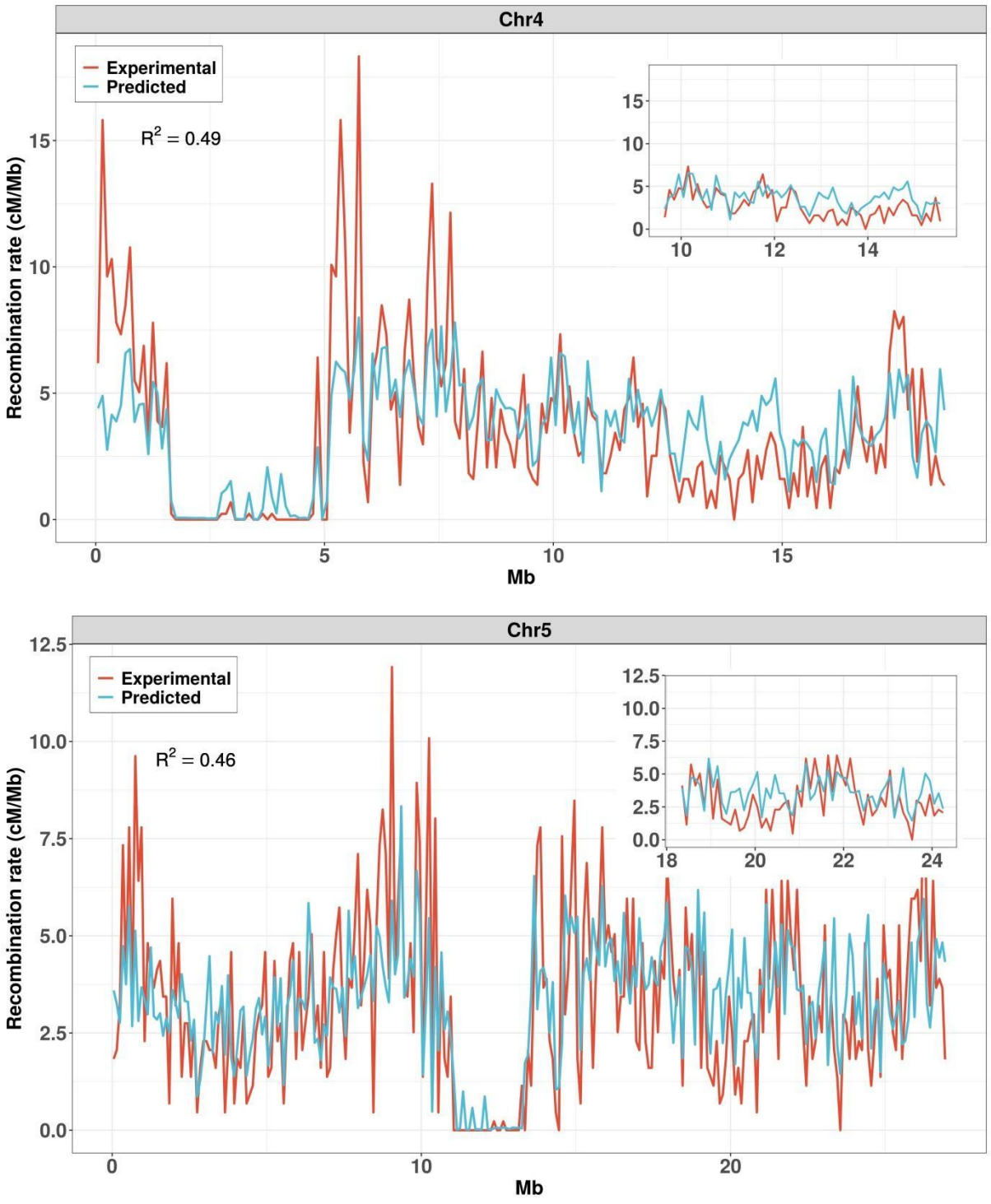
Supplementary Figure S7. Another framework to test whether recombination rate is suppressed by low SNP density. In this approach (different from the one in Main), we compare two hypotheses, H0 and H1. Under H0, we assume that there is an (unknown) “reference” recombination landscape, likely driven by genomic or epigenomic features, but common to all 5 F2 populations of Blackwell et al. (2020). (In Main, this reference landscape was implicitly assumed to be constant.) Under H1, the common landscape is further modulated by the divergence between the homologs present, thus differently in each cross and each bin. This modulation is parametrized via the function (a + b x) exp(- cx) where x is the SNP density of the bin in the considered cross. Because high SNP density is expected to lead to suppressed recombination, the test is only applied to data belonging to the first two quantiles of SNP density. We confront H0 to H1 by asking whether a good fit to the data necessitates the modulation effect. We thus compare the chi-square goodness of fit using H1 to what would be expected if there were no causal suppressive effect (the H0 hypothesis). That distribution is obtained by shuffling in each bin the values of SNP density between crosses to decorrelate recombination rate from any SNP density effect. The figure displays the histogram of the chi-square values under H0 where for each shuffling we have adjusted the parameters a, b, and c to minimize the chi-square for that shuffle. Also, the red line gives the chi-square value in the unshuffled data, corresponding to H1, showing that the recombination rate modulation, when using the SNPs between the parents of each separate cross, improves the fit far more than expected by chance (p-value $\leqslant 0.001$).



Supplementary Figure S8. Scatterplots of experimental and predicted recombination rate when the 15 parameter model calibration is done using bin sizes ranging from 50 to 500 kb. The x-axis specifies the recombination rate predicted by our quantitative model that incorporates 10 chromatin states along with contextual modulating effects, having a total of 15 adjustable parameters. The y-axis corresponds to the experimental recombination rate as produced from the Rowan *et al*. (2019) dataset. R^2 is the fraction of the variance explained by the model; it inevitably increases as bin size decreases because the CO numbers per Mb are more subject to stochastic noise.


Chr2
Experimental
Predicted
R^2 = 0.47
Recombination rate (cM/Mb)
Mb


Chr3
Experimental
Predicted
R^2 = 0.59
Recombination rate (cM/Mb)
Mb



Supplementary Figure S9. Experimental and predicted recombination landscapes of chromosomes 2 to 5. Landscapes using 100 kb bins were produced from the Rowan *et al*. dataset (red) and from our quantitative model with 15 adjustable parameters (blue). Each inset shows a corresponding zoom within the right arm. R^2 is the fraction of the recombination rate variance that is explained by the model.