Ubiquitous recombination gradients within plant genic regions shaped by recombination hotspots

Thomas Brazier[1*] **and Sylvain Glémin**[1,2]

[1]University of Rennes, CNRS, ECOBIO (Ecosystems, Biodiversity, Evolution), Rennes, France

[2]Department of Ecology and Genetics, Evolutionary Biology Center and Science for Life Laboratory, Uppsala University, Uppsala, Sweden

Corresponding author: thomas.brazier@univ-rennes.fr

November 18, 2023

Abstract

During meiosis of many eukaryote species, crossovers tend to occur within narrow regions called recombination hotspots. In plants it is generally thought that gene regulatory sequences, especially promoters and 5'-3' untranslated regions, are enriched in hotspots, but this has been characterized in a handful of species only. We also lack a clear description of fine-scale variation in recombination rates within genic regions and little is known about hotspot position and intensity in plants. To address this question we constructed fine-scale recombination maps from genetic polymorphism data and inferred recombination hotspots in eleven plant species. We detected gradients of recombination both in 5' and 3' of genic regions in most species, yet gradients varied in intensity and shape depending on specific hotspot locations and gene structure. To further characterize recombination gradients, we decomposed them according to gene structure by rank and number of exons. We generalised the previously observed pattern that recombination hotspots are organised around the boundaries of coding sequences. However our results suggest that variations among species were driven more by hotspot location among and within genes than by differences in size or intensity among species. Our results shed light on the variation in recombination rates at a very fine scale, beyond genes averages across the whole genome, revealing the diversity and complexity of genic recombination gradients emerging from the interaction between hotspot location and gene structure.

Introduction

Meiotic recombination is a general feature of sexually reproducing species. During meiosis, crossovers (COs) ensure proper segregation of homologous chromosomes and reshuffle alleles from the two parental genomes. The distribution of CO is not homogeneous along chromosomes and varies at large and short scales [1–4]. In many species, recombination tends to occur within narrow regions called recombination hotspots [5–8].

In many mammals, the location of recombination hotspots is directed by the PRDM9 protein [9, 10]. The zinc-finger domain that binds to specific DNA motifs is evolving fast, thus changing recombination landscapes over short evolutionary times [11, 12]. The PRDM9 gene is believed to be conserved across several vertebrate lineages because it redirects recombination away from the transcribed region to avoid deleterious mutations and chromosomal rearrangements [9, 13, 14]. Meanwhile, plants, as fungi and many animals, do not have a PRDM9-like system [13]. In species where PRDM9 is lacking, recombination was found to occur preferentially in gene regulatory sequences, especially in promoters and 5'-3' untranslated regions (UTRs) [15–20]. Because this was found in very phylogenetically divergent species such as plants, yeasts or birds [17, 21, 22] and when PRDM9 is experimentally knock-outed [16], it is thought to be the ancestral mechanism for locating recombination hotspots. Mechanistically, recombination hotspots seem globally associated with nucleosome occupancy and methylation patterns that define flanking regions of genes [4, 17, 23]. However, the spreading of recombination hotspots within coding sequences is recurrent across eukaryotes, despite lower recombination rates within genes compared to flanking regions on average, and the ultimate determinants of the distribution of crossovers along the coding sequence are not clearly resolved [20, 24–28].

In a few plant species it was observed that recombination is not uniformly reduced within coding sequences [25, 29–32]. In rice, recombination hotspots clearly overlap genes between the Transcription Starting Site (TSS) and the Transcription Termination Site (TTS), potentially inflating recombination in the first and last exons [32]. In *Mimulus guttatus*, recombination is higher in the first exon than in 5' noncoding sequences [25]. Even if recombination hotspots strictly remain in flanking regions, recombination tracts span hundreds of base pairs and are likely to end within genes. Indeed, 5' ends of genes experience gene conversion gradients [33–35] and the combination of GC-biased gene conversion (gBGC) and 5'-3' recombination gradients would provide an explanation for GC gradients along genes commonly observed amongst plant species [36–39]. Therefore actual GC gradients are indirect evidence that recombination gradients could be widespread and organised as a function of the number of exons in addition to the distance to TSS and TTS, as observed for GC content [38]. However, direct evidence is scarce. So far, the fine-scale variation in recombination along gene bodies has not been carefully assessed outside a fistful of species, including two plants [15, 17, 25, 40] and we ignore if it is a general tendency. Global 5'-3' decreasing gradients of recombination at the beginning of genes have been observed in *Arabidopsis thaliana* and *Mimulus guttatus*, and rice, maize and potato to a lower extent [17, 25, 31, 32, 41], but we lack a clear description of recombination gradients from start to end that also takes gene length and structure into account.

Because the shape and determinants of recombination gradients in genic regions have not been explicitly addressed in previous studies, we tackled this question by constructing fine-scale recombination landscapes in eleven plant species with a large range of genome size and recombination rate and covering several angiosperm

families in both monocots and dicots. Fine-scale recombination rates can be inferred from patterns of linkage disequilibrium in polymorphism data [42]. Owing to tiny distances between genetic markers (1-2 kb) and to the large number of meioses observed in the genealogy of a population, variation in population-scaled recombination rates can be measured within genes and can be used to infer the precise hotspot locations [17, 20, 43, 44]. We estimated recombination rates and the proportion of hotspot overlap in genomic features (exons, introns, UTRs and flanking regions) to assess if recombination in gene promoters and terminators is a common pattern in plants. Are gene promoters and terminators always enriched in crossovers? Is this positioning in 5' and 3' ends creating a gradient of recombination within coding sequences? How do hotspot intensity and gene structure (exons/introns) interact to shape the recombination gradient? And finally, how does it affect the recombination rate of genes?

Results

LD-based recombination landscapes

In order to achieve a fine resolution at a gene scale in LD-based recombination maps, we gathered high-density polymorphism datasets in eleven flowering plant species (Table 1, Table S1). We identified species of interest, representing the diversity of plant genomes (small/large chromosomes, eudicots/monocots) and broad-scale recombination patterns (low/high genome-wide recombination rates), from a previous study [1]. We also used a human dataset [45] to compare the results with a species with hotspots targeted outside genic regions due to the PRDM9 mechanism. To ensure the comparability of the results we run the same pipeline on this human dataset instead of directly using published recombination maps. After filtering SNPs (minor allele frequency < 0.05, missing data per site < 0.1, bi-allelic SNPs, genotype quality score ≥ 30), we kept only datasets with at least 0.5 SNP/kb (range 0.6-11.8 SNP/kb, Table 1).

Fine-scale recombination landscapes have been estimated with LDhat [42] using a custom pipeline (`https://github.com/ThomasBrazier/ldhat-recombination-pipeline.git` v1.0). We checked for hidden population structure and sampled individuals within a single consistent genetic group with the highest polymorphism level, accordingly. We considered haploid genomes for selfing species. The population size parameter (θ_π) was estimated from the sampled population. For computational limits a maximum of forty diploid genomes per species (80 haploid genomes for selfing species, autosomes only) were sampled for LDhat analyses since the quality of estimates does not improve much over twenty individuals [46]. The convergence of LDhat MCMC chains have been assessed with LDhat's diagnostic tools. We checked the reliability of LD-based recombination landscapes by visual comparison (Fig. S1) between LD-based landscapes and broad-scale pedigree-based landscapes (Marey maps) of Brazier and Glémin [1]. Only three species showed signs of limited resolution (*Camellia sinensis*, *Phaseolus vulgaris* and *Glycine max*). In these species we identified a few genomic regions with a high constant recombination rate but the global landscapes were highly heterogeneous.

The mean genome-wide population-scaled recombination rates (weighted by intervals width) ranged from 0.18 to 8.4 ρ/kb across species. Most of recombination rate estimates values were lower than 5, yet the tails of the distributions of ρ/kb were consistently skewed towards high values across species (Fig. S2). We estimated

a ratio ρ/θ of 0.83 for human. It is close to the ratio of 0.94 found by Kong et al. with pedigree data (2002, 2012), giving us confidence in our estimates for the eleven plants which were in a range close to human. The ratio ρ/θ were between 0.71 and 6.99, except for *Camellia sinensis* (ratio $\sim$ 21.6), suggesting that LDhat estimates for this species may be less reliable than for the eleven other species.

Table 1: Summary statistics of twelve datasets (eleven plants and one human). Species, reference of the original polymorphism data, number of SNPs after filtering, genome length (Mb), SNP density (number of SNPs per kb), number of chromosomes of the genome (number of chromosomes removed from analyses in parenthesis), number of diploid genomes sampled, mean and median ρ/kb, estimated $\hat{\theta}_\pi$ and ratio ρ/θ.

Species	Reference	# SNPs	Mb	SNP/kb	# chr.	# diploid genomes	mean ρ/kb	median ρ/kb	$\hat{\theta}_\pi$	ρ/θ
Arabidopsis thaliana	The 1001 Genomes Consortium [49]	639,942	119	5.4	5	40	1.26	0.84	0.0014	0.92
Camellia sinensis	Zhang et al. [50]	1,811,170	2985	0.6	15	40	3.89	3.14	0.0002	21.57
Citrullus lanatus	Guo et al. [51]	362,063	362	1.0	11	40	0.27	0.0007	0.0004	0.71
Glycine max	Yang et al. [52]	4,254,477	914	4.7	20 (1)	40	5.23	0.43	0.0012	4.44
Homo sapiens	Sudmant et al. [45]	6,584,022	2796	2.3	22	40	0.70	0.007	0.0008	0.83
Malus sieversii	Sun et al. [53]	6,086,695	516	11.8	13	37	8.43	3.40	0.0036	2.32
Oryza sativa indica subsp.	Wang et al. [54]	2,208,563	344	6.4	11	40	0.87	0.23	0.0021	0.42
Phaseolus vulgaris	Wu et al. [55]	1,823,344	515	3.5	11 (1)	40	7.42	0.28	0.0011	6.99
Populus tremula	Liu et al. [56]	1,557,963	288	5.4	14	36	2.06	1.47	0.0011	1.87
Sorghum bicolor	Lozano et al. [57]	423,782	683	0.6	10	32	0.18	0.02	0.0002	1.11
Spinacia oleracea	Cai et al. [58]	3,009,618	879	3.4	6	40	0.84	0.42	0.0011	0.79
Triticum aestivum B subgenome	Zhou et al. [59]	6,173,805	9140	0.7	7	40	0.35	0.003	0.0004	0.86

Recombination hotspots are general but vary in intensity among species

Based on the LDhot statistical framework we inferred LD-based recombination hotspots in eleven plant species as well as in human in a standardized and comparable manner (Table 2, Fig. 1A). We applied two levels of filtering to the hotspot call. As recombination hotspots are generally defined as narrow peaks of recombination over a short genomic range, we first filtered out all hotspots larger than 10 kb without making assumptions on their intensity (soft filtering). In order to improve our power to detect fine scale patterns and associations with short targets (e.g. TSS, TTS), we did a harder filtering removing every hotspot larger than 10 kb and with an intensity lower than 4 and higher than 200 (putatively false hotspots due to variance in LDhat estimates). The hotspot intensity was measured as the peak rate divided by the background recombination rate (the mean background rate in a 50 kb window around the hotspot centre). The two levels of filtering drastically reduced the number of hotspots but in similar proportions across species (Table 2). Hotspot filtering had the same impact for all species. Hotspot size was reduced to approximately 6-8 kb and the hotspot intensity increased, as expected by the two criteria we chose (Fig. 1A).

The genome wide hotspot density range was of the same order of magnitude between plants and human (12.1-42.5 hotspots/Mb for unfiltered data), independently of the filtering strategy (Table 2). The number of hotspots per species was significantly correlated with genome size (Spearman's rank correlation test, ρ = 0.89, ρ = 0.87, ρ = 0.89 for unfiltered, soft and hard filtered datasets, p $<$ 0.001 for the three; Table 2) but not significantly correlated with SNP density (Spearman's rank correlation test, ρ = -0.26, p = 0.42, ρ = -0.20, p = 0.51, ρ = -0.24, p = 0.46 for the three datasets, respectively), suggesting that the number of hotspots is more closely associated with differences of genome size than differences in data quality. These results were supported by the fact that recombination rates varied over many orders of magnitude at a fine scale and were concentrated in a short fraction of the genome (Fig. 1B, Fig. S2). For most species 80% of the total recombination was concentrated within 20% of the genome. The species-specific patterns were roughly the same between plants and human. Our results for human, rice and *Arabidopsis thaliana* were globally

similar to state-of-the-art previous studies [17, 20, 43, 60], both for the shape of hotspots and the cumulative distribution function (Fig. 1A,B) and the number of hotspots (Table 2). Three species (*Arabidopsis thaliana*, *Malus sieversii* and *Populus tremula*) showed slightly less heterogeneity than the rest of the dataset (Fig. 1B). We noticed more variability between *Oryza sativa* chromosomes than for other species (Fig. 1B). Overall our results showed that recombination was globally concentrated in a short fraction of the genome across eleven plant species, suggesting that recombination hotspots may be a general feature in plants, similarly to human.

Table 2: Number of hotspots (density of hotspots per Mb). Statistics are given for the unfiltered, soft and hard filtering datasets. The number of hotspots already known in the literature is given at the end.

Species	# hotspots	density/Mb	# (soft)	% (soft)	density/Mb (soft)	# (hard)	% (hard)	density/Mb (hard)	# (literature)	Reference
Arabidopsis thaliana	5,059	42.5	2,587	51.1	21.7	1,123	22.2	9.4	8,448	[17]
Camellia sinensis	41,085	13.8	13,643	33.2	4.6	7,137	17.4	2.4		
Citrullus lanatus	8,232	22.7	4,117	50	11.4	2,036	24.7	5.6		
Glycine max	34,528	36.4	20,398	59.1	21.5	10,223	29.6	10.8		
Homo sapiens	94,851	32.9	53,946	56.9	18.7	34,645	36.5	12	25,000-50,000	[20]
Malus sieversii	27,334	41.9	13,000	47.6	19.9	3,605	13.2	5.5		
Oryza sativa	10,847	29.1	5,567	51.3	14.9	3,934	36.3	10.5	14,125	[32]
Phaseolus vulgaris	13,477	26.2	7,277	54	14.1	3,355	24.9	6.5		
Populus tremula	13,856	38.3	6,689	48.3	18.5	2,349	17	6.5		
Sorghum bicolor	12,264	17.9	6,449	52.6	9.4	3,080	25.1	4.5		
Spinacia oleracea	32,220	36.7	21,142	65.6	24.1	9,865	30.6	11.2		
Triticum aestivum	122,595	12.1	65,045	53.1	6.4	53,454	43.6	5.3		

Despite qualitatively similar sizes, hotspots varied in intensity among species (Fig. 1A,C, Fig. S3). Three species exhibited a lower level of hotspot intensity (*Arabidopsis thaliana*, *Malus sieversii* and *Populus tremula*, mean intensity = 8.1, 5.2 and 5.6, respectively) and *Triticum aestivum* exhibited the higher level of hotspot intensity (mean intensity = 67.9). Hotspot intensities above 200 were not rare for many species. After soft filtering, the normalized hotspot peak rate of *Citrullus lanatus*, *Populus tremula* and *Arabidopsis thaliana* and two other species were largely above the genome-wide recombination rate, suggesting that the contribution of hotspots to the genome-wide rate was more important in these species (Fig. 1C). On the contrary, the normalized hotspot peak rate of *Phaseolus vulgaris* was even lower than the genome-wide recombination rate. The six other species varied between these two extremes, with human in the middle. For unfiltered hotspots, eleven out of twelve species (except *Phaseolus vulgaris*) showed higher recombination rates at hotspot centre than the genome wide average recombination rate (Fig. 1A). Overall, the species-specific intensity of hotspots and their contribution to the genome wide recombination rate seemed to be distributed along a continuum from intense strongly contributing hotspots to more diffuse hotspots weakly contributing.

By comparing the background recombination rate (± 50 kb) and the genome-wide rate, we assessed that hotspot detection was sometimes biased in favor of genomic regions recombining on average slightly less or more than the genome wide average (see the difference between soft lines and the dashed line in Fig. 1A). For the unfiltered and soft filtered datasets, the bias was absent in three species and weak in eight other species, suggesting that we were not strongly limited to detect hotspots in regions experiencing extremely high or low recombination rates. On the contrary, in *Phaseolus vulgaris* the detection of hotspots was particularly biased towards lowly recombining regions. In *Phaseolus vulgaris*, *Glycine max*, *Camellia sinensis* and *Malus sieversii* the bias was more pronounced for the hard filtered dataset, probably because intense hotspots are more easily detected on a low recombining background. In general the soft filtered dataset seemed relatively more reliable for the exhaustive search of hotspot location along the genome since hard filtering was biased towards low recombining regions in a few species. However the power of the hard filtering dataset to test

for associations between hotspots and target genomic features should not be negatively affected, as only the colder hotspots remained underdetected. Subsequently we present results for the soft filtered dataset.

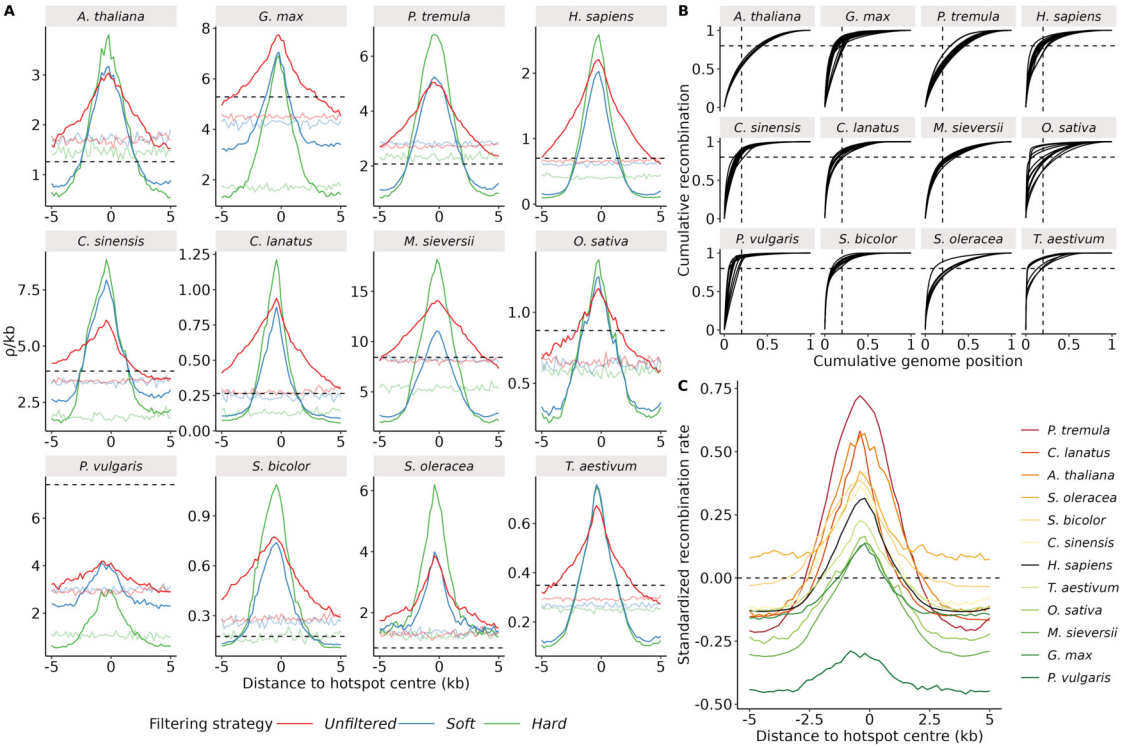

Figure 1: The distribution of recombination is heterogeneous at a fine scale and concentrated within recombination hotspots. (A) The recombination rate (ρ/kb) as a function of the genomic distance to the crossover hotspot centre (kb) with three different filtering strategies. Crossover hotspots were inferred with LDhot. The LDhot results were filtered to remove putatively false hotspots. Soft filtering removed hotspots larger than 10 kb. Hard filtering removed hotspots larger than 10 kb and hotspots with an intensity lower than 4 or higher than 200. Hard lines are the recombination rates around hotspot centre (midpoint of the hotspot interval) and soft lines are the control by randomly resampling hotspot intervals within a $\pm$ 50 kb neighboring region, hence soft lines represent the variance of the background recombination rates around hotspots. The dashed horizontal line is the mean genome-wide recombination rate. (B) The cumulative ordered recombination fraction as a function of the cumulative genome fraction for each chromosomes and species. We represented dashed vertical lines at 20% of the genome and dashed horizontal lines at 80% of the total recombination. (C) Normalized recombination rates (centered-reduced recombination rates, i.e. the hotspot recombination rate minus the species genome-wide recombination rate, all divided by the genome-wide standard deviation) as a function of the distance to the hotspot centre for the soft filtered dataset. Species in the legend ordered by descending peak height. Human is in black.

Recombination is globally concentrated in 5' and 3' ends of genes

When comparing genic and intergenic recombination rates, we observed three different patterns (Fig. 2A, Fig. S4A). *Glycine max*, *Malus sieversii*, *Oryza sativa*, *Phaseolus vulgaris* and human had lower recombination

rates in genic regions than intergenic ones. On the contrary, genic regions were enriched in recombination in *Populus tremula*, *Sorghum bicolor* and *Triticum aestivum*. Other species showed no clear differences between genic and intergenic rates. Flanking regions around gene features generally correspond to promoters and other cis-regulatory elements and are supposed to be enriched in crossovers. In order to remove any effect of the flanking regions, we also looked at intergenic regions after removing ± 3 kb flanking regions around genes (intergenic with a buffer) and did not observe clear differences in the previous patterns, despite significant differences between intergenic and intergenic with a buffer in a few species. On average, we did not observe a general pattern between genic and intergenic regions, but we clearly detected variation among species at this scale.

Interestingly, genic regions were globally enriched in recombination hotspots (Table 3, Table S2). In ten plant species, and so as in human, the number of genic hotspots (overlapping a gene feature) was significantly higher than expected (random expectation computed by 1,000 iterations of random shuffling of hotspots ranges). In *Triticum aestivum* the number of genic hotspots was less than expected just by chance.

In general, recombination within the coding region was slightly lower than within Untranslated Regions (UTRs) and the flanking regions, excepted for *Populus tremula*, *Glycine max* and *Camellia sinensis* which showed on average higher recombination rates within the coding part. In eight plant species and in human, recombination was enriched at the 5' and 3' ends of genes (Fig. 2B). Mean recombination rates (weighted by sequence length) were lower in exon/intron features than in surrounding regions, such as UTRs and flanking regions (the unweighted mean yielded qualitatively similar patterns; Fig. S4B). In *Populus tremula*, the recombination was higher at the 5' and 3' ends, with the highest recombination rates within exons. Intriguingly, for *Camellia sinensis* genic recombination was higher in exons than in flanking regions. *Glycine max* showed a lower recombination rate in UTRs than flanking regions and exons/introns. We lacked information to annotate UTRs in *Citrullus lanatus* but recombination rates were reduced in the transcribed region. We did not detected any clear pattern for differences in averaged recombination rates between exons and introns.

Table 3: Number of genic and intergenic hotspots, i.e. hotspots overlapping or not a gene feature (soft filtered data). The expected number of genic hotspots was computed by 1,000 random shuffling of hotspots ranges.

Species	# Intergenic hotspots	# Genic hotspots	# Expected genic (95% C.I.)
Arabidopsis thaliana	192	2395	617 (577-655)
Camellia sinensis	7947	5696	782 (729-837)
Citrullus lanatus	2100	2017	587 (545-627)
Glycine max	14226	6172	5065 (4959-5182)
Homo sapiens	40846	13100	11583 (11419-11751)
Malus sieversii	6976	6024	2801 (2708-2888)
Oryza sativa	2089	3478	1053 (997-1111)
Phaseolus vulgaris	4373	2904	1366 (1305-1427)
Populus tremula	1349	5340	1415 (1359-1476)
Sorghum bicolor	3387	3062	785.5 (734-835)
Spinacia oleracea	15471	5671	5392 (5277-5511)
Triticum aestivum	61872	3173	9749 (9566-9920)

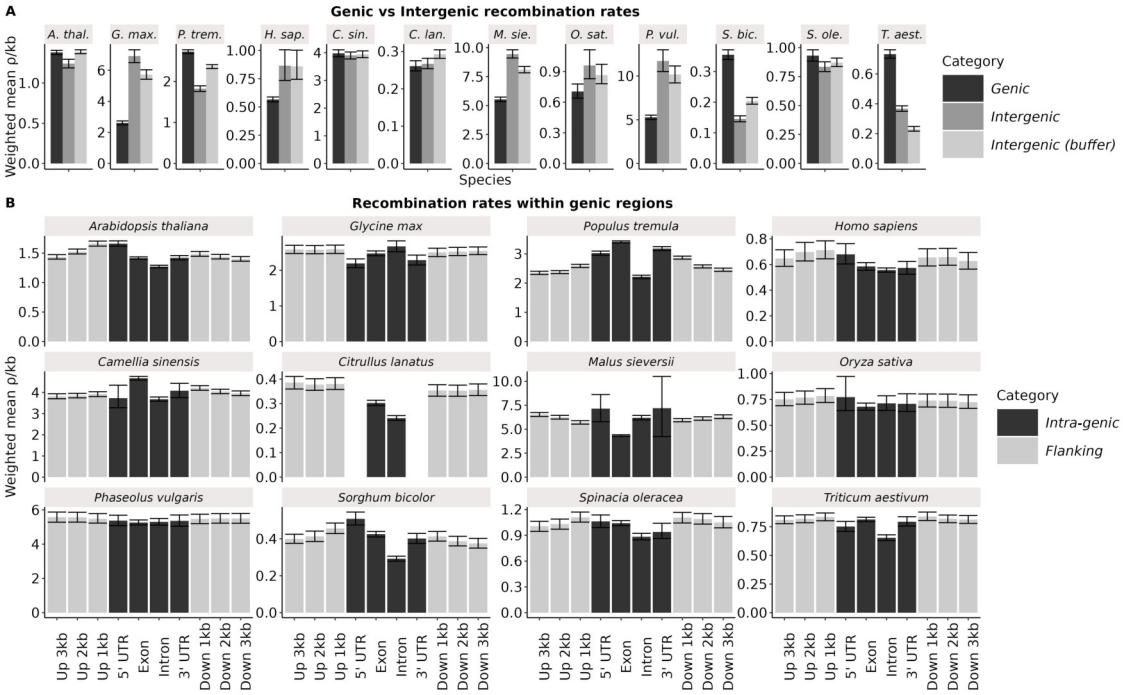

Figure 2: Variations in recombination rates (ρ/kb) around and within genes. (A) Recombination rates vary between genic and intergenic regions. In order to remove an effect of 5' and 3' regulatory regions at the proximity of genes, buffered intergenic regions were defined by excluding 3 kb flanking regions upstream and downstream of genes. (B) Within genic regions, the recombination rate varies between genomic features as well as in 5' upstream and 3' downstream flanking regions. The weighted mean recombination rate (mean recombination rate weighted by the sequence length in kb) and 95% confidence intervals were estimated by 1,000 bootstraps. Annotations of UTRs were not available for *Citrullus lanatus*.

Characterising 5'-3' gradients of recombination in genes

Based on Figure 2, we chose three species (*Arabidopsis thaliana*, *Populus tremula*, *Glycine max*) illustrating the diversity of recombination patterns to characterise more precisely 5'-3' genic gradients (Fig. 3). In *A. thaliana*, the 5' flanking region and UTR exhibited a plateau with higher recombination rates whereas a 5'-3' steep decreasing gradient began just after the start codon of the coding sequence. The 3' end of the gene showed a small peak within the coding sequence just before the Transcription Terminator Site (TTS), but weaker than the 5' end plateau. This pattern was in adequacy with previous results considered as the standard pattern in plants [17]. In *G. max* we observed a 5' peak centred on the ATG/TSS, yet weaker than in *A. thaliana* (note the different scales for the recombination rate). *P. tremula* showed marked differences with the pattern observed in *A. thaliana*. We observed two narrow peaks inside the coding sequence, centered on the CDS start instead of TTS and on the 3' UTR instead of TTS. Contrary to *A. thaliana* and *G. max* we also observed a higher peak in 5' than in 3'. In other species we also found gradients in 5' TSS and 3' TTS, yet many of them were weak (Fig. S5).

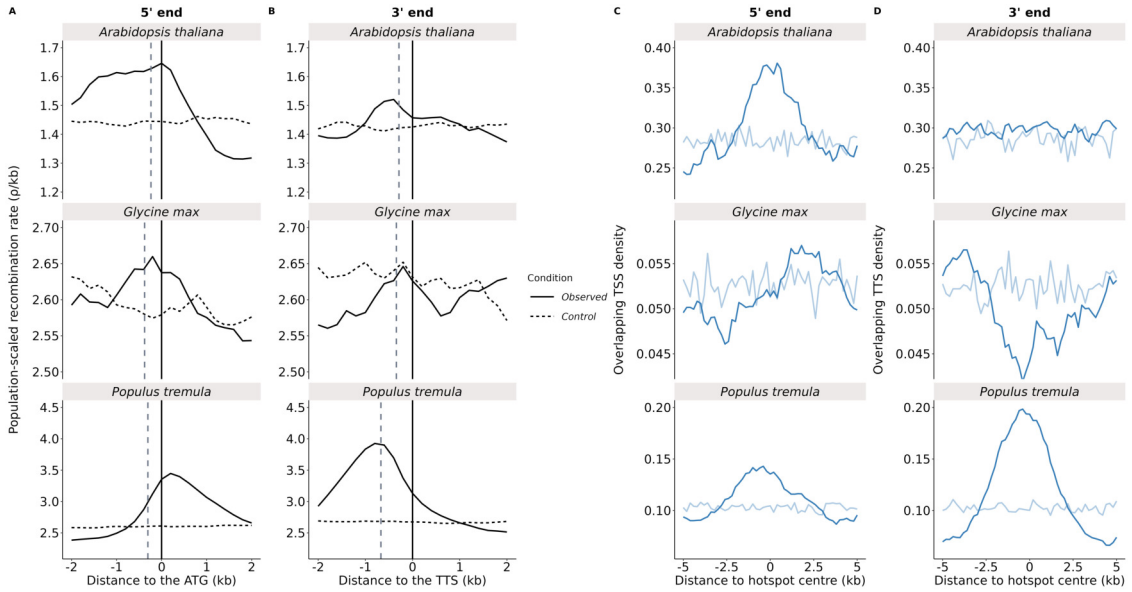

Figure 3: Recombination gradients along genes are mostly around Transcription Starting Sites (TSS) and Transcription Terminating Sites (TTS). (A, B) The recombination rate (ρ/kb) was estimated in 200 bp windows as a function of the distance to the ATG start codon (start of first CDS) or the TTS and averaged among introns and exons. The vertical dashed lines are respectively the mean distance to the TSS (i.e. the 5' UTR, panel A) and the mean 3' UTR length (panel B). Random controls by randomly resampling intervals in the genome and re-scaled to the mean of the gradient. (C, D) The density of TSS and TTS as a function of the distance to hotspot centre (soft filtered hotspots). For smoothing, TSS and TTS were defined as regions of 1 kb centered on the gene start and end position respectively. Soft lines are a control by randomly resampling hotspot intervals in a $\pm$ 50 kb neighboring region.

The peaks of recombination and the decreasing gradients around them were largely correlated with the presence of hotspots in these regions (Fig. 3C, D). We plotted the density of TSS/TTS around the hotspot centre to graphically test for an enrichment in crossovers at this genomic features. In order to improve our power to detect an association we used the hard filtered hotspot dataset. In *Arabidopsis thaliana*, hotspots were associated with TSS but not TTS. In *Glycine max*, there was no clear association between hotspots and TSS and we observed a decrease in TTS overlap within hotspots. In *Populus tremula* hotspots overlap both TSS and especially TTS more frequently than expected by chance.

Hotspots and gene structure shaped the gradient of recombination

We decomposed the gradients in exons/intron ranks to remove spatial averaging effects among genes of different lengths and evaluate the distribution of recombination within gene bodies (Fig. 4). Recombination gradients were polarised from the start to the end of genes. In *A. thaliana* and *G. max*, the average decreasing gradient spanned over the five to seven first exons and reached a minimal plateau. In contrast, in *P. tremula*, after a decrease in 5', recombination rates re-increased from the middle of genes to the 3' end, yielding a U-shape gradient. We observed only slight differences in gradients among exons and introns (Fig. 4, Fig. S6).

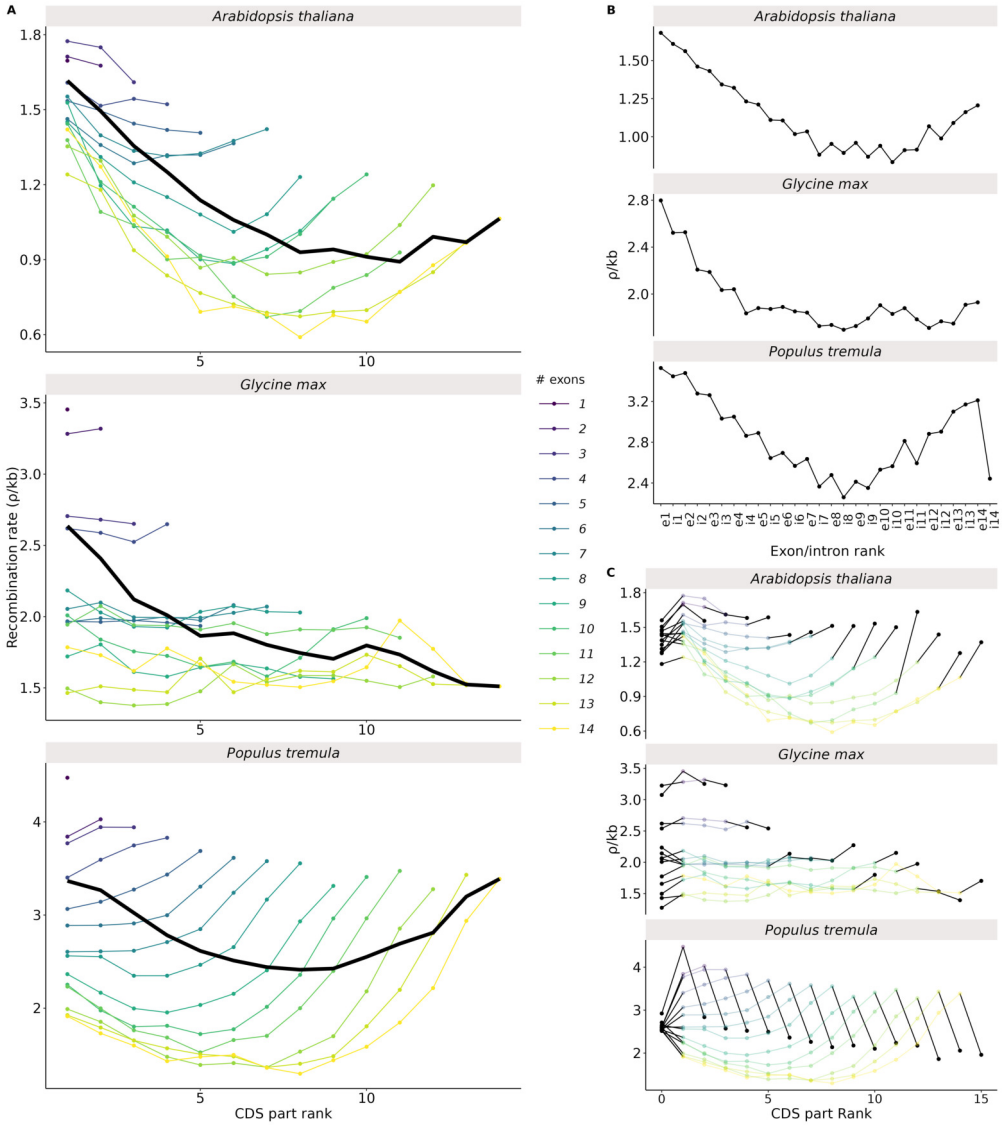

Figure 4: Gradients of recombination (ρ/kb) along genes as a function of exon/intron rank. (A) Comparison of independent gradients as a function of exon rank (CDS part) with genes pooled by their number of exons. The black line is the averaged gradient (all genes pooled). (B) Average gradient of recombination rate as a function of exon of rank i (e_i) or intron of rank i (i_i). (C) Gradients in flanking regions representing the differences of recombination rates between genic regions and their flanking non-coding regions. On each side of the gene, the recombination rate of the 5' and 3' flanking regions (3 kb windows) is represented by a black dot and the difference between the gene part and its close flanking region is represented by a black line.

Since recombination hotspots were mostly in the 5' and 3' ends of genes, we evaluated how gene structure (rank and number of exons) could shape differences in gradients among species as in Ressayre et al. [38]. We decomposed the averaged gradients across ranks into independent gradients for each exon-number class of genes (Fig. 4A). In all three species, shorter genes (genes with fewer exons) recombined more than longer genes on average (Fig. 4A, Fig. S7). However, it was not entirely driven by lower recombination rates in

the middle of genes. The first exon(s) of shorter genes experienced higher recombination rates than the first exons of longer genes. Gradients per gene class were more or less parallel to the average gradient in *A. thaliana*, though a slight increase was detected in 3' ends of longer genes (longer than seven exons). On the contrary, per gene class gradients were flat in *G. max* and the average gradient was shaped by an excess of recombination in shorter genes (shorter than five exons). The structure of *P. tremula* gradients was more complex than the two previous patterns. The U-shape average gradient was produced by a combination of per gene class gradients increasing in 3' and higher recombination rates in the first exons of shorter genes.

We checked if recombination rates at genes boundaries were correlated with the local recombination rate within flanking regions. In *Arabidopsis thaliana* and *Populus tremula*, recombination rates at 5' and 3' flanking regions were similar among gene sizes (number of exons) and decoupled from variations at the boundaries of the transcript sequence (Fig. 4C). This pattern was particularly clear for *P. tremula* for which almost all of the variation in recombination was concentrated within the genic region. In *A. thaliana* we noted a 5' flanking recombination rate consistently lower than the recombination rate of the first exon but a more elevated recombination rate in 3' flanking regions than in the last exon. In *Glycine max*, we observed no differences in recombination between flanking regions and the flat genic gradients per gene class. In the three species recombination rates were similar between the 5' and 3' ends.

Diversity of gradients across species

We applied the same analyses to our eleven plant species in order to identify common patterns across species. We also analysed the human dataset as a control species where hotspots are mainly located outside genic regions. We observed a diversity of patterns, more or less similar to the three species we detailed above. Recombination patterns were mostly determined by differences in hotspot location and intensity. Average 5'-3' recombination gradients were common across the eleven plant species and even in human (Fig. 5A, black lines), except in *T. aestivum* for which the power to detect gradients was limited (see methods for details).

As for the three species we chose as an example, we also found three main different patterns. In *Arabidopsis thaliana*, *Sorghum bicolor*, *Malus sieversii* and *Citrullus lanatus* the detailed gradients per gene class were parallel to the average gradient and matched what was already known in plants [17, 25]. In *Glycine max*, *Phaseolus vulgaris*, *Oryza sativa* (less clear because of outliers) and *Spinacia oleracea* the average gradients observed were shaped exclusively by differences in recombination rates between genes with a different number of exons. Gradients per gene class were flat, suggesting a very weak or null gradient in individual genes or that we lacked the resolution to detect a gradient. Finally, in *Populus tremula* the gradient was clearly driven by hotspots at 3' ends combined with differences in recombination between gene classes, thereby shaping a U-shaped gradient. In *Camellia sinensis* the gradient was mostly increasing towards the 3' end but without a 5' gradient. For the detailed gene class gradients, we did not observed recombination gradients in *Homo sapiens* and even the poorly resolved *T. aestivum* showed more structure between gene classes than human. As a consequence of hotspot positioning in 5' and 3' in most species (Fig. 3C,D, Fig. S8), it seemed general among plants that shorter genes recombine more on average (Fig. 5A, Fig. S7).

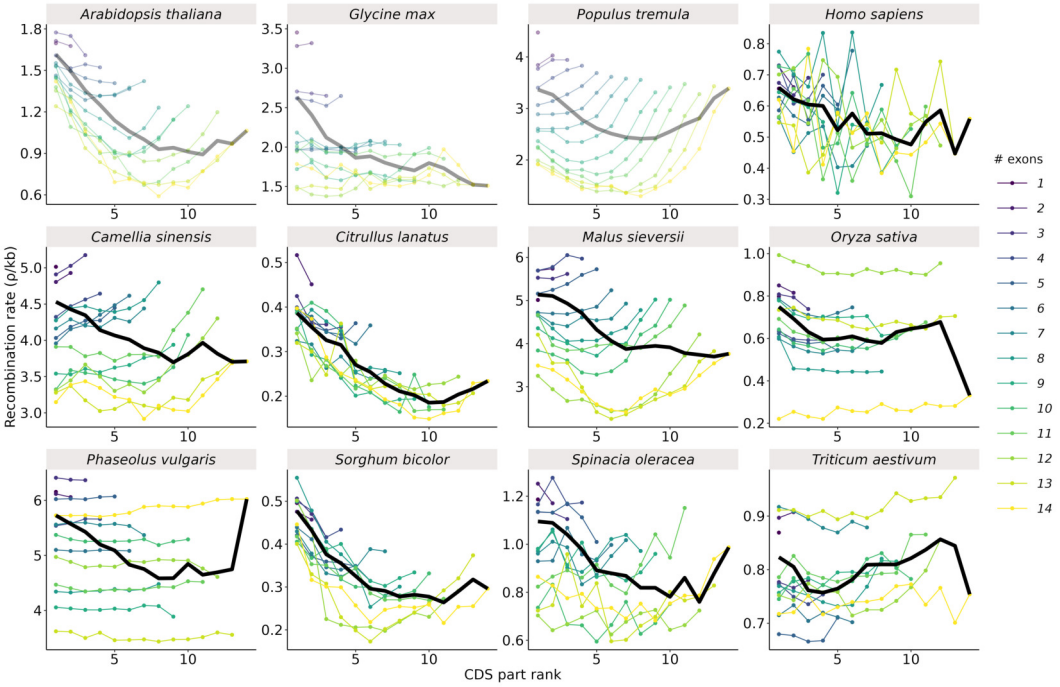

Figure 5: Diversity of recombination gradients across plant species. The recombination rate (ρ/kb) was estimated in exons (CDS part) as a function of their rank and genes were grouped by their number of exons. The black line is the averaged gradient (all genes pooled). The three reference species already shown above are transparent. All protein-coding genes were used.

Gradients were not clear in some of the species, particularly *Oryza sativa*, *Triticum aestivum* and *Spinacia oleracea*. To test whether it was a genuine biological signal or simply an artefact due to noisy data we applied two filtering strategies. First, we only kept the most polymorphic genes which clearly improved gradients in *O. sativa* and *S. oleracea* (Fig. 6A). As hotspot intensities could be flatten by a high genome-wide or regional background recombination rate (Fig. 1B, D), we also filtered out the 25% highly recombining genes for each species thus revealing even more resolved gradients in *O. sativa* and *S. oleracea* and a steep gradient in the three first exons of human genes (Fig. 6B). *T. aestivum* gradients were not improved by these filtering steps, but the level of LD disequilibrium was high in this species and the LD decay was very slow, even in genic regions, suggesting a lack of power to detect fine-scale patterns (Fig. S9). Overall, among eleven plant species, filtering genes by polymorphism or recombination rate revealed more resolved and complex gradients than initially observed (Fig. S10, Fig. S11).

In order to support the ubiquity of recombination gradients in our eleven plant species, we assessed if the diversity of recombination gradients we observed could be parallel to hotspot gradients (Fig. 6C, Fig. S12). Hotspot gradients were clearer and more consistent among species than recombination gradients, with a few exceptions (Fig. 6C). Hotspot gradients were detected in all twelve species (Fig. S11), which suggested that with a sufficiently polymorphic dataset we should be able to detect recombination gradients in every species. Overall, filtered gradients and hotspot distribution suggested that recombination gradients could be a general pattern among plants.

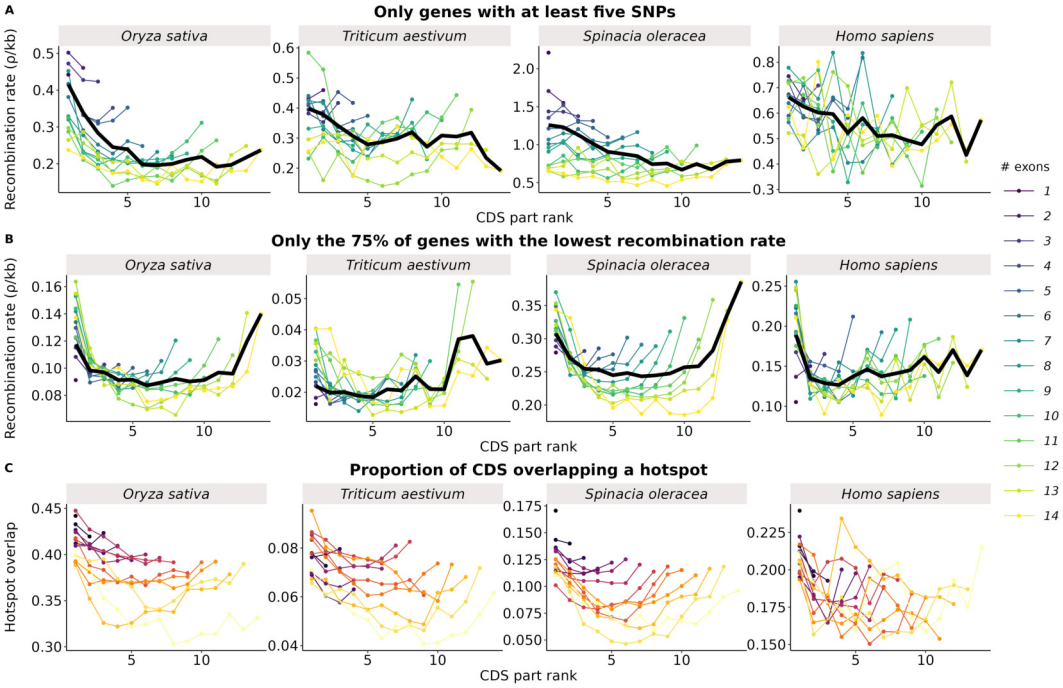

Figure 6: Improved recombination gradients in the less resolved species by filtering genes or mapping hotspot overlap.The recombination rate (ρ/kb) was estimated in exons (CDS part) as a function of their rank and genes were grouped by their number of exons. The black line is the averaged gradient (all genes pooled). (A) Only genes overlapping at least five SNPs were retained (*Oryza sativa* = 14,020 genes, 51.7 %, *Triticum aestivum* = 10,658 genes, 18.6 %, *Spinacia oleracea* = 15,833 genes, 58.1 %, *Homo sapiens* = 12,378 genes, 86.0 %). (B) Only the less recombining genes were retained. For each species, only genes with recombination rates below the third quartile were retained (*Oryza sativa* = 20,289 genes, *Triticum aestivum* = 43,047 genes, *Spinacia oleracea* = 20,430 genes, *Homo sapiens* = 10,786 genes). (C) Gradients of hotspot overlap as a function of exon rank (CDS part). The proportion of CDS overlapping a hotspot in a given class of exon is the number of CDS overlapping at least one hotspot (soft filtered hotspots) divided by the total number of CDS of that class.

Discussion

We characterized recombination patterns in genic regions in eleven plant species, including nine species not described before. We generalised the previously observed pattern that recombination hotspots are organised around the boundaries of coding sequences, both at the 5' and 3' ends of genes. We leveraged high resolution of recombination maps at the exon scale to characterize detailed recombination gradients exhibiting more diversity than assumed until now from only a few model species, involving the interaction between hotspot location and intensity and gene structure.

Recombination varies around and within coding sequences

At the gene scale, the location of recombination hotspots was often concentrated in gene 5' and 3' ends, as previously known for a few plants only [17, 25, 31, 32, 41]. We observed that recombination can be targeted towards diverse domains both in 5' and 3' ends (UTRs, TSS/TTS, promoters, flanking) and was not systematically associated with promoters as previously found [25, 31, 61]. Despite divergent patterns, it is clear that the genic region is finely structured according to the different genomic domains composing it.

Within 5' and 3' ends enriched in COs in *A. thaliana*, chromatin accessibility redirects meiotic recombination towards gene promoters and terminators [17], and the same mechanism also likely occurs in other species [3]. In plants as in yeast, double-strand breaks (DSB) hotspots, the precursors of COs, are found both in TSS and TTS where nucleosome occupancy is reduced whereas COs form close to H2A.Z and H3K4me3 histone marks which are epigenetic modifications of chromatin packaging also involved in the regulation of gene expression [17, 62, 63]. Low methylation is also usually associated with CO hotspots [17]. Both TSS and TTS are hypomethylated in *A. thaliana*, *G. max* and *P. tremula* [64–66], which fits well with our results. These findings are globally in agreement with the tethered-loop/axis model, which postulates that DSBs form globally in chromatin loops bound to the chromosome axis whereas COs occur only in nucleosome-depleted regions that are susceptible to being tethered to the central chromosome axis by recombination-promoting factors [4, 67]. Indeed, gene promoters and terminators are located in chromatin loops and COs most probably concentrate here due to their highly conserved chromatin state and DNA accessibility [4, 68]. However, *P. tremula* and *C. sinensis* do not conform exactly to this canonical model as CO hotspots preferentially occur inside rather than outside coding sequences as in *Mimulus guttatus* [25]. In $PRDM9^{-/-}$ mice, the centre of PRDM9-independent hotspots is on the +1 nucleosome positioned downstream of the TSS [16]. For these species, the detailed description of chromatin and histone patterns during meiosis (especially H2A.Z and H3K4me3) as well as nucleosome maps could be useful to better understand this non-canonical pattern [3, 4].

Although we focused on genic regions, we noticed that recombination rates were sometimes higher within intergenic than within genic regions. Intergenic regions could be more prone to mapping errors, which could artificially inflate recombination rates. If it is a true signal, it would suggest the existence of a mechanism, different form PRDM9, directing recombination towards intergenic regions.

Gradients are shaped by hotspots and gene structure

Recombination gradients are shaped by the distance to crossover hotspots which are close to or even overlap the boundaries of the coding sequence, as already observed in rice, maize or *Mimulus guttatus* [19, 25, 32]. DSB hotspots are typically 1-2 kb wide in plants and a CO is susceptible to occur at each end of the DSB [4]. COs sometimes occur away from recombination initiation sites. In the *a1* maize gene, recombination events resolve in the 5' ends of the transcribed sequence, despite initiation in promoters [69]. In yeast, the density of CO boundaries tends to decay with the distance to the DSB [62]. Our results suggest that hotspots affect recombination in a slightly larger region (± 2-3 kb from the hotspot centre) than DSBs, though the precision may be affected by the relatively large size of some hotspot estimations (see the number of hotspots > 10 kb in Table 2). While the DSB hotspot position remains tightly controlled, recombination tracts can spread further in coding sequences as CO tract length can exceed exon length in plants [70]. In maize, 20% of CO tracts are above 200 bp, which is the average exon length, whereas in *A. thaliana* the average tract length is ∼300-400 bp, up to 1,300 bp [19, 61]. These features are compatible with the recombination gradients we observed in nine species. It also coincides with what is already known of gene conversion gradients in plants, though CO tracts are thought to be longer than non-crossover (NCO) tracts [24, 33, 35, 61].

The intensity and location of hotspots are globally responsible for the polarity of recombination within genes. The estimates of recombination rates show a large variance and can achieve high values within a narrow range (Fig. S2, Fig. 1), thus even a small proportion of hotspots overlapping 5' and 3' ends is sufficient to significantly affect the average recombination rate. More importantly, recombination hotspots seem similar in size but their intensities vary between species (Fig. 1, Fig. S3). In *G. max*, the intensity of hotspots compared to the genome-wide average recombination rate is lower than in the two other species we analysed in details (Fig. 1A,C), and it could explain why gradients are less pronounced in the former species. By comparison, in human also there is variation around and within the coding sequence but it is hard to detect a gradient and recombination rates are not clearly structured in exons compared to plant species (less structured than the flattest gradient observed in *T. aestivum*). The lack of gradients in human is expected when recombination hotspots are directed away from genic regions [16]. However, PRDM9 hotspots are dominant but not exclusive [71, 72] and a fraction of CO hotspots can also occur in promoters. We identified a slight enrichment of promoters in a few human CO hotspots consistent with this idea, but without generating recombination gradients as in plants.

Hotspot position seems to shape the recombination gradient towards the 5' or 3' end. Though the TSS is usually the preferential location of CO hotspots, the increasing gradient towards 3' end in *P. tremula* and *C. sinensis* is associated with the location of a significant fraction of recombination hotspots around the TTS [18]. Besides, TSS and TTS do not seem to be enriched in hotspots in *G. max*, which may explain the flat gradients, contrary to *A. thaliana* and *P. tremula*.

In addition to hotspot location, the structure of genes also matters to shape recombination gradients. An important novelty of our work is to decompose the gradients by exon rank and gene class (number of exons). It was already known that GC gradients were shaped by exon rank and gene class in plants and it was proposed that they could be generated by similar underlying recombination gradients [38]. This hypothesis

has not been directly tested so far but our results are globally in agreement with it. Introns are spacing out exons from recombination hotspots in both 5' and 3' ends, lowering recombination in the middle of long genes. Interestingly, in *A. thaliana*, recombination gradients per rank and gene class fit relatively well with GC gradients whereas it is less clear in rice (compare our Fig. 5 with Fig. 5 in Ressayre et al. [38]).

Finally, gene length not only affects recombination in the middle of genes but it is also lowered recombination for all exons positions. As the proportion of hotspot is invariant with gene length, the same amount of recombination is automatically spread over more exons, making average ρ/kb more or less negatively proportional to gene length. The toy model we propose below can explain this observation simply because shorter genes are on a closer influence of hotspots on both sides than longer genes (Fig. 7). A direct consequence is that shorter genes recombine more than longer genes on average, as we observed in nine out of twelve species.

Reconciliation of the apparent diversity of recombination gradients with a single hotspot model

After filtering genes to improve our power to detect gradients, it seems that recombination gradients are a general feature of plant genic regions. Ten out of eleven plant species experienced a recombination gradient and our power analysis (see methods) showed that in *T. aestivum* the absence of gradient can be explained by a lack of power. Moreover, filtered gradients converged towards the 5' oriented *Arabidopsis*-like or 3' oriented *Populus*-like gradients and flat gradients, as in *Glycine max*, finally appeared as 5' oriented. Flat *Glycine*-like gradients most probably reflect a limited ability to measure gradients, whereas the biological reality is more likely a continuum of patterns between the 5' and 3' patterns.

All these findings can be summarized by formalizing a simple toy model (Fig. 7) initially proposed by Choi and Henderson (2015, see their Figure 2). Hotspots originate with a same total rate in genic regions, distributed either in 5' (in proportion p_5) or in 3' (in proportion p_3 = 1 - p_5) and with recombination tract extending within genes with a given characteristic distance (d_5 or d_3). The proportions p_5 and p_3 may depend on the species (for example $p_5 > p_3$ in *A. thaliana* and $p_3 > p_5$ in *P. tremula*). The rate of decay of recombination (determined by d_5 and d_3) can also vary, for example depending on the heat or size of hotspots. Overall, this model can qualitatively explain all observed patterns, including the flat patterns observed in *G. max* and *P. vulgaris*. This suggests that a same mechanism controlling the fine-scale distribution of recombination could be conserved across angiosperms with only quantitative changes in some parameters. We suggest that such a model could serve as a baseline to test against in further studies on the determinants of recombination landscapes.

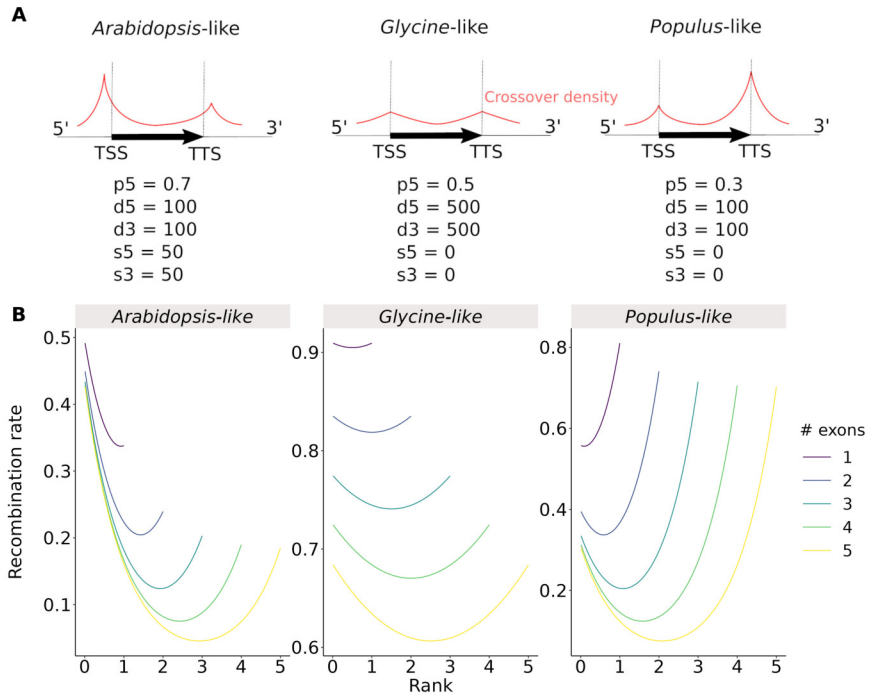

Figure 7: Toy model of the recombination gradient. (A) Conceptual model of recombination gradients adapted from Choi and Henderson [18]. Recombination rates exponentially decay around the TSS and TTS. (B) Quantitative toy model, where hotspots are in 5' and 3', and recombination rates exponentially decay as a function of the distance to the hotspot according to the following equation (1) $r_i = p_5 \exp \frac{x_i+s_5}{d_5} + (1-p_5) \exp \frac{x_i-L+s_3}{d_3}$. Parameters are defined as the relative position of hotspots as a function of the start/stop codon (s_5/s_3), which can accommodate that recombination starts within or outside genes, the range of the hotspot characterised by a characteristic distance controlling the decay in recombination (d_5/d_3) and the relative proportion of hotspots in 5' and 3' (p_5 and $p_3 = 1 - p_5$). The recombination rate r_i at a given position x_i in a gene of size L is estimated by equation 1. The values of parameters (annotated in panel A) are set such as in *Arabidopsis*-like, 5' and 3' hotspots are short, outside the coding sequence, and 5' is dominant. In the flat gradient, 5' and 3' hotspots overlap the gene and 5'/3' are equivalent. In *Populus*-like, 5' and 3' are short, placed in TSS and TTS, and 3' is dominant.

Conclusion

It was known that many organisms, including plants, had recombination hotspots [8, 73, 74], sometimes inducing recombination gradients in genic regions. Our results confirm that it is likely a rather general feature of plant genome but also highlight an unexpected diversity of patterns. Interestingly we observed more differences in hotspot intensity among plants than between humand and *A. thaliana*, for example. We also showed that taking gene structure into account is necessary to unravel the complexity of these recombination gradients that likely emerged from both the location of recombination hotspots and the distribution of genes per class of exon numbers. We suggest that detailed characterisation of recombination and its known covariates (e.g. methylation, chromatin state) at such a fine scale will be helpful to better understand the evolution of recombination hotspots themselves [3, 75], patterns of base composition and gBGC in genic regions [37], or the evolution of intron structure [76].

Materials and Methods

To produce fine-scale recombination maps we analyzed previously published polymorphism datasets in eleven plant species and human (Table 1) starting from VCF files. We identified datasets by literature search with the keywords "resequencing", "high density SNP", "polymorphism data", "variant database", "Whole Genome Sequencing" or "WGS" in association with "plant", "angiosperms" and a set of plant species names. We also searched directly for public genomic databases (e.g. 1001 genomes). We used a previous study of 57 plant species to identify species with interesting and contrasting characteristics, such as monocots/eudicots, small/large genomes, low/high recombination rates, and homogeneous/heterogeneous recombination landscapes [1]. Before downstream analyses, we kept only bi-allelic SNPs and filtered them with the same quality criteria for all datasets (minor allele frequency higher than 0.05, less than 10% of missing data per site, quality score ≥ 30 when the information was available). When the quality score was not available in the published VCF, we checked in the original study that the dataset we re-used was already filtered by a similar quality score.

The genome assemblies and corresponding annotations (GFF files) on which SNPs were originally mapped were retrieved from public databases, mostly NCBI (Table S1). GFFs with gene, mRNA, CDS and exon features were parsed using custom Python scripts to infer introns, 5' and 3' UTRs, flanking sequences (1-3 kb upstream/downstream, in consecutive windows of 1 kb) and intergenic sequences according to the GFF3 specifications. Scripts are freely available as a Python package (`https://github.com/ThomasBrazier/PiSlice.git` v1.0). Only protein-coding genes were retained when this information was annotated. For datasets without information about gene biotype, we considered all genes as protein coding. In order to avoid low sample size for longer genes, we removed genes with more than 14 exons as in Ressayre et al. (2015). When more than one splicing variant has been annotated, the splicing variant with the longest CDS sequence was retained, as it is often also the most expressed sequence. We considered only the CDS part of exons.

All statistical analyses were performed with R version 4.0.4 [77]. Genomic coordinates were manipulated with the R package GenomicRanges [78]. We used a custom pipeline for data processing, SNP filtering, statistical phasing, population structure and re-sampling, recombination rates estimates with LDhat 2.2 [40, 42] and hotspot detection with LDhot [43]. The Snakemake pipeline is freely available (`https://github.com/ThomasBrazier/ldhat-recombination-pipeline.git` v1.0) and the method is presented below.

Population sampling

Population-scaled recombination rates and hotspot detection are based on the assumption of a population without genetic structure and absence of gene flow from other populations [79, 80]. For each dataset, we inferred population structure and sampled one single genetic population as homogeneous as possible and well distinct from other possible genetic clusters. As much as possible, we tried to sample a single large population ($>$ 40 diploid individuals or 80 haploid genomes for selfing species) of wild/landrace individuals with the highest possible polymorphism level. Results and metadata of the original publications were used for a pre-sampling. The genetic structure within the pre-sampled population was assessed with FastSTRUCTURE for K = 1-7 [81]. FastSTRUCTURE was run on a random subset of 100,000 SNPs for computation issues. The best model (best value of K) was inferred by FastSTRUCTURE's algorithm for multiple choices of

K and validated by the visual assessment of admixture plots. If a population substructure was positively assessed, individuals were clustered into different genetic populations according to their mean admixture proportion. Clusters with higher proportions of missing data and low polymorphism (heterozygosity and $\hat{\theta}_\pi$) were discarded and we kept individuals in the largest remaining genetic cluster.

Population-scaled recombination rates

After population sampling, the VCF was phased with Shapeit2, with a phasing window size of 2 Mb [82]. Selfing species were treated as haploid individuals, by randomly sampling only one phased haplotype per individual. When the sample size was sufficient, at most 40 diploid individuals (or 80 haploid sequences for selfing species) were randomly subset for LDhat 2.2 for computational issues.

The population-scaled recombination rates ρ were estimated with LDhat 2.2 [40, 42]. For the estimate of genetic diversity required by LDhat we used the genome averaged $\hat{\theta}_\pi$ estimated with VCFtools on the population dataset [83]. The per-base mutation rate was set to $1.28.10^{-8}$ for all species. When $\hat{\theta}_\pi$ was in the same order of magnitude as a pre-generated look-up table ($\theta = 0.01$ or 0.001), the look-up table was generated from this pre-existing one for computation issues. Otherwise, a complete look-up table was generated with the θ estimate. The maximum value of ρ was set to 100, as suggested in the LDhat manual. Recombination rates between pairwise adjacent SNPs were estimated with the LDhat 'interval' program, using different values of block penalty (bpen = 5, 15, 25). Since the quality of recombination landscapes was similar between different block penalties, despite a smoothing effect for higher values, we kept estimates with the smaller block penalty (bpen = 5). We ran the MCMC algorithm for 10,000,000 iterations, sampling every 5,000 iterations, with a burn-in of 1,000,000 first iterations. The convergence of MCMC sampling chains was checked with LDhat diagnostic plots. For species with an already published Marey map in Brazier and Glémin (2022), the visual congruence between the LD map and the Marey map was used as a qualitative validation of our estimates.

Hotspot detection

Recombination hotspots were inferred with LDhot using default settings [43]. We tested the significance of putative hotspots in 3 kb windows, with a step size of 1 kb along the genome. The background window was ± 50 kb around the hotspot centre. The hotspots inferred were then summarised using the LDhot summary program with a significance cutoff of 0.001 for calling a hotspot and 0.01 for merging adjacent hotspots. In order to control for hotspot quality and reduce false positive hotspots, we compared the raw hotspot dataset to filtered datasets after soft and hard filtering strategies. As hotspots are generally defined as narrow genomic regions, the soft filtering approach removed every hotspot larger than 10 kb. The intensity of a hotspot was estimated by dividing the recombination peak rate by the background recombination rate (± 50 kb around the hotspot centre). Low intensity hotspots were considered to be most likely false positive calls. In addition, we considered that extreme values of hotspot intensity were most probably artefactual due to high variance in the LDhat recombination map caused by assembly and/or snp calling errors. The hard filtering removed hotspots larger than 10 kb or with intensity lower than 4 or higher than 200.

Recombination rates and hotspots in genomic features

The average ρ/kb for each genomic feature was calculated as the weighted arithmetic mean of the ρ/kb estimates intersecting this genomic interval (R 'weighted.mean' function). The weights were the number of nucleotides spanned by each interval. The number of hotspots overlapping a given genomic feature was counted with 'countOverlap' and 'findOverlap' from the GenomicRanges R package [78]. The mean and the median of ρ/kb within genomic features and 95% confidence intervals were estimated by bootstrap (1,000 iterations).

Gradients of recombination

To assess a putative gradient of recombination at the 5' end of genes, we estimated LD-based recombination rates (ρ/kb) in consecutive intervals (200 bp) along a region $\pm$ 5 kb around the TSS, ATG codon (start of the coding sequence) and TTS. Annotations of start codons are generally more consistent and reliable than Transcription Starting Sites (TSS) among genomes that have been annotated using different methodologies. Intervals overlapping another gene were discarded. We also separated the average gradient from exon/intron-specific gradients. Each interval in a gene was marked as exon-specific (intron-specific, respectively) only if an exon (intron) overlapped at least 70% of the interval. A random control of ρ/kb was calculated by resampling the exact same intervals at a random position in the genome.

We also produced gradients of recombination by estimating recombination rates as a function of exon/intron ordinal rank within genes. We computed the mean, weighted mean and median ρ/kb for each exon (CDS part) and intron. The recombination rate of exons/introns was averaged by exon/intron ordinal rank for the average gradient, and by rank and gene size (number of exons) for detailed gradients. As annotation quality and sample size decline for longer genes, we removed all genes longer than 14 exons or 10 kb.

We evaluated the performance of our method to produce gradients by testing whether it had the power to detect a true gradient of recombination or the capacity to reject an artifactual gradient. Firstly, we assessed whether the gradient observed could be an artefact. LDhat recombination rates are estimated within intervals between adjacent SNPs. As a consequence, a lack of SNPs at the beginning of genes where recombination hotspots are supposed to be located could reduce our resolution to accurately position hotspots in this region, thus leading to an artificial leakage of high ρ/kb values at a larger scale than the true biological size of the hotspot. To evaluate this, we re-calculated the gradient of ρ/kb as a function of the rank, but this time we kept only ρ/kb estimates beginning in the genomic interval (exon/intron). Thus, we were sure to measure only the actual recombination within the interval. We observed the same patterns using this restricted set of ρ values confirming that the observed gradients were a true biological signal (Fig S13A,B). Secondly, we assessed the power to detect a true gradient, if there was one, for a given empirical dataset. For each gene, the ranks of exons were sorted by descending ρ/kb values, to simulate the maximal negative 5'-3' gradient that could be observed from empirical observations, if there is one. Ten out of eleven datasets had the necessary power to detect true gradients of recombination, except for *Triticum aestivum* lacking resolution (Fig S13B).

Acknowledgments

We wish to thank Laurent Duret, Pierre-Alexandre Gagnaire, Marie Raynaud, Julien Joseph, Nicolas Lartillot and all the other members of the HotRec ANR project, as well as members of the lab and people met outside for the discussion of the results. We used the GENOUEST computing facility for analyses. **Funding:** Agence Nationale de la Recherche, Grant ANR-19-CE12472 0019 / HotRec. **Author contributions:** Original idea: S.G.; Data analyses: T.B.; First draft: T.B.; Editing and revisions: T.B., S.G. **Competing interests:** The authors declare no conflicts of interest. **Data and materials availability:** Analysis scripts and documentation are available at **https://github.com/ThomasBrazier/landrec-gradients**. Data produced in this study are available at **https://osf.io/3aekw/**.

References

1. Brazier, T. & Glémin, S. Diversity and Determinants of Recombination Landscapes in Flowering Plants. *PLOS Genetics* **18** (ed Henderson, I. R.) e1010141 (2022).
2. Haenel, Q., Laurentino, T. G., Roesti, M. & Berner, D. Meta-Analysis of Chromosome-Scale Crossover Rate Variation in Eukaryotes and Its Significance to Evolutionary Genomics. *Molecular Ecology* **27,** 2477–2497 (2018).
3. Lloyd, A. Crossover Patterning in Plants. *Plant Reproduction* (2022).
4. Tock, A. J. & Henderson, I. R. Hotspots for Initiation of Meiotic Recombination. *Frontiers in Genetics* **9,** 521 (2018).
5. de Massy, B. Initiation of Meiotic Recombination: How and Where? Conservation and Specificities Among Eukaryotes. *Annual Review of Genetics* **47,** 563–599 (2013).
6. Kauppi, L., Jeffreys, A. J. & Keeney, S. Where the Crossovers Are: Recombination Distributions in Mammals. *Nature Reviews Genetics* **5,** 413–424 (2004).
7. Lam, I. & Keeney, S. Nonparadoxical Evolutionary Stability of the Recombination Initiation Landscape in Yeast. *Science* **350,** 932–937 (2015).
8. Mézard, C. Meiotic Recombination Hotspots in Plants. *Biochemical Society Transactions* **34,** 531–534 (2006).
9. Baudat, F. *et al.* PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327,** 836–840 (2010).
10. Grey, C., Baudat, F. & de Massy, B. PRDM9, a Driver of the Genetic Map. *PLOS Genetics* **14** (ed Cohen, P. E.) e1007479 (2018).
11. Boulton, A., Myers, R. S. & Redfield, R. J. The Hotspot Conversion Paradox and the Evolution of Meiotic Recombination. *Proceedings of the National Academy of Sciences* **94,** 8058–8063 (1997).
12. Latrille, T., Duret, L. & Lartillot, N. The Red Queen Model of Recombination Hot-Spot Evolution: A Theoretical Investigation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372,** 20160463 (2017).

13. Baker, Z. *et al.* Repeated Losses of PRDM9-directed Recombination despite the Conservation of PRDM9 across Vertebrates. *eLife* **6,** e24133 (2017).

14. Smagulova, F. *et al.* Genome-Wide Analysis Reveals Novel Molecular Features of Mouse Recombination Hotspots. *Nature* **472,** 375–378 (2011).

15. Auton, A. *et al.* Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genetics* **9** (ed Henderson, I.) e1003984 (2013).

16. Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D. & Petukhova, G. V. Genetic Recombination Is Directed Away from Functional Genomic Elements in Mice. *Nature* **485,** 642–645 (2012).

17. Choi, K. *et al.* Arabidopsis Meiotic Crossover Hot Spots Overlap with H2A.Z Nucleosomes at Gene Promoters. *Nature Genetics* **45,** 1327–1336 (2013).

18. Choi, K. & Henderson, I. R. Meiotic Recombination Hotspots - a Comparative View. *The Plant Journal* **83,** 52–61 (2015).

19. Li, X., Li, L. & Yan, J. Dissecting Meiotic Recombination Based on Tetrad Analysis by Single-Microspore Sequencing in Maize. *Nature Communications* **6,** 6648 (2015).

20. Myers, S. A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. *Science* **310,** 321–324 (2005).

21. Pan, J. *et al.* A Hierarchical Combination of Factors Shapes the Genome-wide Topography of Yeast Meiotic Recombination Initiation. *Cell* **144,** 719–731 (2011).

22. Singhal, S. *et al.* Stable Recombination Hotspots in Birds. *Science* **350,** 928–932 (2015).

23. Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N. & Levy, A. A. DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in *Arabidopsis*. *The Plant Cell* **27,** 2427–2436 (2015).

24. Dooner, H. K. & He, L. Polarized Gene Conversion at the Bz Locus of Maize. *Proceedings of the National Academy of Sciences* **111,** 13918–13923 (2014).

25. Hellsten, U. *et al.* Fine-Scale Variation in Meiotic Recombination in *Mimulus* Inferred from Population Shotgun Sequencing. *Proceedings of the National Academy of Sciences* **110,** 19478–19482 (2013).

26. McVean, G. A. T. The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. *Science* **304,** 581–584 (2004).

27. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread Genomic Signatures of Natural Selection in Hominid Evolution. *PLoS Genetics* **5** (ed Nachman, M. W.) e1000471 (2009).

28. Smeds, L., Mugal, C. F., Qvarnström, A. & Ellegren, H. High-Resolution Mapping of Crossover and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. *PLOS Genetics* **12** (ed Henderson, I.) e1006044 (2016).

29. Dooner, H. K. & Martínez-Férez, I. M. Recombination Occurs Uniformly within the Bronze Gene, a Meiotic Recombination Hotspot in the Maize Genome. *The Plant Cell* **9,** 1633–1646 (1997).

30. Fridman, E., Pleban, T. & Zamir, D. A Recombination Hotspot Delimits a Wild-Species Quantitative Trait Locus for Tomato Sugar Content to 484 Bp within an Invertase Gene. *Proceedings of the National Academy of Sciences* **97,** 4718–4723 (2000).

31. He, Y. *et al.* Genomic Features Shaping the Landscape of Meiotic Double-Strand-Break Hotspots in Maize. *Proceedings of the National Academy of Sciences* **114,** 12231–12236 (2017).

32. Marand, A. P. *et al.* Historical Meiotic Crossover Hotspots Fueled Patterns of Evolutionary Divergence in Rice. *The Plant Cell* **31,** 645–662 (2019).

33. Detloff, P., White, M. A. & Petes, T. D. Analysis of a Gene Conversion Gradient at the HIS4 Locus in *Saccharomyces Cerevisiae*. *Genetics* **132,** 113–123 (1992).

34. Malone, R. E., Bullard, S., Lundquist, S., Kim, S. & Tarkowski, T. A Meiotic Gene Conversion Gradient Opposite to the Direction of Transcription. *Nature* **359,** 154–155 (1992).

35. Schultes, N. P. & Szostak, J. W. Decreasing Gradients of Gene Conversion on Both Sides of the Initiation Site for Meiotic Recombination at the ARG4 Locus in Yeast. *Genetics* **126,** 813–822 (1990).

36. Clément, Y. *et al.* Evolutionary Forces Affecting Synonymous Variations in Plant Genomes. *PLOS Genetics* **13** (ed Copenhaver, G. P.) e1006799 (2017).

37. Glémin, S., Clément, Y., David, J. & Ressayre, A. GC Content Evolution in Coding Regions of Angiosperm Genomes: A Unifying Hypothesis. *Trends in Genetics* **30,** 263–270 (2014).

38. Ressayre, A. *et al.* Introns Structure Patterns of Variation in Nucleotide Composition in *Arabidopsis Thaliana* and Rice Protein-Coding Genes. *Genome Biology and Evolution* **7,** 2913–2928 (2015).

39. Serres-Giardi, L., Belkhir, K., David, J. & Glémin, S. Patterns and Evolution of Nucleotide Landscapes in Seed Plants. *The Plant Cell* **24,** 1379–1397 (2012).

40. Auton, A. *et al.* A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science* **336,** 193–198 (2012).

41. Marand, A. P. *et al.* Meiotic Crossovers Are Associated with Open Chromatin and Enriched with Stowaway Transposons in Potato. *Genome Biology* **18,** 203 (2017).

42. Auton, A. & McVean, G. Recombination Rate Estimation in the Presence of Hotspots. *Genome Research* **17,** 1219–1227 (2007).

43. Auton, A., Myers, S. & McVean, G. Identifying Recombination Hotspots Using Population Genetic Data. *arXiv:1403.4264 [q-bio]* (2014).

44. Stukenbrock, E. H. & Dutheil, J. Y. Fine-Scale Recombination Maps of Fungal Plant Pathogens Reveal Dynamic Recombination Landscapes and Intragenic Hotspots. *Genetics* **208,** 1209–1229 (2018).

45. Sudmant, P. H. *et al.* Global Diversity, Population Stratification, and Selection of Human Copy-Number Variation. *Science* **349,** aab3761 (2015).

46. Raynaud, M., Gagnaire, P.-A. & Galtier, N. Performance and Limitations of Linkage-Disequilibrium-Based Methods for Inferring the Genomic Landscape of Recombination and Detecting Hotspots: A Simulation Study. *Peer Community Journal* **3,** e27 (2023).

47. Kong, A. *et al.* A High-Resolution Recombination Map of the Human Genome. *Nature Genetics* **31,** 241–247 (2002).

48. Kong, A. *et al.* Rate of de Novo Mutations and the Importance of Father's Age to Disease Risk. *Nature* **488,** 471–475 (2012).

49. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at & 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis Thaliana. *Cell* **166,** 481–491 (2016).

50. Zhang, X. *et al.* Haplotype-Resolved Genome Assembly Provides Insights into Evolutionary History of the Tea Plant Camellia Sinensis. *Nature Genetics* **53,** 1250–1259 (2021).

51. Guo, S. *et al.* Resequencing of 414 Cultivated and Wild Watermelon Accessions Identifies Selection for Fruit Quality Traits. *Nature Genetics* **51,** 1616–1623 (2019).

52. Yang, C. *et al.* Resequencing 250 Soybean Accessions: New Insights into Genes Associated with Agronomic Traits and Genetic Networks. *Genomics, Proteomics & Bioinformatics,* S1672022921001601 (2021).

53. Sun, X. *et al.* Phased Diploid Genome Assemblies and Pan-Genomes Provide Insights into the Genetic History of Apple Domestication. *Nature Genetics* (2020).

54. Wang, W. *et al.* Genomic Variation in 3,010 Diverse Accessions of Asian Cultivated Rice. *Nature* **557,** 43–49 (2018).

55. Wu, J. *et al.* Resequencing of 683 Common Bean Genotypes Identifies Yield Component Trait Associations across a North–South Cline. *Nature Genetics* **52,** 118–125 (2020).

56. Liu, S. *et al.* Demographic History and Natural Selection Shape Patterns of Deleterious Mutation Load and Barriers to Introgression across *Populus* Genome. *Molecular Biology and Evolution* (ed Kim, Y.) msac008 (2022).

57. Lozano, R. *et al. Comparative Evolutionary Analysis and Prediction of Deleterious Mutation Patterns between Sorghum and Maize* Preprint (Genomics, 2019).

58. Cai, X. *et al.* Genomic Analyses Provide Insights into Spinach Domestication and the Genetic Basis of Agronomic Traits. *Nature Communications* **12,** 7246 (2021).

59. Zhou, Y. *et al.* Triticum Population Sequencing Provides Insights into Wheat Adaptation. *Nature Genetics* **52,** 1412–1422 (2020).

60. Marand, A. P. *et al.* Historical Meiotic Crossover Hotspots Fueled Patterns of Evolutionary Divergence in Rice. *The Plant Cell* **31,** 645–662 (2019).

61. Wijnker, E. *et al.* The Genomic Landscape of Meiotic Crossovers and Gene Conversions in *Arabidopsis Thaliana. eLife* **2,** e01426 (2013).

62. Jessop, L., Allers, T. & Lichten, M. Infrequent Co-conversion of Markers Flanking a Meiotic Recombination Initiation Site in *Saccharomyces Cerevisiae. Genetics* **169,** 1353–1367 (2005).

63. Kianian, P. M. A. *et al.* High-Resolution Crossover Mapping Reveals Similarities and Differences of Male and Female Recombination in Maize. *Nature Communications* **9,** 2370 (2018).

64. Song, Q.-X. *et al.* Genome-Wide Analysis of DNA Methylation in Soybean. *Molecular Plant* **6,** 1961–1974 (2013).

65. Yelina, N. E. *et al.* DNA Methylation Epigenetically Silences Crossover Hot Spots and Controls Chromosomal Domains of Meiotic Recombination in *Arabidopsis. Genes & Development* **29,** 2183–2202 (2015).

66. Vining, K. J. *et al.* Dynamic DNA Cytosine Methylation in the *Populus Trichocarpa* Genome: Tissue-Level Variation and Relationship to Gene Expression. *BMC Genomics* **13,** 27 (2012).

67. Blat, Y., Protacio, R. U., Hunter, N. & Kleckner, N. Physical and Functional Interactions among Basic Chromosome Organizational Features Govern Early Steps of Meiotic Chiasma Formation. *Cell* **111,** 791–802 (2002).

68. Cooper, T. J., Garcia, V. & Neale, M. J. Meiotic DSB Patterning: A Multifaceted Process. *Cell Cycle* **15,** 13–21 (2016).

69. Xu, X., Hsia, A. P., Zhang, L., Nikolau, B. J. & Schnable, P. S. Meiotic Recombination Break Points Resolve at High Rates at the 5' End of a Maize Coding Sequence. *The Plant Cell* **7,** 2151–2161 (1995).

70. Wang, Y. & Copenhaver, G. P. Meiotic Recombination: Mixing It Up in Plants. *Annual Review of Plant Biology* **69,** 577–609 (2018).

71. Schield, D. R. *et al.* Snake Recombination Landscapes Are Concentrated in Functional Regions despite PRDM9. *Molecular Biology and Evolution* **37** (ed Su, B.) 1272–1294 (2020).

72. Hoge, C. R. *et al. Patterns of Recombination in Snakes Reveal a Tug of War between PRDM9 and Promoter-like Features* Preprint (Evolutionary Biology, 2023).

73. Choi, K. & Henderson, I. R. Meiotic Recombination Hotspots - a Comparative View. *The Plant Journal* **83,** 52–61 (2015).

74. Buard, J. & de Massy, B. Playing Hide and Seek with Mammalian Meiotic Crossover Hotspots. *Trends in Genetics* **23,** 301–309 (2007).

75. Dluzewska, J., Szymanska, M. & Ziolkowski, P. A. Where to Cross Over? Defining Crossover Sites in Plants. *Frontiers in Genetics* **9,** 609 (2018).

76. Duret, L. Why Do Genes Have Introns? Recombination Might Add a New Piece to the Puzzle. *Trends in Genetics* **17,** 172–175 (2001).

77. R Core Team. *R: A Language and Environment for Statistical Computing* Manual. R Foundation for Statistical Computing (Vienna, Austria, 2022).

78. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology* **9** (ed Prlic, A.) e1003118 (2013).

79. Dapper, A. L. & Payseur, B. A. Effects of Demographic History on the Detection of Recombination Hotspots from Linkage Disequilibrium. *Molecular Biology and Evolution* **35,** 335–353 (2018).

80. Samuk, K. *et al.* Gene Flow and Selection Interact to Promote Adaptive Divergence in Regions of Low Recombination. *Molecular Ecology* **26,** 4378–4390 (2017).

81. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics* **197,** 573–589 (2014).

82. O'Connell, J. *et al.* A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics* **10** (ed Gibson, G.) e1004234 (2014).

83. Danecek, P. *et al.* The Variant Call Format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

675 ## Supplementary figures

Supplementary PDF

Figure S1: Chromosome recombination maps. (A) Fine-scale recombination maps estimated with LDhat. Mean ρ/kb for each inter-SNP interval. (B) Broad-scale recombination maps (cM/Mb, 100 kb windows) estimated from Marey maps in [1]. The grey ribbon is the 95% confidence interval estimated by 1,000 bootstraps on SNP markers.

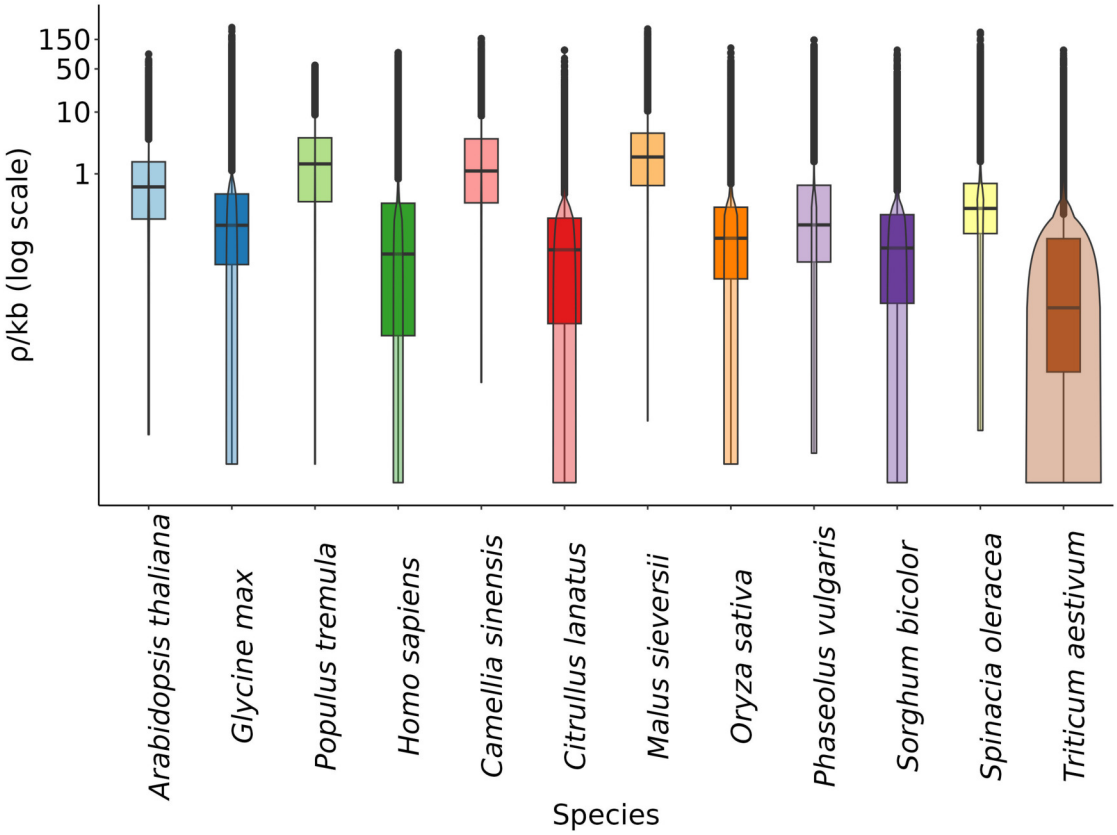

Figure S2: Species distributions of LDhat estimates (ρ/kb, log10 scale).

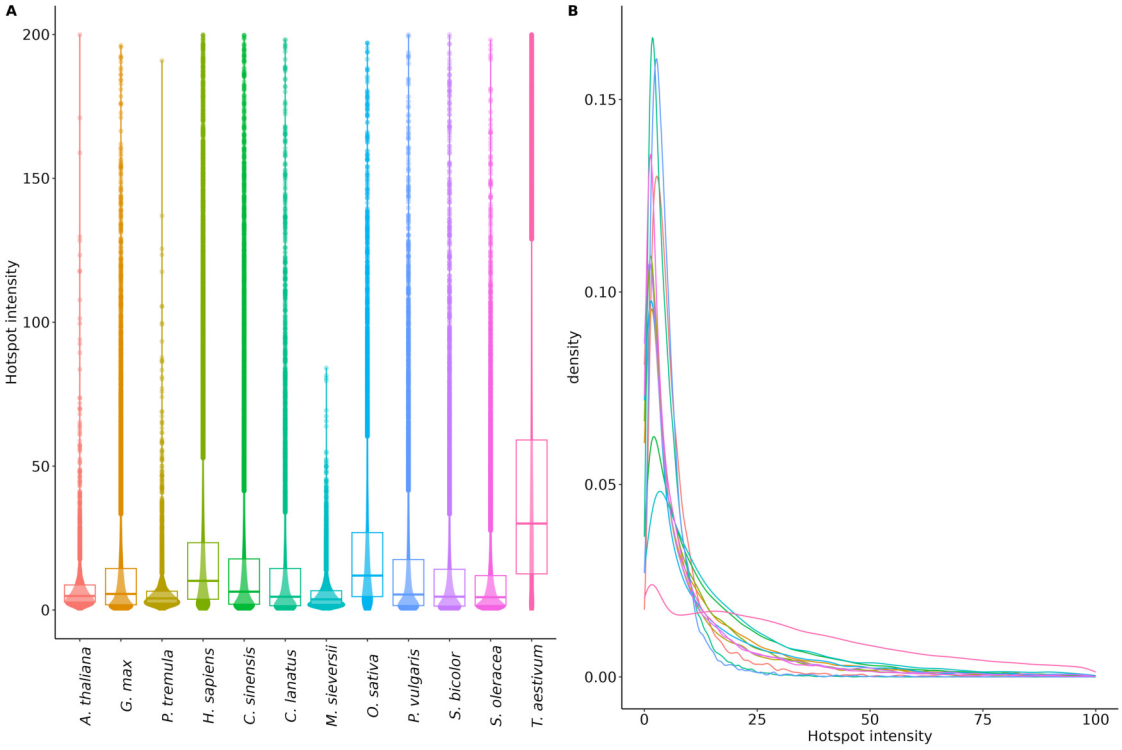

Figure S3: Hotspot intensity across species. (A) Distributions of hotspot intensity in the unfiltered dataset inferred by LDhot. The upper range is limited to 200 ρ/kb and higher values are not shown for clarity. (B) Density of hotspot intensity in the unfiltered LDhot dataset.

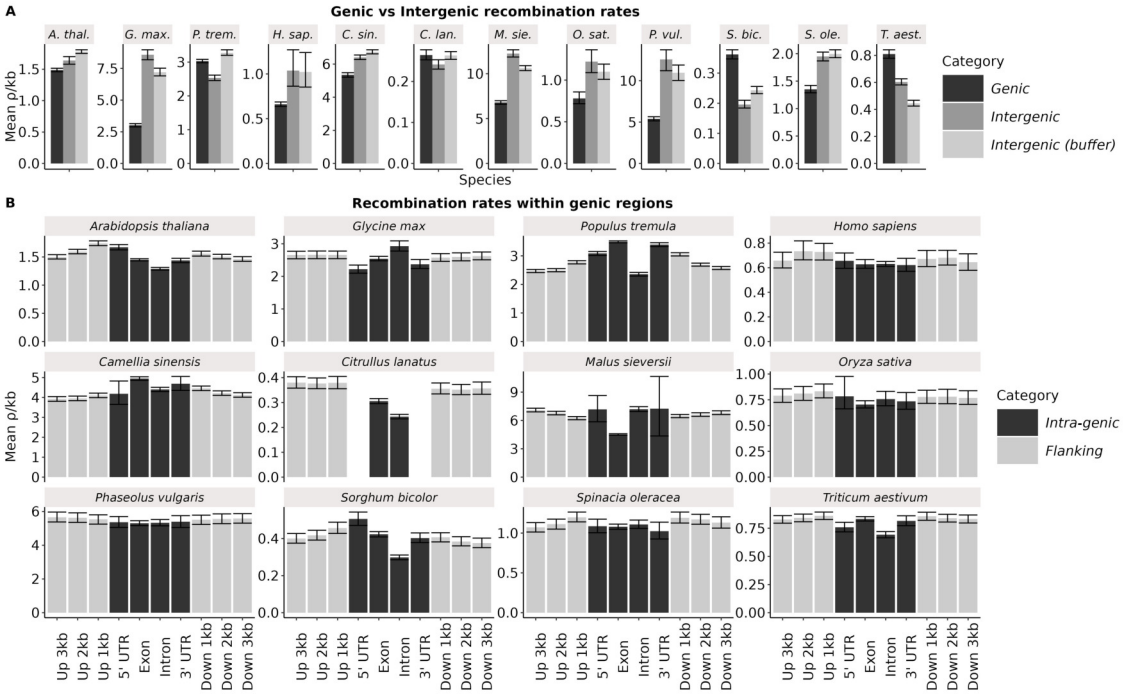

Figure S4: Variations in mean recombination rates (ρ/kb) around and within genes. (A) Recombination rates vary between genic and intergenic regions. In order to remove an effect of 5' and 3' regulatory regions at the proximity of genes, buffered intergenic regions were defined by excluding 3 kb flanking regions upstream and downstream of genes. (B) Within genic regions, the recombination rate varies between genomic features as well as in upstream and downstream flanking regions. The mean recombination rate and 95% confidence intervals were estimated by 1,000 bootstraps. Annotations of UTRs were not available for *Citrullus lanatus*.

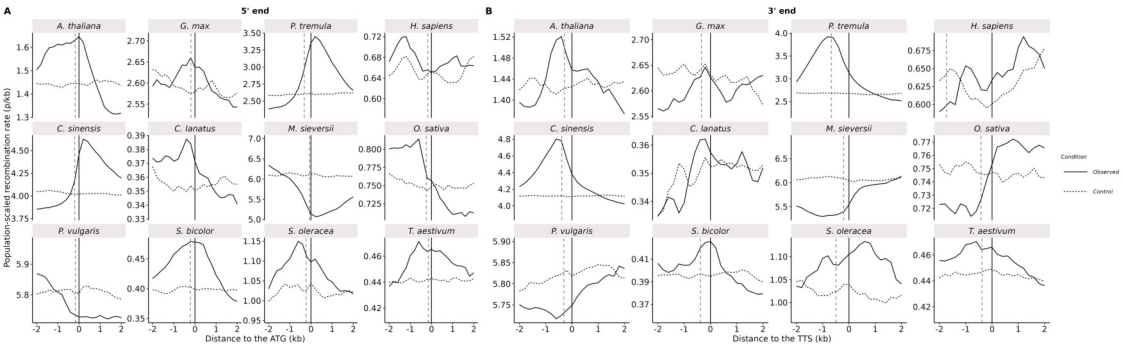

Figure S5: Recombination gradients along genes are mostly around Transcription Starting Sites (TSS) and Transcription Terminating Sites (TTS) among twelve species (including eleven plants). (A, B) The recombination rate (ρ/kb) was estimated in 200 bp windows as a function of the distance to the ATG start codon (start of first CDS) or the TTS and averaged among introns and exons. The vertical dashed black line is the mean distance to the TSS (i.e. the 5' UTR) and the vertical dashed grey line is the mean 3' UTR length. Random controls by randomly resampling intervals in the genome and re-scaled to the mean of the gradient.

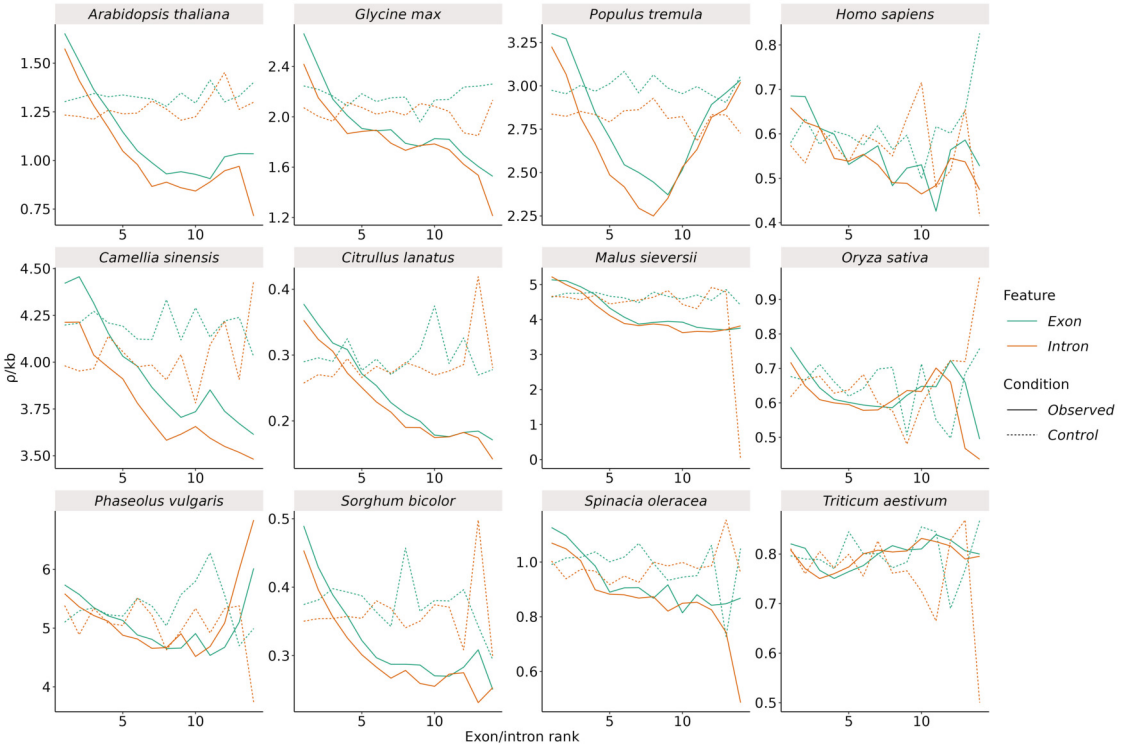

Figure S6: Genic recombination gradients of ρ/kb decomposed among introns and exons. The recombination rate (ρ/kb) was estimated in exons/introns as a function of exon/intron rank. Random controls by randomly resampling exon/introns intervals in the genome.

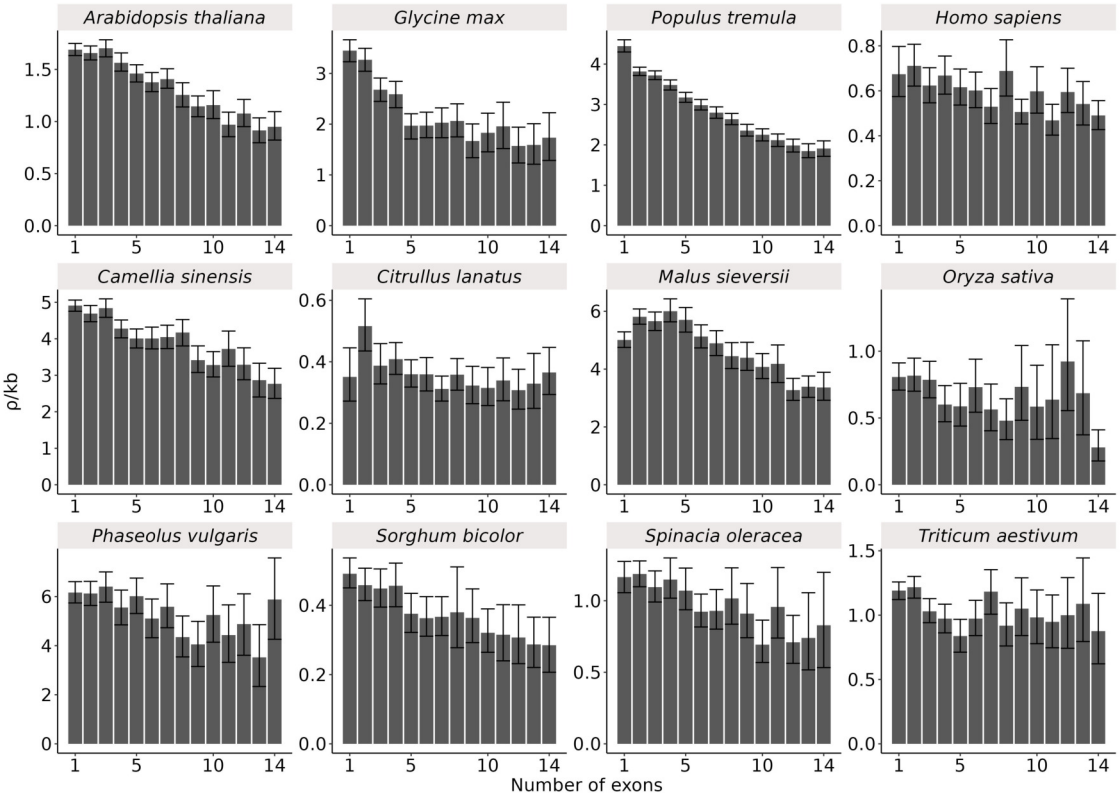

Figure S7: Weighted mean recombination rate (ρ/kb, weighted by the interval genomic size) as a function of gene size (i.e. number of exons). The 95 % confidence intervals were estimated by 1,000 bootstraps.

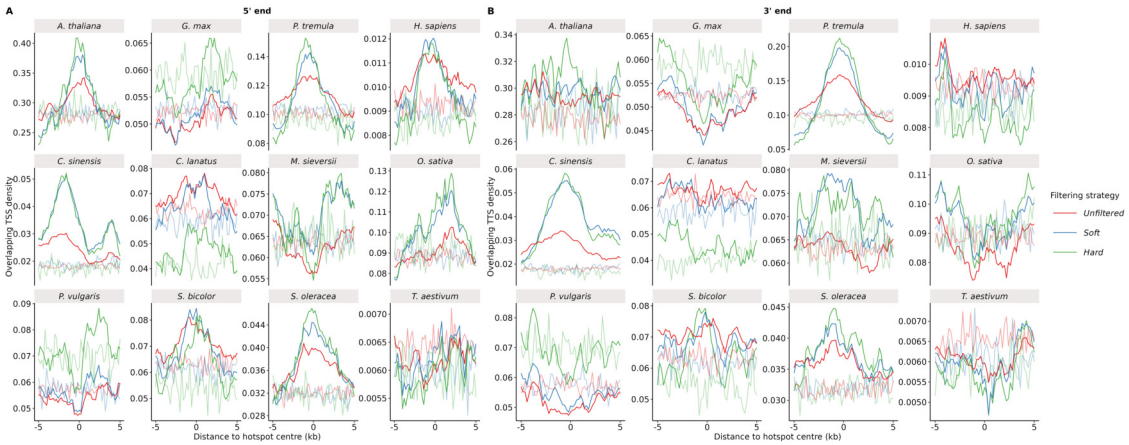

Figure S8: The density of TSS/TTS overlap around hotspot centre, expressed as a function of the distance to hotspot centre, for the three filtering strategies. For smoothing, TSS and TTS were defined as regions of 1 kb centered on the gene start and end position respectively. Soft lines are a control by randomly resampling hotspot intervals in a ± 50 kb neighboring region.

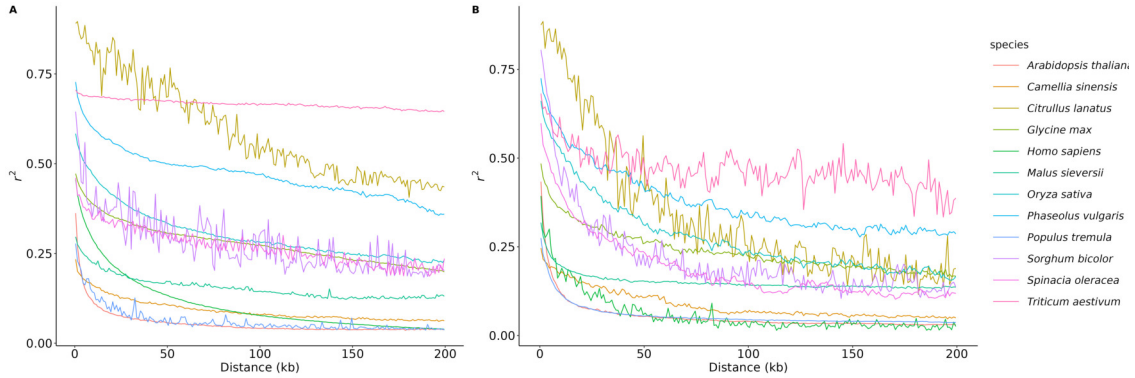

Figure S9: LD decay (r^2) as a function of the pairwise distance between markers. (A) r^2 measured genome wide. (B) r^2 measured between SNPs in genic regions.

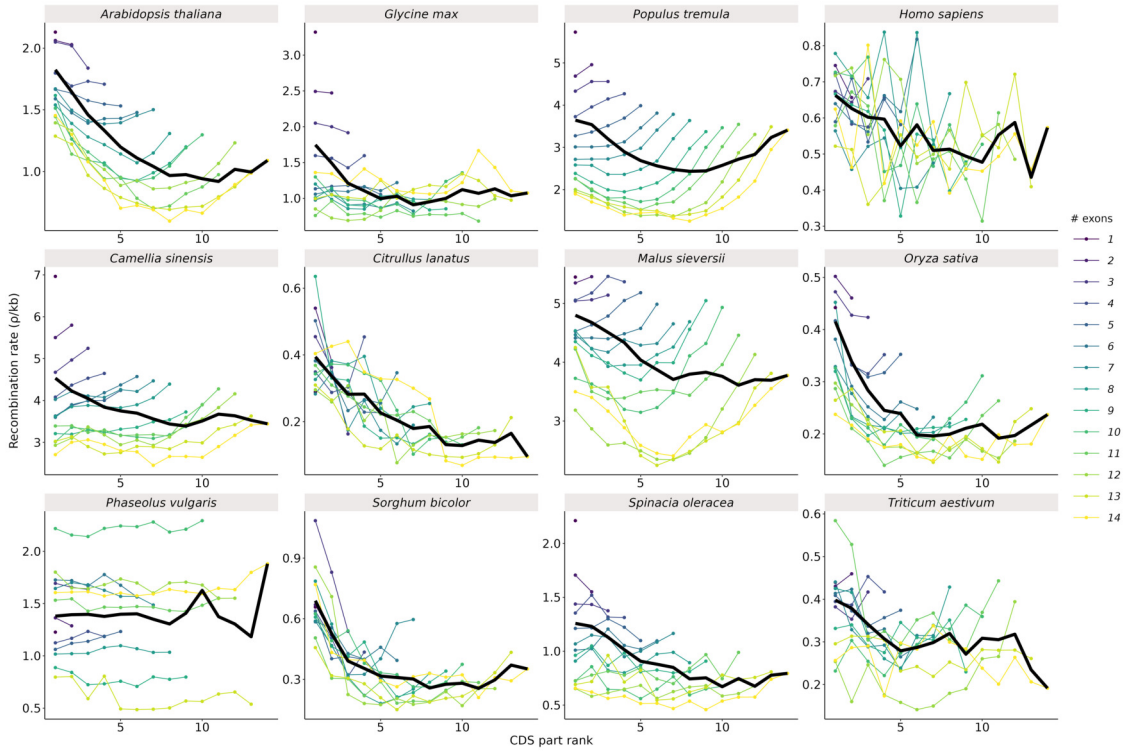

Figure S10: Genic recombination gradients of ρ/kb for genes overlapping at least five SNPs. The recombination rate (ρ/kb) was estimated in exons (CDS part) as a function of their rank and genes were grouped by their number of exons. The black line is the averaged gradient (all genes pooled).

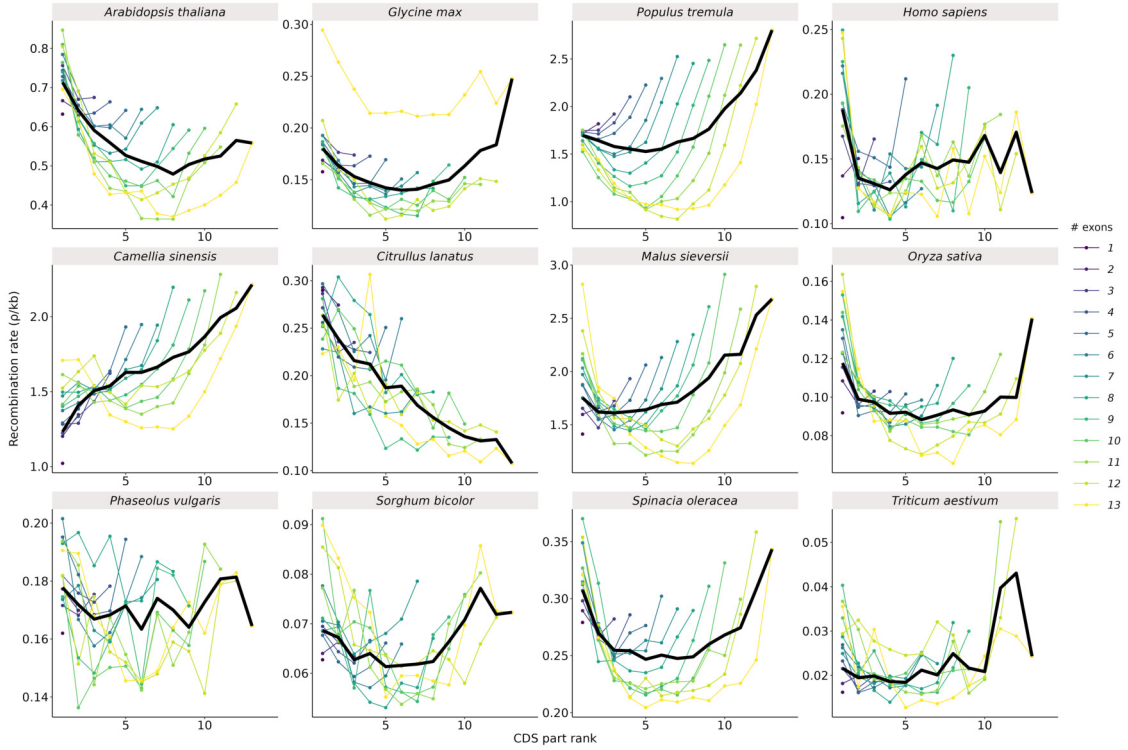

Figure S11: Genic recombination gradients of ρ/kb for the 75% of genes with the lower recombination rate. The recombination rate (ρ/kb) was estimated in exons (CDS part) as a function of their rank and genes were grouped by their number of exons. The black line is the averaged gradient (all genes pooled).

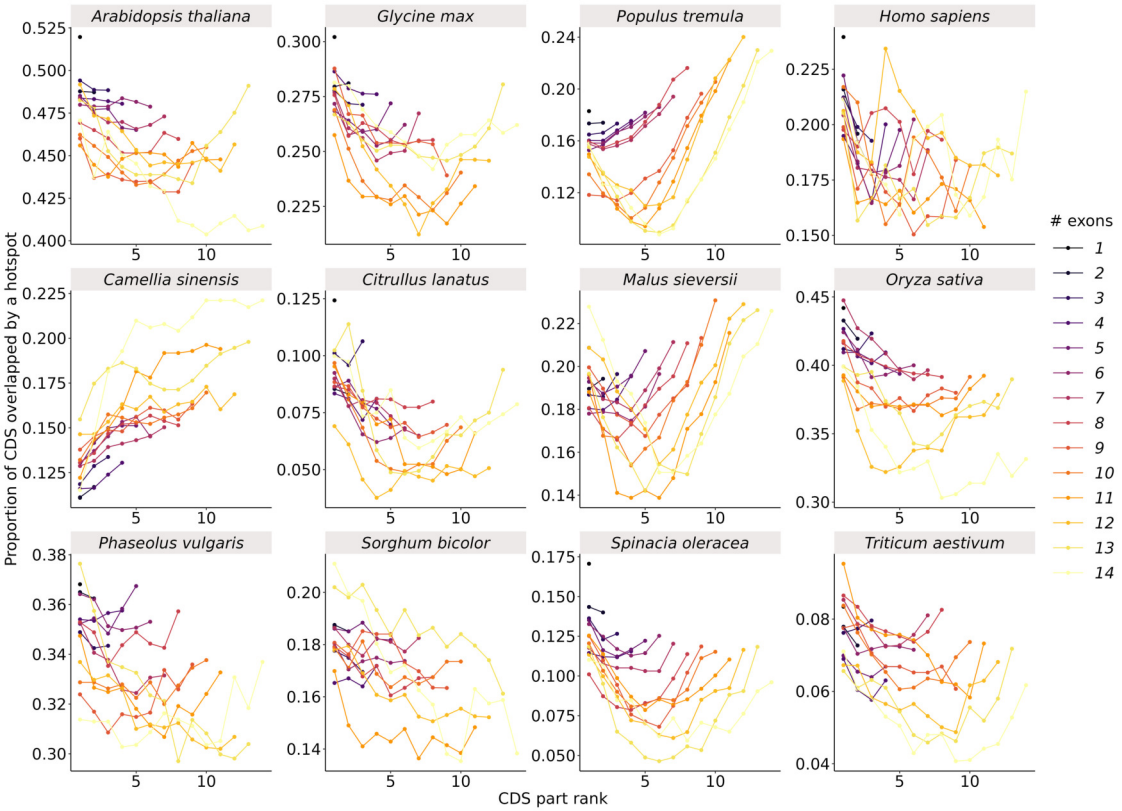

Figure S12: Gradients of hotspot overlap as a function of exon rank (CDS part). The proportion of CDS overlapping a hotspot in a given class of exon is the number of CDS overlapping at least one hotspot (soft filtered hotspots) divided by the total number of CDS of that class.

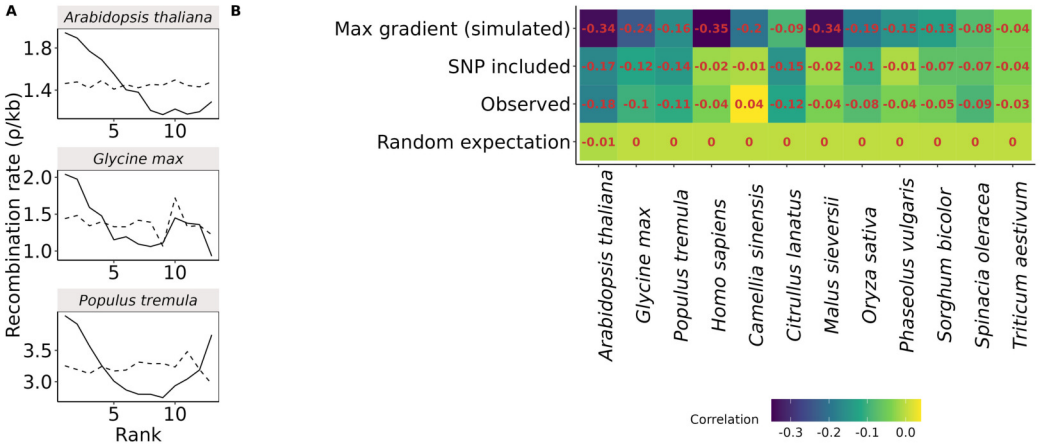

Figure S13: Evaluation of the performance of the method. (A) Strict gradients were estimated only with ρ/kb actually measured within the genomic interval. Random controls by resampling ranks (dashed lines). (B) Spearman's correlation between ρ/kb and exon rank. From bottom to top, the random expectation was computed by resampling ranks with replacement; correlation in actual LD estimates (Observed); Strict gradient with only ρ/kb actually measured within the given exon (SNP included); a maximal 'ideal' gradient was simulated by reordering ranks by descending ρ/kb (Max gradient).

676 ## Supplementary tables

Table S1: Metadata of the twelve datasets. Dataset name used during analyses, author of the original study, year of the original study, mating system, number of SNP before filtering, number of SNPs after filtering, total number of individuals in the dataset, number of individuals sampled for LDhat analyses, accession of the reference genome, number of chromosomes, number of genes, public database where to access original data (if relevant), direct link to the original data (if relevant), doi of the original study.

Table S2: Total number of genic vs intergenic hotspots for the three filtering strategies. Hotspots were considered as genic (intergenic, respectively) if they overlap (don't overlap) a gene interval. The expected number of genic hotspots ($\pm$ 95% confidence interval) was calculated by randomly resampling genic intervals.