Performance and limitations of linkage-disequilibrium-based methods for inferring the genomic landscape of recombination and detecting hotspots: a simulation study

Marie Raynaud[1,*], Pierre-Alexandre Gagnaire[1], Nicolas Galtier[1]
[1]ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France
*To whom correspondence should be addressed: marie.raynaud@umontpellier.fr

Abstract

Knowledge of recombination rate variation along the genome provides important insights into genome and phenotypic evolution. Population genomic approaches offer an attractive way to infer the population-scaled recombination rate $\rho=4N_e r$ using the linkage disequilibrium information contained in DNA sequence polymorphism data. Such methods have been used on a broad range of plant and animal species to build genome-wide recombination maps. However, the reliability of these inferences has only been assessed under a restrictive set of conditions. Here, we evaluated the ability of one of the most widely used coalescent-based approaches, *LDhelmet*, to infer a biologically-realistic genomic landscape of recombination. Using simulations, we specifically assessed the impact of empirical (sample size, block penalty) and evolutionary parameters (effective population size (N_e), mutation and recombination rates) on inferred map quality. We report reasonably good correlations between simulated and inferred landscapes, but point to limitations when it comes to detecting recombination hotspots. False positives and false negatives considerably confound fine-scale patterns of inferred recombination under a wide array of conditions, particularly when N_e is small and the mutation/recombination rate ratio is low, to the extent that maps inferred from populations sharing the same recombination landscape appear uncorrelated. We thus address a message of caution to users of such approaches, while also recognizing their importance and potential, particularly in species with less complex landscapes for which LD-based approaches should provide high quality recombination maps.

Key words

Population-scaled recombination rate, LDhelmet, simulations, linkage disequilibrium, recombination landscapes, recombination hotspots

Introduction

Recombination is highly conserved among sexually reproducing species of eukaryotes. This fundamental mechanism of meiosis is essential for the proper segregation of homologous chromosomes during the reductional division. Recombination involves crossing over events (CO) that play a crucial evolutionary role by allowing genetic mixing and generating new combinations of alleles (Baudat and de Massy 2007; Cromie et al. 2001; Capilla et al. 2016). Measuring the rate at which recombination occurs and the magnitude of its variation along the genome has important implications for fundamental research in molecular biology and evolution, but also for applied genomics such as genome-wide association studies (GWAS) (Morris et al. 2013; Hunter et al. 2016). Several approaches have been developed to reconstruct genome-wide recombination maps (reviewed in Peñalba and Wolf 2020). Cytological methods, like ChIP-seq, target protein-DNA complexes directly involved in the formation of double-strand breaks (DSB) and CO during meiosis (Pratto et al. 2014). Gamete typing methods analyse the meiotic products of a diploid individual (reviewed in Carrington and Cullen 2004; Dréau et al. 2019; Sun et al. 2019). Methods based on pedigree analysis reconstruct the gametic phase from patterns of allele inheritance in bi-parental crosses (Lander and Green 1987; Kong et al. 2002; Kodama et al. 2014; Rastas 2017). All these approaches have the advantage of providing direct estimates of the recombination rate. However, by focusing on CO that occurred in a few individuals or families across one or a couple of generations, they remain intrinsically limited in resolution due to the small number of recombination events that occur per chromosome per generation (Clark et al. 2010; Peñalba and Wolf, 2020).

Another type of approach uses genome sequence data from natural samples to take advantage of the large number of recombination events that have occurred during the history of the considered species/population. Instead of directly observing crossover products, these methods detect the footprints left by historical recombination events on patterns of haplotype segregation and linkage disequilibrium (LD) (reviewed in Stumpf and McVean 2003). The recombination rate and its variation across the genome are inferred via coalescent-based analysis of DNA sequence polymorphism data (Chan et al. 2012; Kamm et al. 2016; Li and Stephens 2003; McVean et al. 2004; Spence and Song 2019). The resulting LD maps have been widely used to evaluate the genomic impact of natural selection and admixture, and to perform genome-wide association studies (GWAS) (*e.g.* Chan et al. 2012; The International HapMap Consortium 2007). These approaches provide an accessible and attractive way of describing recombination landscapes - *i.e.* the variation of recombination rates along the genome - particularly in non-model taxa where direct methods are often difficult to implement

(Auton et al. 2012, 2013; Melamed-Bessudo et al. 2016; Shanfelter et al. 2019; Singhal et al. 2015; Shield et al. 2020).

Direct and indirect methods have revealed considerable variation in recombination rate at different scales along the genome, particularly in vertebrates. At a large scale (of the megabase order), recombination tends to be concentrated in subtelomeric regions compared to centromeric and centro-chromosomal regions, a pattern shared among many species of plants and animals (Auton et al. 2012; Capilla et al. 2016; Danguy des Déserts et al. 2021; Haenel et al. 2018; Melamed-Bessudo et al. 2016). At a finer scale (of the kilobase order), recombination events often cluster in small regions of about 2 kb, called recombination hotspots (Choi and Henderson, 2015; Kim et al. 2007; Mancera et al. 2008; Myers et al. 2005; Shanfelter et al. 2019; Singhal et al. 2015; Schield et al. 2020). Two distinct regulatory systems of recombination hotspot location have been described to date, with major implications on the evolutionary dynamics of recombination landscapes. In passerine birds (Singhal et al. 2015), dogs (Auton et al. 2013; Axelsson et al. 2012) and some teleost fishes (Shanfelter et al. 2019; Baker et al. 2017), recombination hotspots tend to be found in CpG-islands / promoter-like regions, and are highly conserved between closely-related species (Singhal et al. 2015). In contrast, in humans (Myers et al. 2005, 2010), apes (Auton et al. 2012 ; Great Ape Genome Project 2016) and mice (Booker et al. 2017), hotspot location is directed by the PRDM9 protein, which binds specific DNA motifs and triggers the formation of DSBs (Baudat et al. 2010; Grey et al. 2018; Simon Myers et al. 2010; Oliver et al. 2009; Parvanov et al. 2010). In these taxa, hotspots are mostly located away from genes (Auton et al. 2012; Baker et al. 2017), and show little or no conservation between closely related species (Myers et al. 2005, 2010; Auton et al. 2012; Booker et al. 2017) due to self-destruction by gene conversion and rapid turnover of PRDM9 alleles (Coop and Myers 2007; Latrille et al. 2017; Lesecque et al. 2014).

Population-based inference methods aim to infer the population recombination rate $\rho = 4N_e r$, r being the per meiosis, per bp recombination rate and N_e the effective population size (Stumpf and McVean, 2003). The ρ parameter reflects the density of population recombination events that segregate in polymorphism data, integrated across time and lineages. Several programs have been developed for reconstructing LD-maps (reviewed in Peñalba and Wolf, 2020); PHASE: Li and Stephens, 2003; LDhat: McVean et al. 2004; LDhelmet: Chan et al. 2012, LDpop: Kamm et al. 2016; pyrho: Spence and Song, 2019), which use coalescent theory based on the ancestral recombination graph (Griffiths et al. 1997; Arenas, 2013) to model and explore the genealogies of the distinct genomic segments. The most popular family of LD-based methods like LDhat (McVean et al. 2004) and its improved version LDhelmet (Chan et al. 2012), implement a pairwise composite likelihood method under a Bayesian framework

using a reversible jump Markov Chain Monte Carlo (rjMCMC) algorithm. They have been used for building fine scale LD-based maps in a broad range of animal taxa including humans (McVean et al. 2004), dogs (Axselsson et al. 2012 ; Auton et al. 2013), fruitfly (Chan et al. 2012), finches (Singhal et al. 2015), bee (Wallberg et al. 2015), stickleback (Shanfelter et al. 2019), rattlesnakes (Schield et al. 2020). In some species, inferred LD-based maps have been validated by assessing their correlation with recombination maps obtained using conventional approaches, confirming their quality (Chan et al. 2012; McVean et al. 2004; Booker et al. 2017; Shanfelter et al. 2019; Singhal et al. 2015). However, as genetic and LD-based maps greatly differ in resolution (pedigree-based inference provide resolution of about 1 cM, while population-based methods can infer recombination events at the kilobase scale, (Peñalba and Wolf, 2020)), such comparisons do not provide qualitative information on the reliability of the detected hotspots. Moreover, the heterogeneity of the data analysed in these papers (in terms of taxonomy, genetic diversity, demography, sample size, and software parameters, among other things) makes it difficult to appreciate the performance and the possible weaknesses of LD-based methods. In fact, the reliability and conditions of application of LD-based methods are still poorly understood and need to be deeply characterised, considering the growing importance of these tools.

The power and sensitivity of LDhat and LDhelmet have been tested by simulations aiming to evaluate the influence of switch error in haplotype phasing (Booker et al. 2017; Singhal et al. 2015), the amount of polymorphism, and the intensity of recombination hotspots (Singhal et al. 2015). These studies simulated simple recombination landscapes assuming either homogeneous recombination rates or a few, well-defined hotspots contrasting with a low-recombination background (Auton & McVean, 2007; Booker et al. 2017; Chan et al. 2012; McVean et al. 2004; Shanfelter et al. 2019; Singhal et al. 2015; Schield et al. 2020). Real recombination landscapes, however, are usually more complex and involve a continuous distribution of recombination hotspot density and intensity across genomic regions. This complexity has not been taken into account so far in studies assessing the performance of LD-map reconstruction methods. We thus lack a comprehensive picture of the ability of these methods to properly recover the biological characteristics of real recombination landscapes. In particular, the proportion of the inferred recombination hotspots that are correct, and the proportion of true hotpots that are missed, have not yet been quantified under a biologically realistic scenario. These are crucial quantities to properly interpret and use reconstructed LD-maps in genomic research.

In this paper, we specifically assessed the performance of the LDhelmet program to detect hotspots while assuming a biologically realistic recombination landscape. We evaluated the influence of methodological parameters including sample size and block penalty, and

species-specific parameters such as the effective population size, mutation rate, and recombination rate. We also considered different definitions of a recombination hotspot relative to its background recombination rate, with the aim of improving the sensitivity of the analysis. We identified the conditions in which LD-based inferences can provide an accurate mapping of hotspots, and the parameters that negatively affect the sensitivity and specificity of their detection within biologically realistic recombination landscapes.

Results

Recombination landscape modelling

Five realistic, heterogeneous recombination landscapes (referred to as “underlying landscapes” in this paper) of 1Mb length were built using the human genome high resolution map of meiotic DSB from Pratto et al. (2014). In order to mimic both broad and fine scale variation in the recombination rate parameter *r,* (Supplementary Figure S1) the first and second half of each landscape were drawn from a gamma distribution of mean 1 cM/Mb and 3 cM/Mb, respectively, with parameters fitted from Pratto et al. (2014) (1-500,000bp: shape=rate=0.1328; 500,001pb-1Mb: shape=0.1598, rate=0.0532). Accordingly, the 5 recombination landscapes generated (Figure 1A, Supplementary Figure S1) showed broad scale differences in recombination peak intensity, with less elevated recombination peaks in the first half compared with the second half of each chromosome. At a fine scale, recombination was concentrated in numerous peaks resembling human recombination hotspots, with about 85% of the recombination concentrated in 15% of the genome. The map lengths in recombination units were about 0.02 Morgan (Supplementary Figure S2).

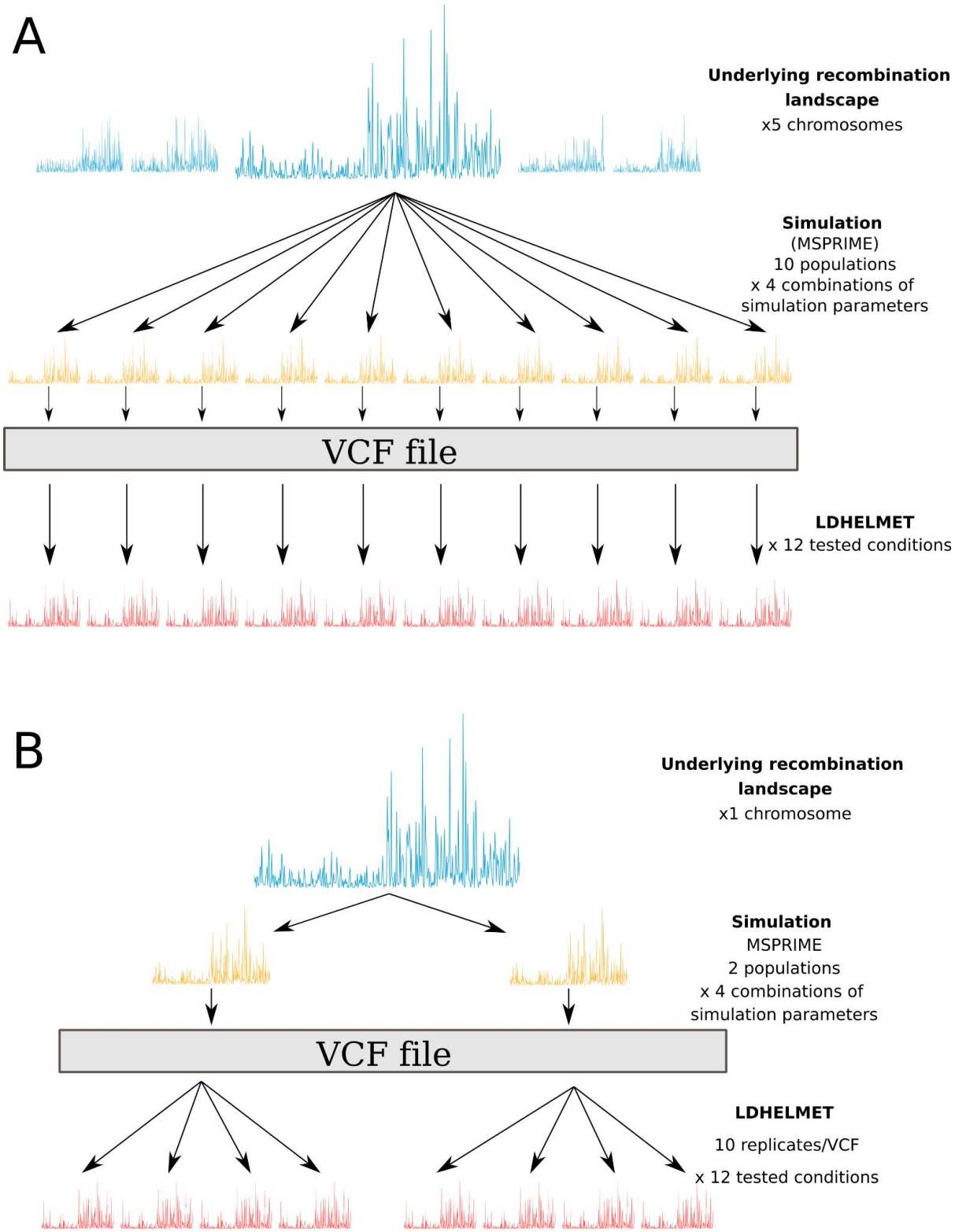


Figure 1. Simulation and inference protocols. **A)** In the first simulation framework, 5 different underlying recombination landscapes were generated based on human empirical data. These 5 landscapes can either be considered as parts of different chromosomes from the same species, or as orthologous parts of chromosomes from different species. For each of the 5 underlying landscapes, 10 recombination landscapes were simulated with MSPRIME for 4 combinations of effective population size (Ne) and sample size (SS) parameters, generating

a VCF file for each simulated population. The VCF files were then used to infer the local population recombination rates using LDhelmet with 3 alternative block penalty values (BP, a key parameter to LDhelmet), representing a total of 12 tested conditions. **B)** In the second simulation framework, only one of the 5 underlying landscapes was used to generate 2 simulated populations for each of the 4 combinations of simulation parameters, before running LDhelmet 10 times in replicate for 3 different values of block penalty, representing a total of 12 tested conditions.

Recombination landscapes of simulated populations (hereafter called "simulated landscapes") were generated in 10 replicates for each underlying landscape, using coalescent simulations with a mutation rate μ=10-8 and 4 combinations of sample sizes (SS=10 or 20) and effective population sizes (N_e=25,000 or 250,000) (Figure 1A). The simulated landscapes were a little shorter than the underlying landscapes (about 0.015-0.018 Morgan), reflecting the occasional occurrence of more than one recombination event between two adjacent SNPs during the simulated coalescent histories (Supplementary Figure S2). They were also highly correlated with the underlying landscapes for each combination of parameters (Spearman's rank correlation > 0.8 at 500bp resolution), showing that the stochasticity of the coalescent process has not resulted in a substantial loss of information about the underlying landscape. As expected from the $\theta=4N_e\mu$ values used in our simulations (θ = 0.001 and 0.01 for N_e = 25,000 and 250,000, respectively), the SNPs density of the large N_e populations was about one order of magnitude higher than for smaller N_e populations (Supplementary Figure S3).

Demographic, species-specific and methodological parameters affect LDhelmet performance

Population-scaled recombination rates (ρ) were inferred from the simulated polymorphism datasets using the program LDhelmet (Chan et al. 2012) (referred to as "inferred landscapes"). Three block penalty (BP) values (*i.e.* the degree of landscape resolution), which determine the number of allowed changes in ρ value within windows of 50 consecutive SNPs were tested (BP=5, 10, or 50) for each of the 4 SS-N_e combinations, resulting in 12 tested conditions (Figure 1A). Underlying and simulated landscapes were converted into population-scale recombination rates ($\rho=4N_er$), and each underlying, simulated and inferred maps were smoothed in 500bp (*i.e.* underlying landscape resolution) and 2,500bp windows (*i.e.* a resolution-level better suited to the SNP density in our low-N_e simulations). The 10 simulated and inferred replicates of each condition were averaged to perform comparisons between mean landscapes. Overall, local recombination rates tended to be overestimated by LDhelmet. but this was especially observed when the local ρ was either very

low (ρ < 10-4) or very high (ρ > 10-2) (Figure 2, Supplementary Figure S4). The mean inferred map lengths calculated across replicates varied substantially among tested conditions (0.017-0.125 M), reaching up to 6 times the length of simulated maps when N_e, SS and BP were low (*i.e.* N_e = 25,000, SS = 10, BP = 5, 10). In contrast, when N_e, SS and BP were high (N_e = 250,000, SS = 20, BP = 50) the inferred map lengths were closer to the expected value of 0.02 M (Supplementary Figure S2). A BP value of 50 produced very smooth recombination maps when N_e was low, which did not capture local-scale variation in recombination rate. By contrast, maps inferred with BP=5 or BP=10 were visually similar and better reflected the local-scale variation of the underlying landscapes (Supplementary Figure S5). Spearman correlation coefficient between the mean simulated and inferred landscapes was lower when N_e and SS were small, with a conspicuous effect of N_e on this measure (Figure 3A). Replicate runs of LDhelmet showed a strong consistency, as revealed by elevated correlations among the 10 replicate landscapes inferred from the same simulated landscape, whatever the condition being tested (Spearman's rho > 0.89, Figure 3B).

Recombination hotspots of the underlying, simulated and inferred landscapes were called using three different threshold values commonly used in the literature (*i.e.* local recombination rate at least 2.5, 5 or 10 times higher than the background rate). True/False positives/negatives rates and discovery rates (TPR, FPR, TDR, FDR, TNR, FNR) were computed under each tested condition. The hotspot detection threshold ratio of 10 between the focal and flanking recombination rates appeared too stringent and yielded a very small number of called hotpots (Supplementary Figure S6). Using a less conservative threshold ratio of 5, we detected 4 to 8 hotspots per Mb in the simulated landscapes, and 5 to 20 per Mb in the inferred landscapes. These numbers reached 40-50 and 20-50 per Mb, respectively, when a threshold of 2.5 was used. Irrespective of the chosen threshold, the number of inferred hotspots tended to be overestimated, especially when N_e was small, and to a lesser extent when SS was small (Supplementary Figure S6). The sensitivity (or TPR) of LDHelmet was medium to high, since depending on conditions, between 50% and 100% of the simulated hotpots were inferred as such. TPR was significantly higher for large N_e and SS, with a more pronounced effect when the detection threshold was set to 2.5 (Figure 4A, Supplementary Figure S7A). The specificity was reasonably high under the best conditions/settings, but dropped dramatically when N_e and/or SS were low, especially for a threshold ratio of 5. In the worst cases, 80% of the inferred hotspots corresponded to non-hotspot windows in the simulated maps (FDR, FPR, Figure 4B, Supplementary Figure S7B and C).

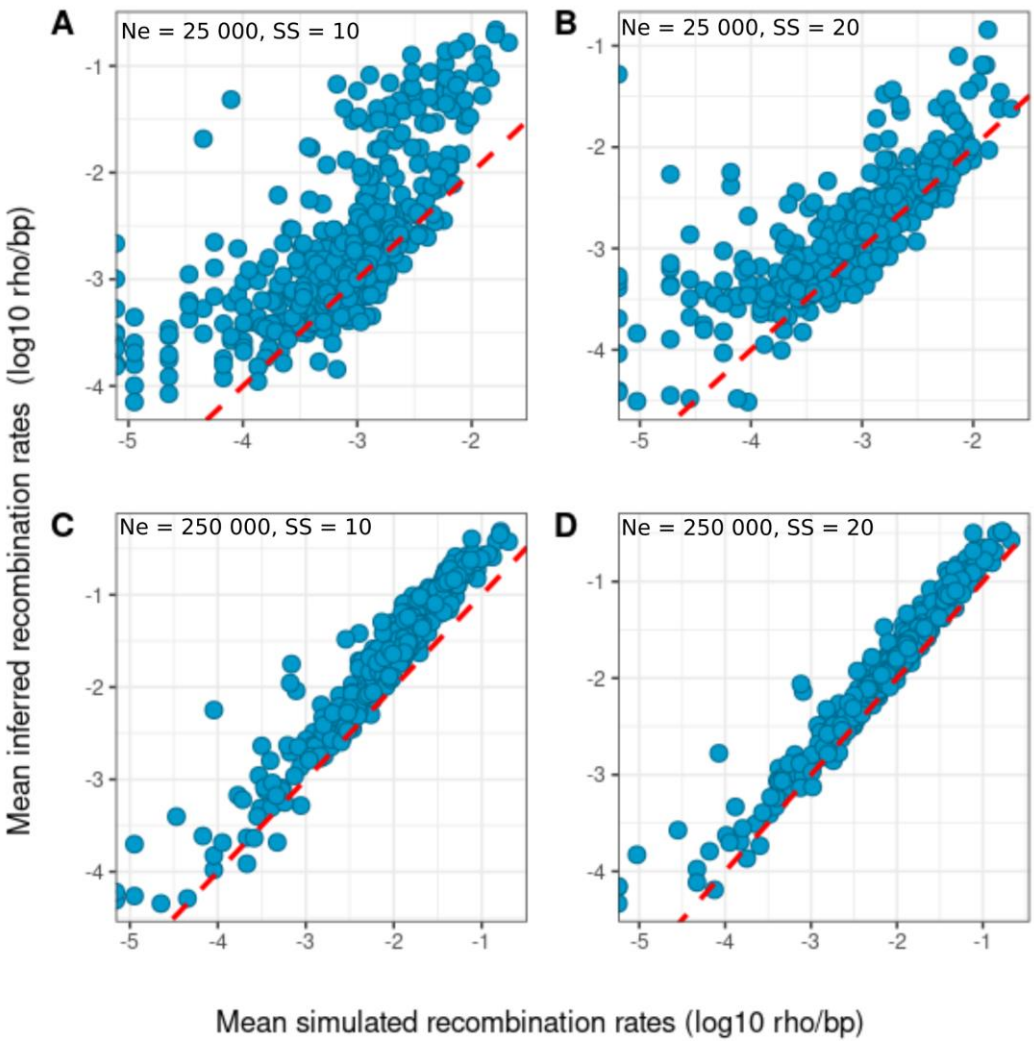


Figure 2. Quality assessment of local recombination rates estimated by LDhelmet, averaged within 2.5kb windows across 10 replicates. The x axis shows the recombination rates of the mean simulated landscapes and the y axis the recombination rates of the mean inferred landscapes, both on a logarithmic scale. Each blue point corresponds to a local 2.5kb-window average calculated across 10 replicate populations obtained under identical simulation parameters. **A)** N_e = 25,000, SS = 10, BP = 5. **B)** N_e = 25,000, SS = 20, BP = 5. **C)** N_e = 250,000, SS = 10, BP = 5. **D)** N_e = 250,000, SS = 20, BP = 5.

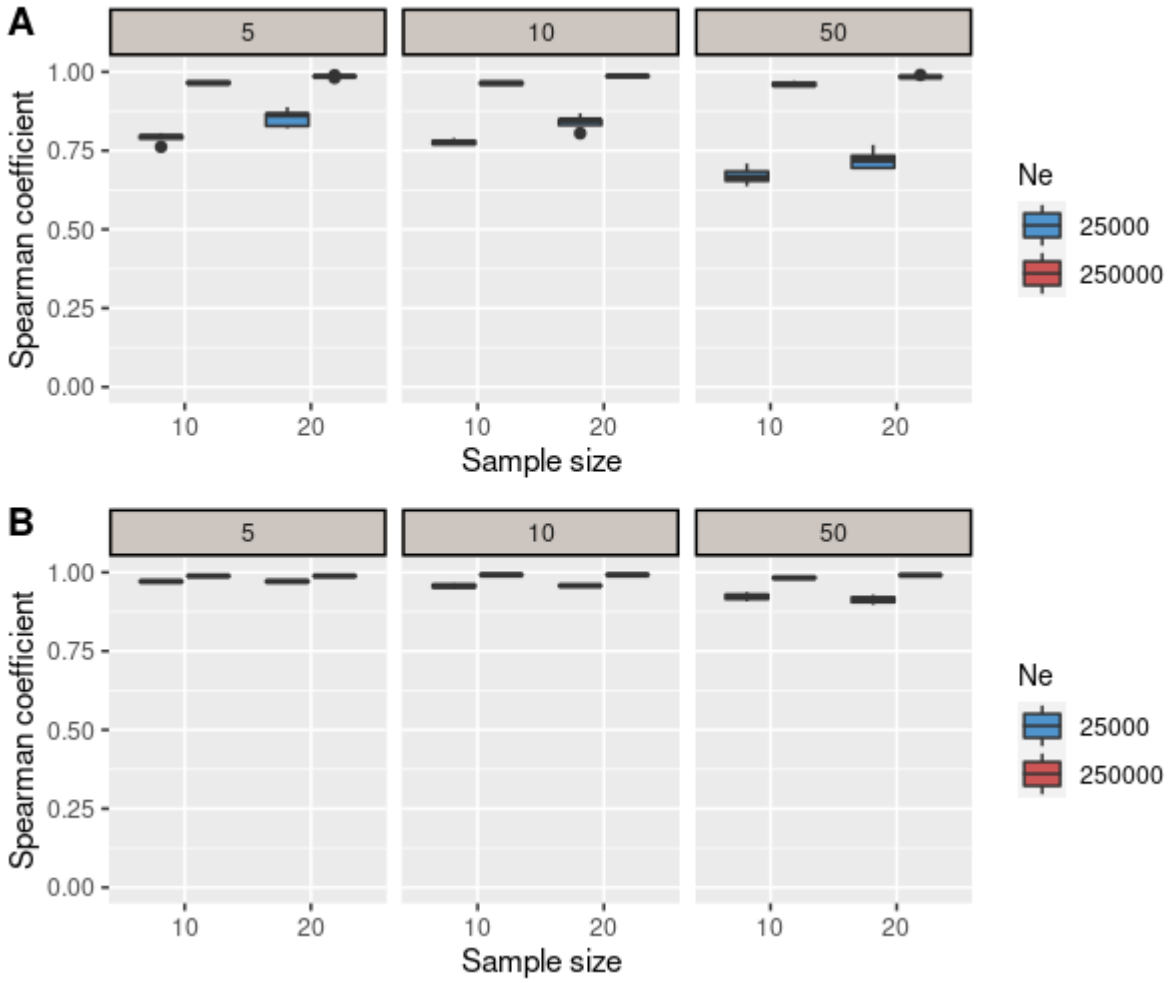


Figure 3. Correlations between simulated and inferred landscapes (A), and between replicates of inferred landscapes (B), for each of the 12 tested conditions. The sample size is shown on the x axis (*i.e.* SS=10 or 20), the N_e parameter is indicated by colour (*i.e.* 25,000 or 250,000) and the LDHelmet BP values correspond to the different panels (*i.e.* BP=5, 10 or 50). **A)** Spearman correlation coefficients between the mean simulated and the mean inferred landscape calculated across the 10 replicate populations originating from each of the five different underlying landscapes (*i.e.* using simulation framework of Figure 1A). **B)** Mean Spearman correlation coefficients calculated between pairwise comparisons among the ten replicates of inferred landscapes, from simulated populations sharing the same underlying landscape (*i.e.* using simulation framework of Figure 1B).

Control analysis

No significant difference in the correlation between simulated and inferred landscapes was found between the first half of the chromosome with a mean r of 1 cM/Mb (referred to as the “cold” region) and the last half with a mean *r* of 3 cM/Mb (the “hot” region). This was also true for the TPR and the FDR, whatever the hotspot detection threshold used (*i.e.* 2.5 or 5) (Student test, $p > 0.05$).

To account for the tendency of LDhelmet to overestimate large recombination rates,

we compared various hotspot detection thresholds applied to the inferred landscapes, for a given definition of what a true hotspot is in the simulated landscape. We aimed to evaluate how these relative definitions can improve the sensitivity (FPR) of hotspot detection while minimising the proportion of type I errors among called hotspots (FDR). When assuming that real hotspots had a 2.5 increased recombination rate compared to their flanking regions in the simulated landscape, the TPR was generally higher if a threshold of 5 was used for the inferred landscape instead of a threshold of 10 (Supplementary Figure S8A), but lower than when using the same threshold of 2.5. The FDR was in turn reduced when using thresholds of 5 and 10 as compared to 2.5 (Supplementary Figure S8B). We note, however, that the low FDR obtained for the 2.5/10 combination was due to a very low number of inferred hotspots when using a very conservative detection threshold. When the real hotspot threshold was assumed to be 5 in the simulated landscape, using a detection threshold of 10 did not improve the TPR and FDR (Student test, $p < 0.05$). Overall, these results suggest that, regardless of the actual relative recombination rate of real hotspots compared to their background (*e.g.*, 2.5- or 5-fold higher), the detection threshold of 5 that is often used in the literature offers a good compromise between sensitivity and FDR.

Differences between populations with different versus identical underlying recombination landscapes

As expected for a comparison between two populations with different underlying recombination landscapes, the mean linear correlation (R^2 coefficient) between inferred landscapes originating from different underlying landscapes was low, between 0.012 and 0.084, and similar to the R^2 between simulated landscapes (0.012-0.017) (Table 1, Supplementary Table S2). A low percentage of 1 to 10% of real hotspots were shared by chance between distinct underlying landscapes, depending on the hotspot definition ratio. Roughly similar proportions of shared hotspots were found between inferred landscapes originating from different underlying landscapes, although these proportions were slightly overestimated for low N_e, SS and BP (Table 1, Supplementary Table S2). A minority of the shared inferred hotspots were TP when $N_e = 25{,}000$ (10-40%). This proportion increased when N_e was large, while not always reaching very high values, depending on conditions (25-100%). Therefore, a non-zero fraction of truly shared hotspots is expected to be found between species with different biological recombination landscapes.

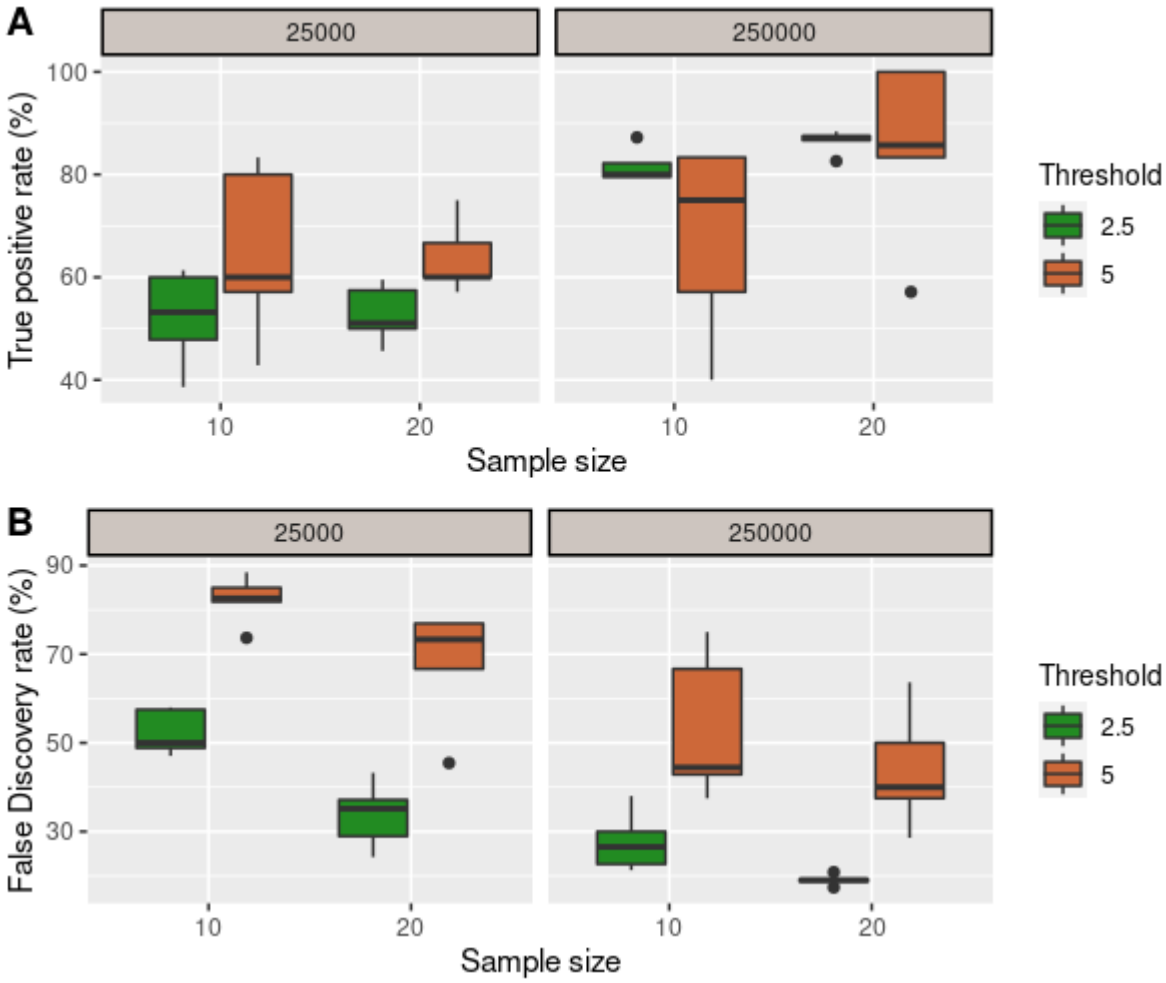


Figure 4. Hotspot detection from inferred landscapes for the four combinations of simulation parameters, and two hotspots detection thresholds as shown by colours (*i.e.* 2.5 and 5). The sample size parameter is shown on the x axis (*i.e.* 10 or 20), the left panel corresponds to conditions where N_e = 25,000, and the right panel to conditions where N_e = 250,000. Only conditions where BP = 5 are shown. **A)** True positive rate (sensitivity). **B)** False discovery rate.

Then we compared simulated populations originating from an identical underlying landscape to check the ability of LDhelmet to recover similar recombination rates between populations with independent coalescent histories. The correlations between simulated landscapes were generally high for both low (R^2>0.7) and large N_e (R^2>0.9) conditions, but the correlations between inferred landscapes were much lower, with R^2 < 0.3 and <0.75 for N_e = 25,000 and 250,000, respectively (Table 1). The proportion of shared hotspots followed the same trend: it was high between simulated landscapes (70 - 90%) and much lower between inferred landscapes - below 30% in the worst cases (Table 1). Thus, one can expect LDhelmet to detect a moderate to low fraction of shared hotpots even between species truly sharing a common recombination landscape, depending on population size, sample size and hotpot definition.

Table 1. Percentages of shared hotspots from simulated and inferred landscapes between populations with either different or identical underlying recombination landscapes. R^2 and mean proportion of shared hotspots are indicated for each type of comparison, including pairwise comparisons among simulated (left) or inferred (right) landscapes originating from either different underlying landscapes (top, following simulation framework from Figure 1A), or from the same underlying landscape (down, following simulation framework from Figure 1B). Only conditions where BP = 5 are shown.

		Mean proportion of shared simulated hotspots			Mean proportion of shared inferred hotspots	
	R^2 log(simuli~simulj)	Threshold = 5	Threshold = 2.5	R^2 log(inferi~inferj)	Threshold = 5	Threshold = 2.5
Different underlying landscapes						
Ne = 25000 SS = 10 BP = 5	0.012	1.715	8.065	0.033	5.385	13.99
Ne = 25000 SS = 20 BP = 5	0.014	2.00	9.045	0.032	3.95	9.735
Ne = 250000 SS = 10 BP = 5	0.017	2.085	8.42	0.013	0.00	11.245
Ne = 250000 SS = 20 BP = 5	0.017	1.715	7.675	0.014	1.17	9.55
Same underlying landscape						
Ne = 25000 SS = 10 BP = 5	0.7463	85.71	79.62	0.2139	30.65	28.55
Ne = 25000 SS = 20 BP = 5	0.7497	80.36	78.99	0.311	14.85	26.9
Ne = 250000 SS = 10 BP = 5	0.9466	80.36	90.42	0.6301	15.3	35.65
Ne = 250000 SS = 20 BP = 5	0.9554	80.36	93.79	0.7514	32.55	49.15

Influence of the μ/r ratio on hotspot detection

The influence of the mutation and recombination rates on the inference of recombination landscapes was assessed by generating two additional underlying landscapes but with a ten times higher (*i.e.* 20 cM/Mb) and lower (*i.e.* 0.2 cM/Mb) average recombination rate, and by producing coalescent simulations under varying mutation rates (*i.e.* 10-9, 10-8 and 10-7). The *μ*/*r* ratio of these 6 new conditions thus equalled 0.1, 1 or 10. For all conditions, Spearman's rank correlation between the mean simulated and the mean inferred landscapes was greater than 0.9, except when *μ* equalled 10-9 (Spearman's $\rho \simeq 0.7$, Table 2). The mean pairwise Spearman's correlation obtained among the 10 inferred replicates was $\simeq 0.8$, except when *μ* or *r* equalled 10-9 ($\simeq$0.5-0.6, Table 2). An increased *μ*/*r* ratio improved the detection of hotspots when *r* was fixed to 10-8, with a higher TPR (up to >80%) and a lower FDR (<5%) when *μ* increased (prop.test, p-value < 0.05 when the threshold is 2.5, Figure 5). The *μ*/*r* ratio did not affect the performances the same way when *μ* was fixed to 10-8: a *μ*/*r* ratio of 10 (*r* = 10-9) yielded lower TPR (< 60%) and higher FDR (> 25%) than a ratio of 1 or 0.1, although these trends were not significant (prop.test, p-value > 0.05, Figure 5).

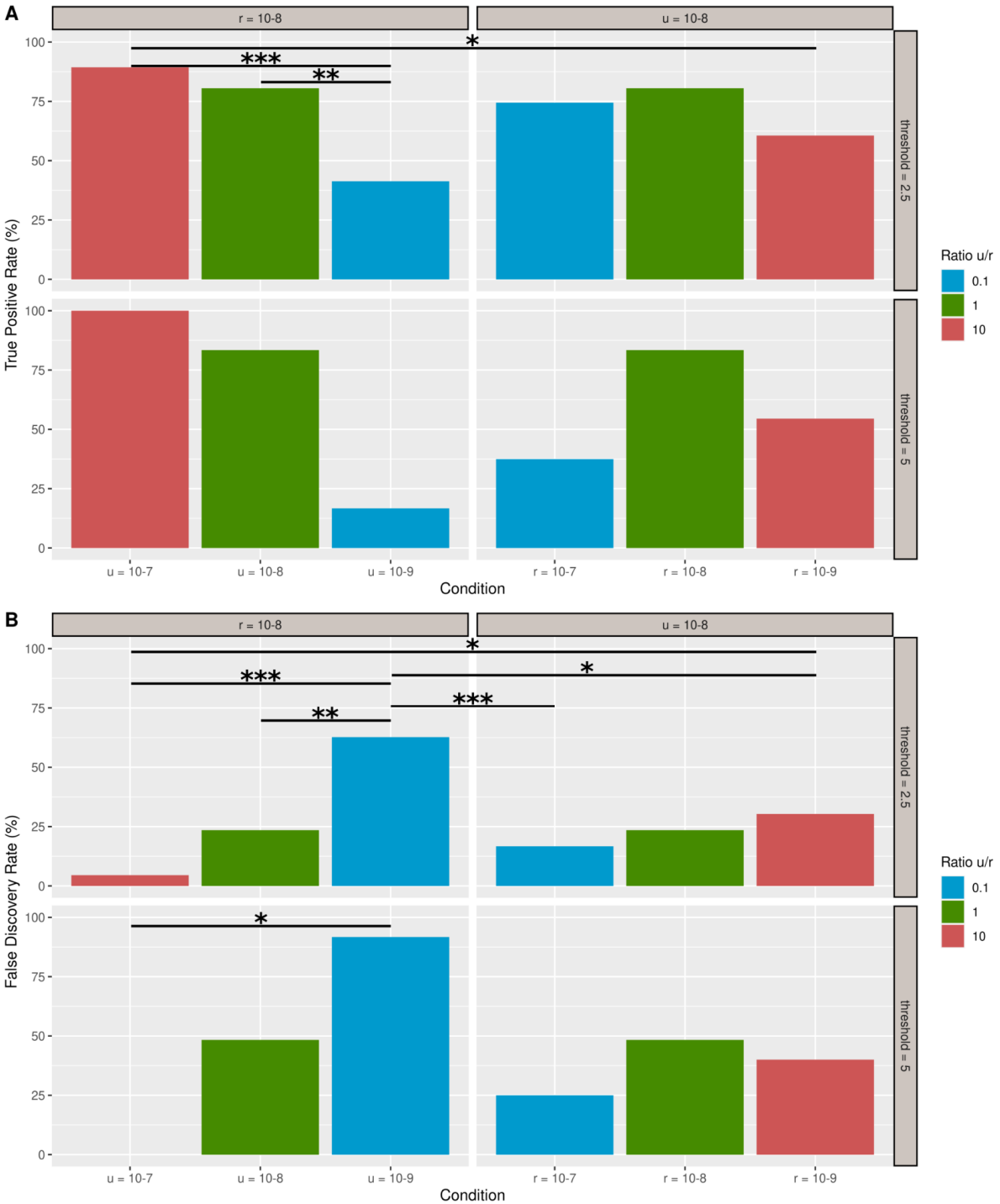


Figure 5. Influence of the *μ*/r ratio on hotspot detection. Colours correspond to different *μ*/r ratios, the left panels show conditions where *r* was fixed to 10-8, and right panels conditions where *μ* was fixed to 10-8. The horizontal panels correspond to the threshold used to detect hotspots. The x axis indicates the different *μ* or *r* values used in the simulations, and the y axis shows the true positive (**A**) and false discovery (**B**) rates. The asterisks show the significance level of the prop.test, with * indicating p-value<0.05, ** <0.01 and *** <0.001.

Table 2. Influence of the *μ*/*r* ratio on Spearman's correlation between mean simulated and mean inferred landscapes, as well as on mean pairwise Spearman's correlation between the 10 inferred replicates, and on hotspot detection metrics. The true positive (sensitivity) and false discovery rates are indicated for the three hotspot detection thresholds tested (*i.e.* 2.5, 5 and 10).

					Threshold = 2.5		Threshold = 5		Threshold = 10	
Ratio	r	u	Correlation simulated-inferred	Mean pairwise correlation	TPR	FDR	TPR	FDR	TPR	FDR
10	10-8	10-7	0.989	0.882	89.36	4.55	100.00	0.00	0	0
1	10-8	10-8	0.951	0.788	82.22	19.57	66.67	42.86	0	100
0.1	10-8	10-9	0.697	0.511	41.30	62.75	16.67	91.67	0	100
0.1	10-7	10-8	0.966	0.771	74.47	16.67	37.50	25.00	0	0
1	10-8	10-8	0.958	0.792	78.72	27.45	100.00	53.85	0	100
10	10-9	10-8	0.932	0.596	60.63	30.30	54.55	40.00	0	100

Discussion

Inferred LD-maps should be interpreted with caution

Inference methods based on linkage disequilibrium provide an attractive way to characterise genomic recombination landscapes from sequence data. As such, they promise to become increasingly popular in empirical studies of eukaryotes. However, their ability to accurately reproduce real recombination landscapes has not been specifically evaluated. It should be recalled, however, that LD-based recombination maps are inferences, not observations; biases and uncertainty must be quantified and taken into account when it comes to interpreting the results. Here, we modelled the biological characteristics of a particularly well-documented recombination landscape, that of humans, as a basis for assessing the impact of methodological and species-specific evolutionary parameters on the performance of the LDhelmet method. Our results send a message of caution regarding the reliability of reconstructed recombination maps and hotspot location.

Indeed we show that the recombination landscapes inferred with LDhelmet differ from real landscapes, sometimes substantially, with Spearman's rank correlation between real and inferred 2.5 kb windows sometimes as low as ~ 0.7 (Figure 3A). Hotspot detection is a particularly tricky and error-prone task: while up to 85% of true hotspots can be detected in the most favourable situations, the FDR ranged from 19% to 82% (Figure 4) according to the type of data and parameters used, meaning that in many cases a majority of the detected hotspots are incorrect calls. These discrepancies lead to a reduction in the apparent overlap in hotspot location between species/populations with identical recombination landscapes, while in turn inflating apparent hotspot sharing in populations with divergent landscapes (Table

1, Table S2). If neglected, these effects might mislead biological interpretations regarding the evolutionary conservation of recombination maps.

In a study of the short time-scale dynamics of recombination landscapes based on LDhelmet, Shanfelter et al. (2019) found only 15% of shared hotspots between two recently-diverged populations of threespine stickleback. A greater overlap in hotspot location was a priori expected given that this species lacks a functional PRDM9 protein, which is responsible for the rapid turnover of recombination landscapes in mammals (Axelsson et al. 2012; Paigen and Petkov, 2018). The authors suggested that a new mechanism of recombination hotspot regulation, different from the two already described in the literature, might be at work in this teleost species. In the light of our results, however, one cannot exclude that the strong divergence between the two reconstructed landscapes is due to a lack of power of the method in the first place. While the sample size of both fish populations was at least 20 individuals, θ was about 0.002, similar to our simulated conditions with a low N_e. Under these conditions, a high FDR and a low proportion of shared hotspots can be expected even if the true underlying maps are identical (Figure 4, Table 1).

It should be recalled that real data sets typically carry less signal and more noise than simulated data sets, meaning that our assessment of the reliability of LDhelmet might be an overoptimistic one. In particular, our data sets are immune from sequencing errors, mapping errors, and, crucially, phasing errors, all of which presumably make the problem of recombination map inference an even harder one.

Guidelines for population-based inference of recombination maps

Our study revealed that whatever the parameters used, the inference of recombination rates by LDhelmet is more reliable for species with large as compared to small effective population size (Figure 2, 3, 4). This might be expected since long-term N_e determines the amount of nucleotide diversity (θ=4$N_e\mu$, Watterson, 1975), so that a higher N_e results in a higher SNPs density and a finer scale characterization of the recombination rate variation along the genome. Moreover, a higher effective size greatly corrects the general tendency of LDhelmet to overestimate the ρ value, especially for low and high recombination rates (Figure 2, Supplementary Figure S5 ; Booker et al. 2017; Singhal et al. 2015). Thus, when studying species with varying effective population sizes, it is recommended to select populations with the largest N_e, for which genetic diversity is greater. The question is then: how to obtain a good-quality recombination map when dealing with low N_e species? The sampling effort also determines, to a lesser extent, the polymorphism level of the dataset (Supplementary Figure S3), improving the accuracy of the inference (Figure 3, 4). A sample size of 20 is

recommended based on our simulations. Moreover, as previously mentioned (Chan et al. 2012; Singhal et al. 2015), the block penalty parameter of LDhelmet, which determines the resolution level of the inferred landscape, also influences the length of the inferred map (*i.e.* a higher BP tends to mitigate the tendency of LDhelmet to overestimate the map length) and the number of detected hotspots (Supplementary Figure S2, S6). Therefore, a small BP, that allows more fine-scale changes in the inferred ρ value, should be used to detect recombination hotspots. The ability of LDhelmet to faithfully reflect the real recombination landscapes is of great importance when it comes to detecting recombination hotspots. To this purpose, the threshold used to decide which region is defined as a "hotspot" is a key parameter that determines the level of detection stringency. If the chosen value is not appropriate, LDhelmet will detect false positives while also missing true hotspots (Figure 4, Supplementary Figure S7). This threshold should thus be adapted to the species studied, using a less stringent threshold in species with lower genome-wide average recombination rate.

Other intrinsic biological variables influence the ability to produce a faithful recombination map, such as the μ/r ratio, which in part determines the power to measure rho at a fine-scale. The among-species variations in genome-average recombination rate r is well documented, ranging from 0.01 to 100 cM/Mb in animals and plants, with vertebrate taxa displaying an average r around 1 cM/Mb (Stapley et al. 2017). As previously mentioned, high and low recombination rates tend to be overestimated by LDhelmet, thus the average r of the studied species is obviously a key parameter to account for. The mutation rate μ also has a key impact on the performance of LDhelmet, since ancestral recombination events can only be detected if properly tagged by flanking mutations. The variation in μ across taxa, and consequently the ratio of μ/r, are much less well known than the variation in r. This ratio, which does not depend on the effective size of the population, is about 1 in humans, which means that two recombination events are separated by one mutation on average. A ratio in favour of mutations ($\mu/r > 1$) will improve the signal, increasing the TP rate and reducing the FD rate (Figure 5, Table 2). But ultimately the performance of LDhelmet is conditioned by r, as low r values provide less power to detect the recombination events, even with $\mu/r = 10$. Thus, the ratio of the mutation and the recombination rate is crucial to build a non-biased recombination map. Therefore, when studying a species for which it appears that this ratio is not favourable, a high rate of false positive hotspots is expected in the inferred population recombination landscape (Figure 5, Table 2), making it difficult to compare maps between closely related species in a meaningful way.

Limitations

The aim of our study was to determine the limits of LD-based methods in inferring biologically realistic recombination landscapes. For this purpose, we used the Pratto et al. (2014) ChIP-seq DMC1 data set to build human-like recombination landscapes including both broad and fine scale variation, reflected by the presence of numerous recombination hotspots of different intensities (Supplementary Figure S1; Myers et al. 2005, 2006; Pratto et al. 2014). However, the distribution of DSB might be more complex and variable than the true distribution of crossing overs. Thus, we may have generated landscapes too heterogeneous to be properly inferred by LDhelmet, reaching the limit of the method. Moreover, hotspots were placed randomly without taking into account the architecture of recombination hotspots along the genome: the proximity of genes and promoter-like regions, GC-rich regions, CpG islands, polymorphic regions, which can explain why a very intense and narrow hotspot is never found within a region of near zero recombination. Besides, we assume a homogeneous polymorphism along the chromosomal segments, which is not true in animal genomes (Ellegren and Galtier, 2016). Consequently, the sensibility of LD-based methods with respect to this architecture was not tested. Moreover, our simulated data were perfectly phased and polarised, which can't be the case when dealing with empirical data, and we didn't simulate various demographic histories that can also bias LD-based methods (Dapper and Payseur, 2018). Finally, we don't know if these simulated landscapes are representative of the diversity of recombination landscapes that exist in the living world.

Indeed, it is likely that the high complexity of the human recombination landscape is not a universal feature in the animal kingdom. Singhal et al. (2015) used LDhelmet for building the recombination map in two species of birds, the zebra finch and the long-tailed finch, that lack a full-length PRDM9 gene copy and diverged about 2.9 Myr. The sample size for both populations was about 20 individuals, and θ (~ 0.01) was about ten times higher than in apes or the threespine stickleback (Shanfelter et al. 2019), thus corresponding to our high N_e simulated conditions. Singhal et al. (2015) found 73% of shared hotspots between the two finch species, which is a higher rate of hotspot sharing than in any of the conditions we simulated. The median estimated recombination rate was of 0.14 cM/Mb in both species of finch, which is seven times lower than the average genomic recombination rate in humans (about 1 cM/Mb, Jensen-Seaman et al. 2004). Combined with the strong polymorphism in those species, we may suppose that birds possess less complex recombination landscapes than humans or compared to what we simulated, which might explain why LDhelmet recovered such a high percentage of shared hotspots in this study.

Conclusion

In the past few years, we have seen a growing interest in recombination rate estimation in functional and evolutionary genomics. Indirect, LD-based approaches raise methodological challenges that are addressed by sophisticated methods such as LDhat or LDhelmet, the reliability of which is still poorly characterised. Our study allows us to guide the users of such methods depending on the characteristics of their species, and calls for caution when it comes to interpreting fine scale differences in recombination rates between species. Extending this approach to a more diverse set of underlying recombination landscapes would help characterise further the reliability of these methods and their range of applicability across data sets and taxa.

Materials and Methods

Our approach separately considers three different layers of information that are involved in the study of recombination landscapes (Figure 1). The first layer that we call the "underlying" recombination landscape corresponds to the true biological distribution of recombination rate (r) across the considered genome. We here used experimental measurements from human studies to model and generate the "underlying" landscapes. The second layer, the population recombination landscape, describes the genomic location of recombination events that occurred during the history of the sample. We used coalescent simulations to produce these population recombination landscapes, thereafter called "simulated" landscapes. Simulated landscapes differ from the underlying landscape due to the stochasticity of the coalescent process, which is inversely proportional to N_e. The third layer, called the "inferred" landscape, corresponds to the output of LDhelmet, *i.e.* an estimate of the population recombination rate between adjacent SNPs. In total we generated five independent replicates of underlying landscapes, and for each of them 10 simulated and 10 inferred landscapes (Figure 1A).

Underlying landscapes

Underlying recombination landscapes were first generated to reproduce the features of the human recombination landscape. These include large-scale variation in the mean background recombination rate and fine-scale variation reflecting the presence of hotspots with varied intensities. Meiotic DSB are the major determinant of crossing over (CO) location along the genome (Li et al. 2019; Pratto et al. 2014). We used the high-resolution map of meiotic DSB obtained using ChIP-seq DMC1 in 5 non-related human genomes (Pratto et al.

2014) to define the genome-wide distribution of recombination rates in our simulations. The five individuals analysed in Pratto et al. (2014) carried different PRDM9 genotypes totalizing about 40,000 hotspots per individual, with distinct genotypes having different sets of DSB hotspots. For each individual, a gamma distribution was fitted to the empirical distribution of hotspot intensity measured by ChIP-seq DMC1 with the R package *figdistribplus* (Delignette-Muller and Dutang, 2015). Extreme ChIP-Seq intensity values (>500) lying above the 97.5th quantile and likely representing technical artefacts were removed. Remaining values were rescaled to 0-100, so as to transform ChIP-Seq intensity values into quantities reflecting the range of recombination rates reported in cM/Mb across the human genome (McVean et al. 2004; Myers et al. 2005). This conversion assumed a linear relation between DMC1 activity and CO frequency (Pratto et al. 2014). We then removed null values and replaced them with small but non-null values (0.001), so that the genome-wide mean recombination rate equaled a target average (e.g. 1 cM/Mb). A Gamma distribution was fitted to these transformed empirical values separately for each of the 5 individuals, before averaging shape and scale parameters across individuals. Targeted genome-wide average value was set to either 1 cM/Mb or 3 cM/Mb, respectively reflecting the average centro-chromosomal and subtelomeric rates in humans. Underlying landscapes of 1 Mb length were built by randomly drawing independent recombination rate values from the fitted distribution and assigning these to non-overlapping windows of 500pb. Values in the first 500kb were drawn from a distribution of mean 1 cM/Mb, while values in the last 500kb were drawn from a distribution of mean 3 cM/Mb. Our approach thus mimics both the large scale variation in recombination rate existing in humans (Buard and de Massy, 2007; Myers et al. 2005; Nachman, 2002; Pratto et al. 2014) and the nearly absence of recombination events outside hotspots (96% of CO occur in hotspots in mice, (Li et al. 2019; Pratto et al. 2014). In total, 5 underlying landscapes were generated (mean r = 2 cM/Mb), which can be considered as independent replicates driven from the same distribution (*i.e.* parts of different chromosomes of the same species, or orthologous chromosome part from closely related species).

Simulated landscapes

For each of the 5 underlying landscapes, 10 simulated landscapes were generated via coalescent simulations using the program MSPRIME (Kelleher and Lohse, 2020), varying the effective population size (N_e= 25,000 or 250,000) and the sample size (SS=10 or 20) and setting the mutation rate to μ = 10-8. These four combinations of simulation parameters were combined with three values of the Block Penalty (BP) parameter of the LDhelmet program (see below), resulting in twelve conditions tested (Supplementary Table S1). For each condition, ten population samples were simulated, to generate independent replicates of the

coalescent history. A VCF file was generated with MSPRIME for each simulated population (Figure 1), which contains the genotypes of variants that segregate in the population sample consisting of 2*n* sequences (with n being the number of samples) following the given underlying recombination landscape.

Inferred landscapes

Recombination rates were estimated for each of the simulated samples with LDhelmet (v1.10, (Chan et al. 2012), Figure 1). Briefly, LDhemet uses phased sequence data to infer the rho parameter locally, using likelihood computation betweens pairs of SNPs and then averaging over 50 consecutive variants to obtain a composite likelihood. The ρ parameter is inferred with a reversible-jump Markov Chain Monte Carlo algorithm using a step function applied to every window of 50 consecutive SNPs and determined by three parameters: the number of change-points, the locations of changes, and the recombination rate value of each constant fragment between two changes. We used VCFtools 0.1.17 (Danecek et al. 2011) and the vcf2fasta function of vcflib (https://github.com/vcflib/vcflib) to convert the SNP data obtained from MSPRIME simulations into the input format to LDhelmet, consisting of FASTA sequences of each individual haplotype. Ancestral states defined as the reference allele of each variant were also used as inputs. Each simulated replicate was analysed with LDhelmet using the following parameters. The haplotype configuration files were created with the find_conf function using the recommended window size of 50 SNPs. The likelihood look-up tables were created with the table_gen function using the recommended grid for the population recombination rate (ρ/pb) (*i.e.* ρ from 0 to 10 by increments of 0.1, then from 10 to 100 by increments of 1), and with the Watterson' $\theta = 4N_e\mu$ parameter corresponding to the condition analysed. The Padé files were created using 11 Padé coefficients as recommended. The Monte Carlo Markov chain was run for 1 million iterations with a burn-in period of 100,000 and a window size of 50 SNPs. An important parameter to LDhelmet is the block penalty (BP), which determines the number of change-points, and thus the variance of the inferred recombination rates at a fine scale (*i.e.* smaller block penalty generates more heterogeneous recombination landscapes). For each of the 4 combinations of Ne and SS simulated, the block penalty was set to either 5, 10 or 50, resulting in 12 combinations tested (Supplementary Table S1). Finally, the population recombination rates between each SNP pair were extracted with the post_to_text function, and were reported in $\rho=4N_er$ per pb unit.

The reliability of the inferences was evaluated in various ways. For each of the 12 conditions, the inferred, simulated and underlying landscapes were compared, in order to assess the ability of LDhelmet to reliably infer the true biological landscape (Figure 1A and

see below Hotspot detection and Statistical Analysis). In a second simulation framework aimed at evaluating the convergence of LDhelmet inferences across replicate runs, LDhelmet was run 10 times using the same parameters on each of two independently simulated VCF files from populations sharing the same underlying landscape (Figure 1B). Finally, for each of the twelve tested conditions, the inferred recombination landscapes of these two populations were compared in order to assess the reproducibility of the LDhelmet inference, *i.e.*, the expected variance between inferred maps in the absence of underlying biological variation (Figure 1B).

Variation in the μ/r ratio

To explore the influence of variation in mutation and recombination rates on the inference of recombination maps, two additional underlying landscapes were generated using the same procedure, this time targeting a ten times higher (*i.e.* 20 cM/Mb) or ten times lower (0.2 cM/Mb) mean recombination rate. Then, using one of the 5 underlying landscapes ($r \sim$ 10-8 M/pb) and the 2 newly generated landscapes with mean r = 10-7 and 10-9 M/pb, respectively, sets of simulations were run with a μ/r ratio of 0.1, 1 and 10. This was achieved by fixing μ to either 10-9, 10-8 or 10-7, while keeping a fixed N_e = 100,000 and SS = 20 (Table 2). For each of the 6 tested combinations of μ and r, 10 populations were simulated. These simulated landscapes were inferred with LDhelmet, using a block penalty of 5.

Hotspot detection

Underlying and simulated landscapes were first converted into population recombination rate landscapes by scaling them by $4N_e$. Underlying, simulated and inferred landscapes were then smoothed at a 500 bp and 2,500 bp resolution using the Python package *scipy.stats*. The former corresponds to the underlying landscape resolution, and the latter to a trade off between the density of segregating sites and the resolution often used in the litterature. For each underlying landscape and each of the 12 tested conditions, a mean simulated landscape and a mean inferred landscape were generated by averaging recombination rates across replicates.

Recombination hotspots of the underlying, simulated and inferred landscapes were called by comparing local vs surrounding recombination rates at each genomic window. A hotspot was defined as a window of 2.5 kb with an average recombination rate either 2.5, 5 or 10 times higher than the 50kb flanking regions (excluding the focal window). Hotspot locations were compared among landscapes using the same threshold values (*i.e.* 2.5/2.5, 5/5, or 10/10), except for three combinations in which lower thresholds were used for simulated compared to inferred landscapes (*i.e.* conditions 2.5/5, 2.5/10 and 5/10).

Statistical analyses

Statistical analyses were run with R 4.0.3. The length of underlying, simulated and inferred maps (*L*) was calculated at the 2.5 kb resolution using the formula:

$$L = \sum(\rho \times win)/4Ne,$$

with ρ the population-scaled recombination rate, *win* the window size resolution used to smooth the maps in bp, and N_e the effective size of the simulated population. Several indices of the sensitivity, specificity, reliability, and repeatability of LDhelmet were computed, using the mean simulated and inferred landscapes from each of the 12 tested conditions. For each condition, Spearman's rank correlation coefficient was calculated between the underlying and the corresponding simulated landscapes, between the simulated and inferred landscapes, and pairwise Spearman's coefficients among the 10 replicates inferred from the two simulated populations sharing the same underlying landscape. True/false positive rates (TPR = TP/(TP+FN) ; FPR = FP/(FP+TN)), true/false negative rates (TNR = TN/(TN+FP) ; FNR = FN/(FN/(TP+FN)), and true/false discovery rates (TDR = TP/(TP+FP) ; FDR = FP/(TP+FP)) were calculated by comparing the simulated and inferred landscapes. The mean pairwise linear correlation (R^2) and the proportion of shared hotspots was calculated between the 5 underlying landscapes, and for each condition and for the three threshold values tested (*i.e.* 2.5, 5 and 10) between the simulated and inferred landscapes from the 5 different underlying landscapes, as well as between the two populations from the same underlying landscape.

The statistical analyses were performed using home-made R scripts available upon request.

Data availability

The Singularity container recipe built to run the simulations is available at: https://github.com/PA-GAGNAIRE/Singularity-Recipes/tree/master/HotRec-Recipes. This recipe contains the installation command lines of the required programs, the scripts used for the simulations, and the five underlying landscapes used in our study.

Acknowledgments

We are grateful to Julien Joseph, Nicolas Lartillot, Frédéric Baudat, Bernard de Massy and Laurent Duret for the discussions and feedback. We also thank Khalid Belkhir, Jimmy Lopez and Mathieu Massaviol for their support to build the Singularity container used to run our simulations. This project was funded by the ANR HotRec ANR-19-CE12-0019.

Competing Interests Statement

The authors declare no conflict of interest.

References

Arenas, M. (2013). The importance and application of the ancestral recombination graph. Frontiers in Genetics, 4. https://doi.org/10.3389/fgene.2013.00206

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E. M., Bowden, R., Aneas, I., Broxholme, J., Humburg, P., Iqbal, Z., Lunter, G., Maller, J., Hernandez, R. D., Melton, C., Venkat, A., Nobrega, M. A., Bontrop, R., … McVean, G. (2012). A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. Science, 336(6078), 193‐198. https://doi.org/10.1126/science.1216872

Auton, A., & McVean, G. (2007). Recombination rate estimation in the presence of hotspots. Genome Research, 17(8), 1219‐1227. https://doi.org/10.1101/gr.6386707

Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J. K., Hayward, J. J., Cohen, P. E., Greally, J. M., Wang, J., Bustamante, C. D., & Boyko, A. R. (2013). Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLoS Genetics, 9(12), e1003984. https://doi.org/10.1371/journal.pgen.1003984

Axelsson, E., Webster, M. T., Ratnakumar, A., The LUPA Consortium, Ponting, C. P., & Lindblad-Toh, K. (2012). Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. Genome Research, 22(1), 51‐63. https://doi.org/10.1101/gr.124123.111

Baker, Z., Schumer, M., Haba, Y., Bashkirova, L., Holland, C., Rosenthal, G. G., & Przeworski, M. (2017). Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. ELife, 6, e24133. https://doi.org/10.7554/eLife.24133

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., & de Massy, B. (2010). PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science, 327(5967), 836‐840. https://doi.org/10.1126/science.1183439

Baudat, F., & de Massy, B. (2007). Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. Chromosome Research, 15(5), 565‐577. https://doi.org/10.1007/s10577-007-1140-3

Booker, T. R., Ness, R. W., & Keightley, P. D. (2017). The Recombination Landscape in Wild House Mice Inferred Using Population Genomic Data. Genetics, 207(1), 297‐309. https://doi.org/10.1534/genetics.117.300063

Buard, J., & de Massy, B. (2007). Playing hide and seek with mammalian meiotic crossover hotspots. Trends in Genetics, 23(6), 301‐309. https://doi.org/10.1016/j.tig.2007.03.014

Capilla, L., Garcia Caldés, M., & Ruiz-Herrera, A. (2016). Mammalian Meiotic Recombination : A Toolbox for Genome Evolution. Cytogenetic and Genome Research, 150(1), 1‐16. https://doi.org/10.1159/000452822

Carrington, M., & Cullen, M. (2004). Justified chauvinism : Advances in defining meiotic recombination through sperm typing. Trends in Genetics, 20(4), 196‐205. https://doi.org/10.1016/j.tig.2004.02.006

Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. PLoS Genetics, 8(12), e1003090. https://doi.org/10.1371/journal.pgen.1003090

Choi, K., & Henderson, I. R. (2015). Meiotic recombination hotspots—A comparative view. The Plant Journal, 83(1), 52‐61. https://doi.org/10.1111/tpj.12870

Clark, A. G., Wang, X., & Matise, T. (2010). Contrasting Methods of Quantifying Fine Structure of Human Recombination. Annual Review of Genomics and Human Genetics, 11(1), 45‐64. https://doi.org/10.1146/annurev-genom-082908-150031

Cromie, G. A., Connelly, J. C., & Leach, D. R. F. (2001). Recombination at Double-Strand Breaks and DNA Ends. Molecular Cell, 8(6), 1163‐1174. https://doi.org/10.1016/S1097-2765(01)00419-1

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156‐2158. https://doi.org/10.1093/bioinformatics/btr330

Danguy des Déserts, A., Bouchet, S., Sourdille, P., & Servin, B. (2021). Evolution of recombination landscapes in diverging populations of bread wheat. Genome Biology and Evolution, evab152. https://doi.org/10.1093/gbe/evab152

Dapper, A. L., & Payseur, B. A. (2018). Effects of Demographic History on the Detection of Recombination Hotspots from Linkage Disequilibrium. Molecular Biology and Evolution, 35(2), 335‐353. https://doi.org/10.1093/molbev/msx272

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus : An R Package for Fitting Distributions. Journal of Statistical Software, 64(4). https://doi.org/10.18637/jss.v064.i04

Dréau, A., Venu, V., Avdievich, E., Gaspar, L., & Jones, F. C. (2019). Genome-wide recombination map construction from single individuals using linked-read sequencing. Nature Communications, 10(1), 4309. https://doi.org/10.1038/s41467-019-12210-9

Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. Nature Reviews Genetics, 17(7), 422‐433. https://doi.org/10.1038/nrg.2016.58

Grey, C., Baudat, F., & de Massy, B. (2018). PRDM9, a driver of the genetic map. PLOS Genetics, 14(8), e1007479. https://doi.org/10.1371/journal.pgen.1007479

Haenel, Q., Laurentino, T. G., Roesti, M., & Berner, D. (2018). Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics.

Molecular Ecology, 27(11), 2477‑2497. https://doi.org/10.1111/mec.14699

Hunter, C. M., Huang, W., Mackay, T. F. C., & Singh, N. D. (2016). The Genetic Architecture of Natural Variation in Recombination Rate in Drosophila melanogaster. PLOS Genetics, 12(4), e1005951. https://doi.org/10.1371/journal.pgen.1005951

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C.-F., Thomas, M. A., Haussler, D., & Jacob, H. J. (2004). Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Research, 14(4), 528‑538. https://doi.org/10.1101/gr.1970304

Kamm, J. A., Spence, J. P., Chan, J., & Song, Y. S. (2016). Two-Locus Likelihoods Under Variable Population Size and Fine-Scale Recombination Rate Estimation. Genetics, 203(3), 1381‑1399. https://doi.org/10.1534/genetics.115.184820

Kelleher, J., & Lohse, K. (2020). Coalescent Simulation with msprime. In J. Y. Dutheil (Éd.), Statistical Population Genomics (Vol. 2090, p. 191‑230). Springer US. https://doi.org/10.1007/978-1-0716-0199-0_9

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., Ecker, J. R., Weigel, D., & Nordborg, M. (2007). Recombination and linkage disequilibrium in Arabidopsis thaliana. Nature Genetics, 39(9), 1151‑1155. https://doi.org/10.1038/ng2115

Kodama, M., Brieuc, M. S. O., Devlin, R. H., Hard, J. J., & Naish, K. A. (2014). Comparative Mapping Between Coho Salmon (Oncorhynchus kisutch) and Three Other Salmonids Suggests a Role for Chromosomal Rearrangements in the Retention of Duplicated Regions Following a Whole Genome Duplication Event. G3 Genes|Genomes|Genetics, 4(9), 1717‑1730. https://doi.org/10.1534/g3.114.012294

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., & Stefansson, K. (2002). A high-resolution recombination map of the human genome. Nature Genetics, 31(3), 241‑247. https://doi.org/10.1038/ng917

Lander, E. S., & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences, 84(8), 2363‑2367. https://doi.org/10.1073/pnas.84.8.2363

Latrille, T., Duret, L., & Lartillot, N. (2017). The Red Queen model of recombination hot-spot evolution : A theoretical investigation. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1736), 20160463. https://doi.org/10.1098/rstb.2016.0463

Lesecque, Y., Glémin, S., Lartillot, N., Mouchiroud, D., & Duret, L. (2014). The Red Queen Model of Recombination Hotspots Evolution in the Light of Archaic and Modern Human Genomes. PLoS Genetics, 10(11), e1004790. https://doi.org/10.1371/journal.pgen.1004790

Li, N., & Stephens, M. (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. Genetics, 165(4), 2213‑2233. https://doi.org/10.1093/genetics/165.4.2213

Li, R., Bitoun, E., Altemose, N., Davies, R. W., Davies, B., & Myers, S. R. (2019). A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nature Communications, 10(1), 3900. https://doi.org/10.1038/s41467-019-11675-y

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., & Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature, 454(7203), 479‑485. https://doi.org/10.1038/nature07135

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The Fine-Scale Structure of Recombination Rate Variation in the Human Genome. Science, 304(5670), 581‑584. https://doi.org/10.1126/science.1092500

Melamed-Bessudo, C., Shilo, S., & Levy, A. A. (2016). Meiotic recombination and genome evolution in plants. Current Opinion in Plant Biology, 30, 82‑87. https://doi.org/10.1016/j.pbi.2016.02.003

Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., Riera-Lizarazu, O., Brown, P. J., Acharya, C. B., Mitchell, S. E., Harriman, J., Glaubitz, J. C., Buckler, E. S., & Kresovich, S. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proceedings of the National Academy of Sciences, 110(2), 453‑458. https://doi.org/10.1073/pnas.1215985110

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science, 310(5746), 321‑324. https://doi.org/10.1126/science.1117196

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., & Donnelly, P. (2010). Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. Science, 327(5967), 876‑879. https://doi.org/10.1126/science.1182363

Myers, S., Spencer, C. C. A., Auton, A., Bottolo, L., Freeman, C., Donnelly, P., & McVean, G. (2006). The distribution and causes of meiotic recombination in the human genome. Biochemical Society Transactions, 34, 5.

Nachman, M. (2002). Variation in recombination rate across the genome: Evidence and implications. Current Opinion in Genetics & Development, 12(6), 657‑663. https://doi.org/10.1016/S0959-437X(02)00358-1

Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., Beatson, S. A., Lunter, G., Malik, H. S., & Ponting, C. P. (2009). Accelerated Evolution of the Prdm9 Speciation Gene across Diverse Metazoan Taxa. PLoS Genetics, 5(12), e1000753.

https://doi.org/10.1371/journal.pgen.1000753

Paigen, K., & Petkov, P. M. (2018). PRDM9 and Its Role in Genetic Recombination. Trends in Genetics, 34(4), 291‐300. https://doi.org/10.1016/j.tig.2017.12.017

Parvanov, E. D., Petkov, P. M., & Paigen, K. (2010). Prdm9 Controls Activation of Mammalian Recombination Hotspots. Science, 327(5967), 835‐835. https://doi.org/10.1126/science.1181495

Peñalba, J. V., & Wolf, J. B. W. (2020). From molecules to populations : Appreciating and estimating recombination rate variation. Nature Reviews Genetics, 21(8), 476‐492. https://doi.org/10.1038/s41576-020-0240-1

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., & Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. Science, 346(6211), 1256442. https://doi.org/10.1126/science.1256442

Rastas, P. (2017). Lep-MAP3 : Robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics, 33(23), 3726‐3732. https://doi.org/10.1093/bioinformatics/btx494

Schield, D. R., Pasquesi, G. I. M., Perry, B. W., Adams, R. H., Nikolakis, Z. L., Westfall, A. K., Orton, R. W., Meik, J. M., Mackessy, S. P., & Castoe, T. A. (2020). Snake Recombination Landscapes Are Concentrated in Functional Regions despite PRDM9. Molecular Biology and Evolution, 37(5), 1272‐1294. https://doi.org/10.1093/molbev/msaa003

Shanfelter, A. F., Archambeault, S. L., & White, M. A. (2019). Divergent Fine-Scale Recombination Landscapes between a Freshwater and Marine Population of Threespine Stickleback Fish. Genome Biology and Evolution, 11(6), 1552‐1572. https://doi.org/10.1093/gbe/evz090

Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., Hooper, D. M., Strand, A. I., Li, Q., Raney, B., Balakrishnan, C. N., Griffith, S. C., McVean, G., & Przeworski, M. (2015). Stable recombination hotspots in birds. Science, 350(6263), 928‐932. https://doi.org/10.1126/science.aad0843

Spence, J. P., & Song, Y. S. (2019). Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. SCIENCE ADVANCES, 15.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., & Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes : Patterns and processes. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1736), 20160455. https://doi.org/10.1098/rstb.2016.0455

Stumpf, M. P. H., & McVean, G. A. T. (2003). Estimating recombination rates from population-genetic data. Nature Reviews Genetics, 4(12), 959‐968. https://doi.org/10.1038/nrg1227

Sun, H., Rowan, B. A., Flood, P. J., Brandt, R., Fuss, J., Hancock, A. M., Michelmore, R. W., Huettel, B., & Schneeberger, K. (2019). Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. Nature Communications, 10(1), 4310. https://doi.org/10.1038/s41467-019-12209-2

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature, 449(7164), 851‐861. https://doi.org/10.1038/nature06258

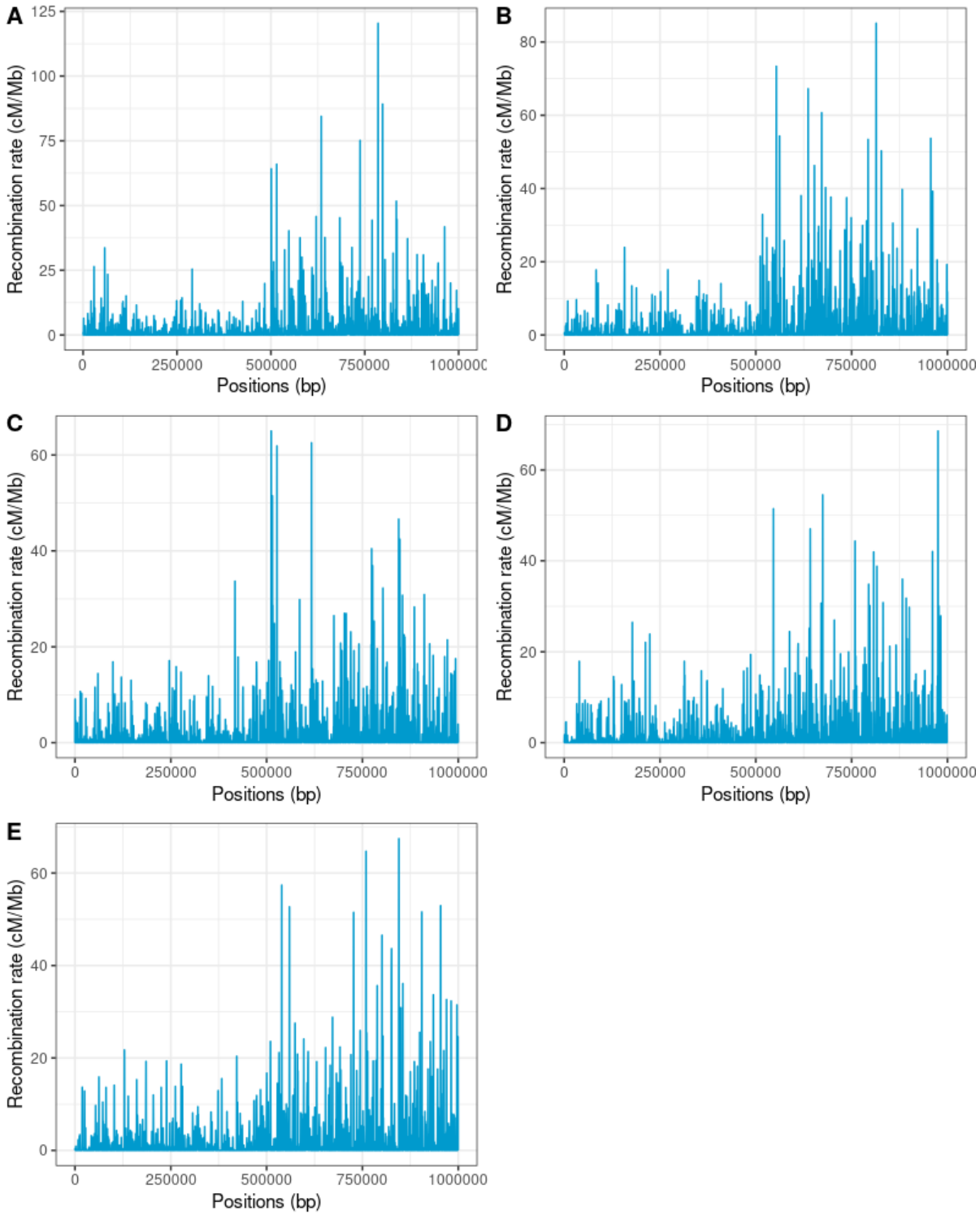
Wallberg, A., Glémin, S., & Webster, M. T. (2015). Extreme Recombination Frequencies Shape Genome Variation and Evolution in the Honeybee, Apis mellifera. PLOS Genetics, 11(4), e1005189. https://doi.org/10.1371/journal.pgen.1005189

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. Theoretical Population Biology, 7(2), 256‐276. https://doi.org/10.1016/0040-5809(75)90020-9

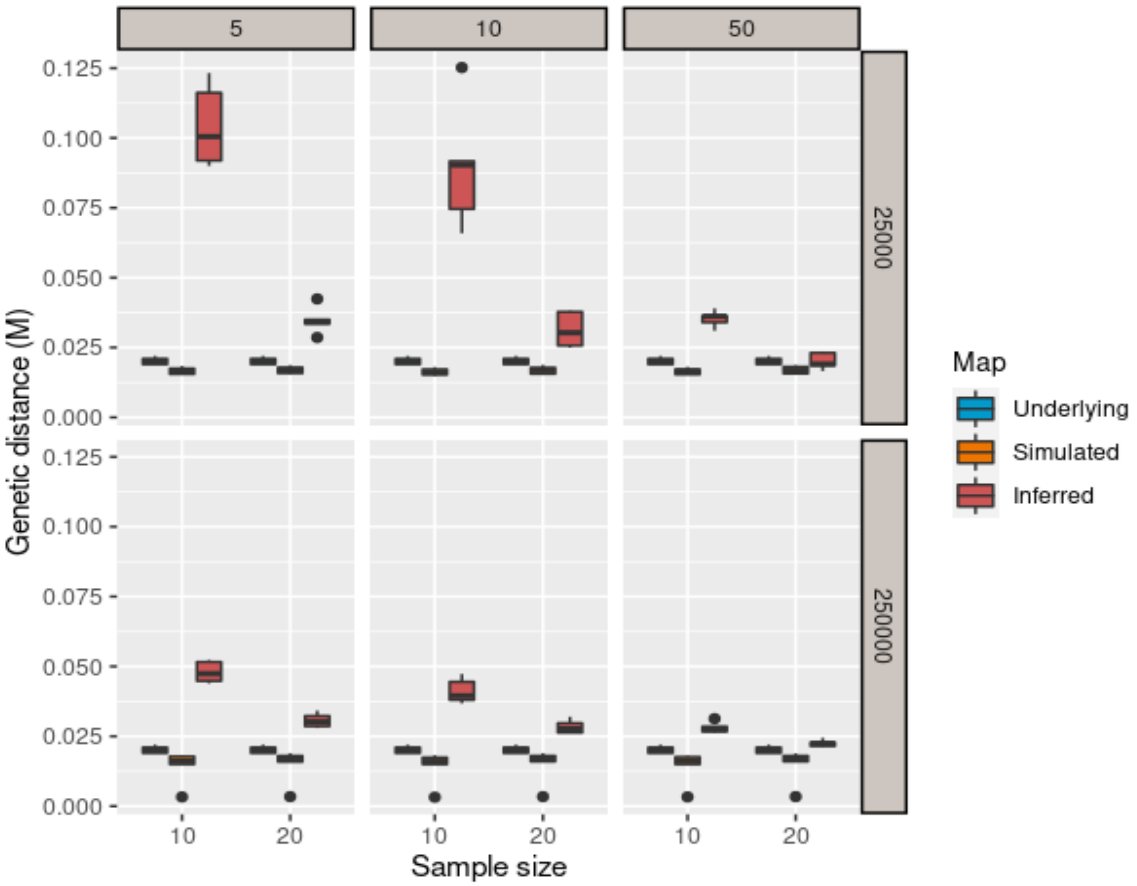
Supplementary Figures

Supplementary Table S1. The 4 combinations of coalescent simulation parameters (N_e and SS) in combination with three values of the block penalty parameter to LDhelmet (BP), resulting in a total of 12 tested conditions in our analysis.

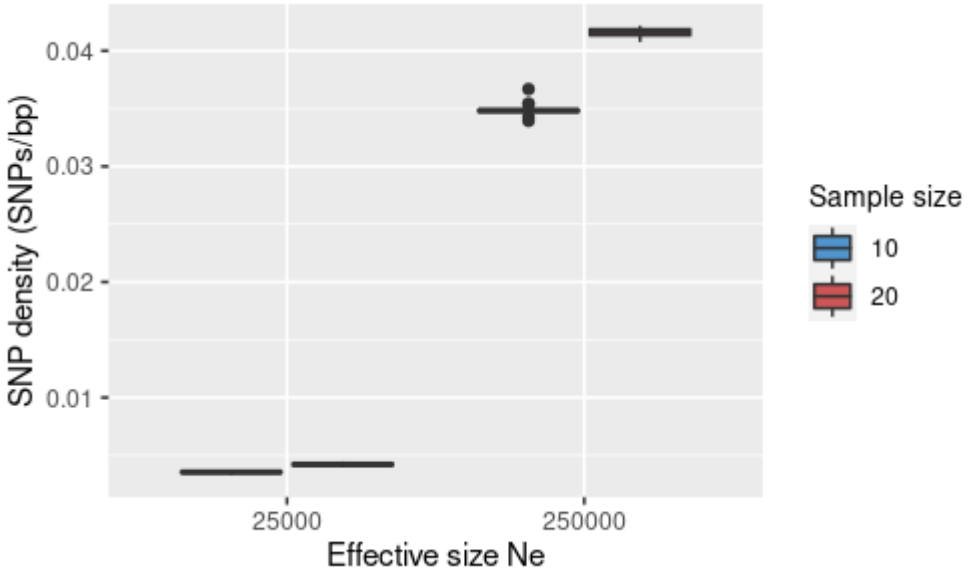
Ne	SS	BP
25000	10	5
		10
		50
	20	5
		10
		50
250000	10	5
		10
		50
	20	5
		10
		50



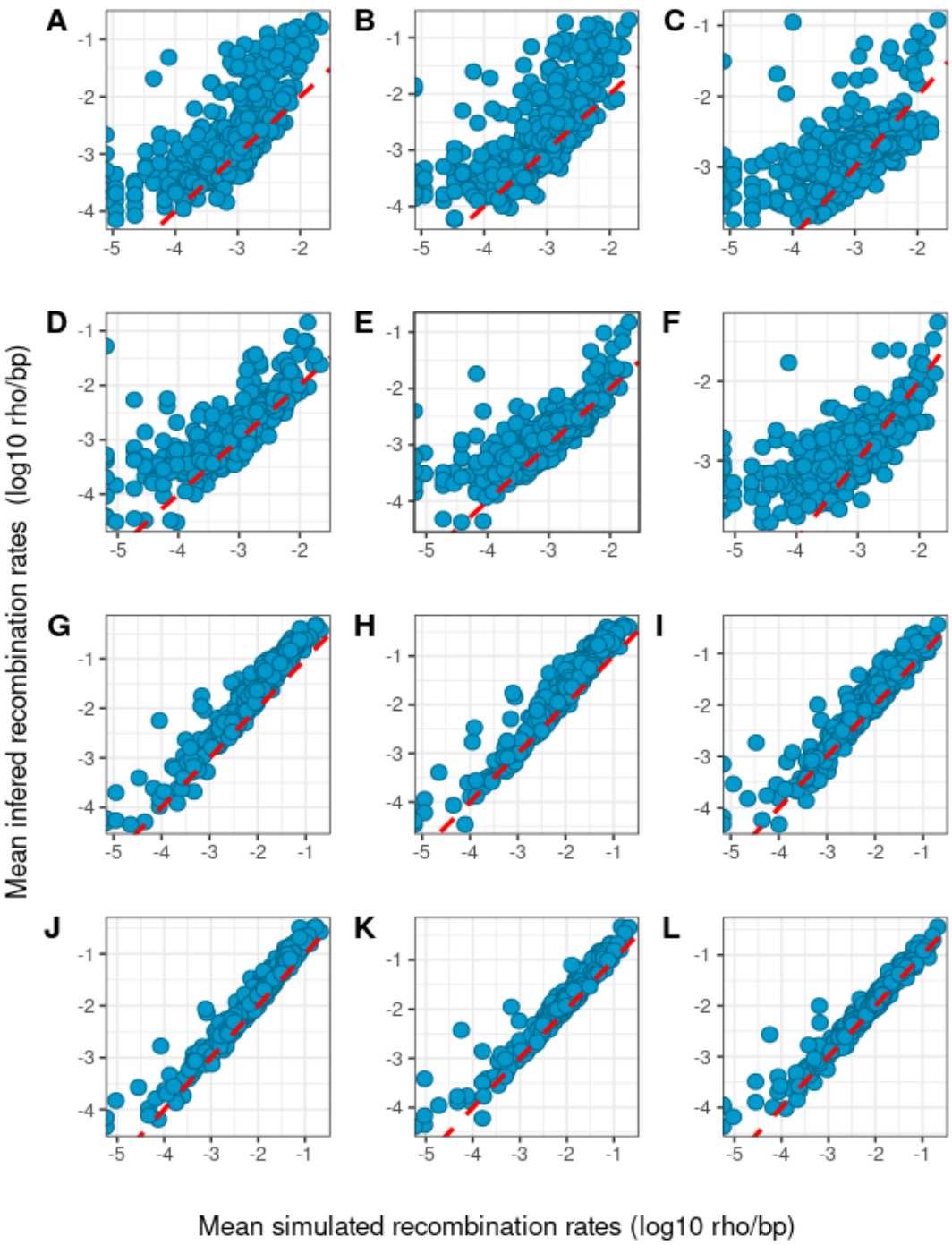
Supplementary Figure S1. The 5 underlying recombination landscapes (represented in units of cM/Mb (y-axis) along a chromosomal region of 1Mb (x axis)) generated using human ChIP-seq data from Pratto et al. (2014).



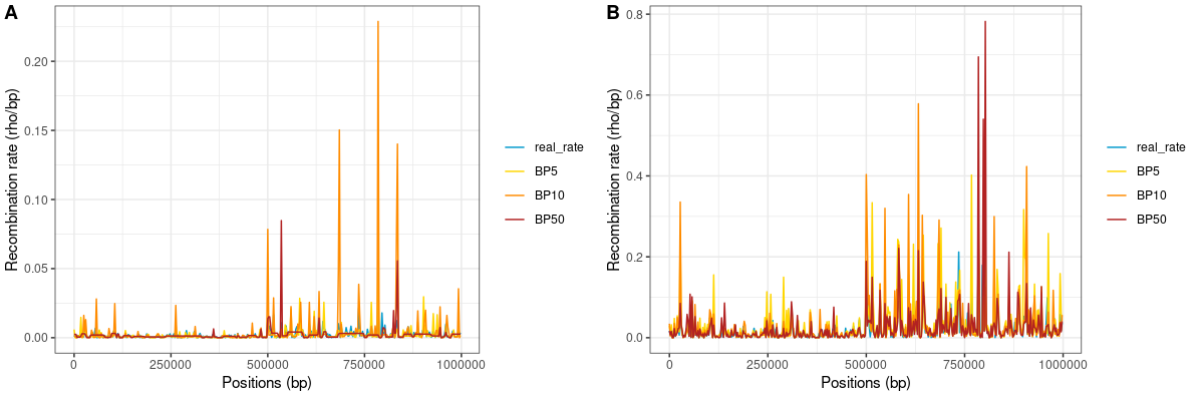
Supplementary Figure S2. Underlying, simulated and inferred map length, analysed at a 2.5 kb resolution.



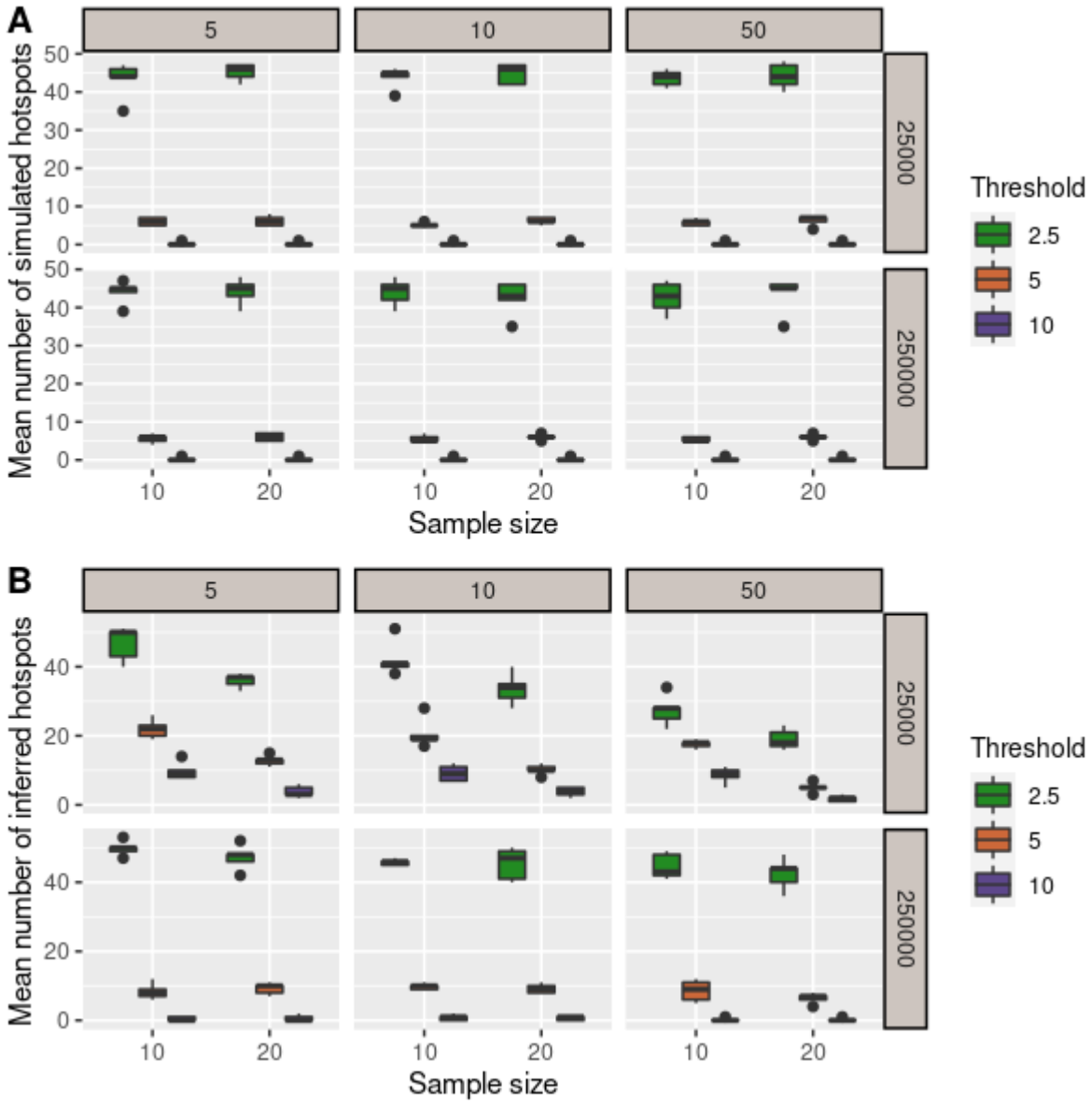
Supplementary Figure S3. SNP density according to coalescent simulation parameters (N_e and SS). The sample size (SS) parameter is shown with different colours (*i.e.* 10, blue or 20, red), and the effective population size (N_e) is shown on the x axis. The mutation parameter was μ = 10-8.



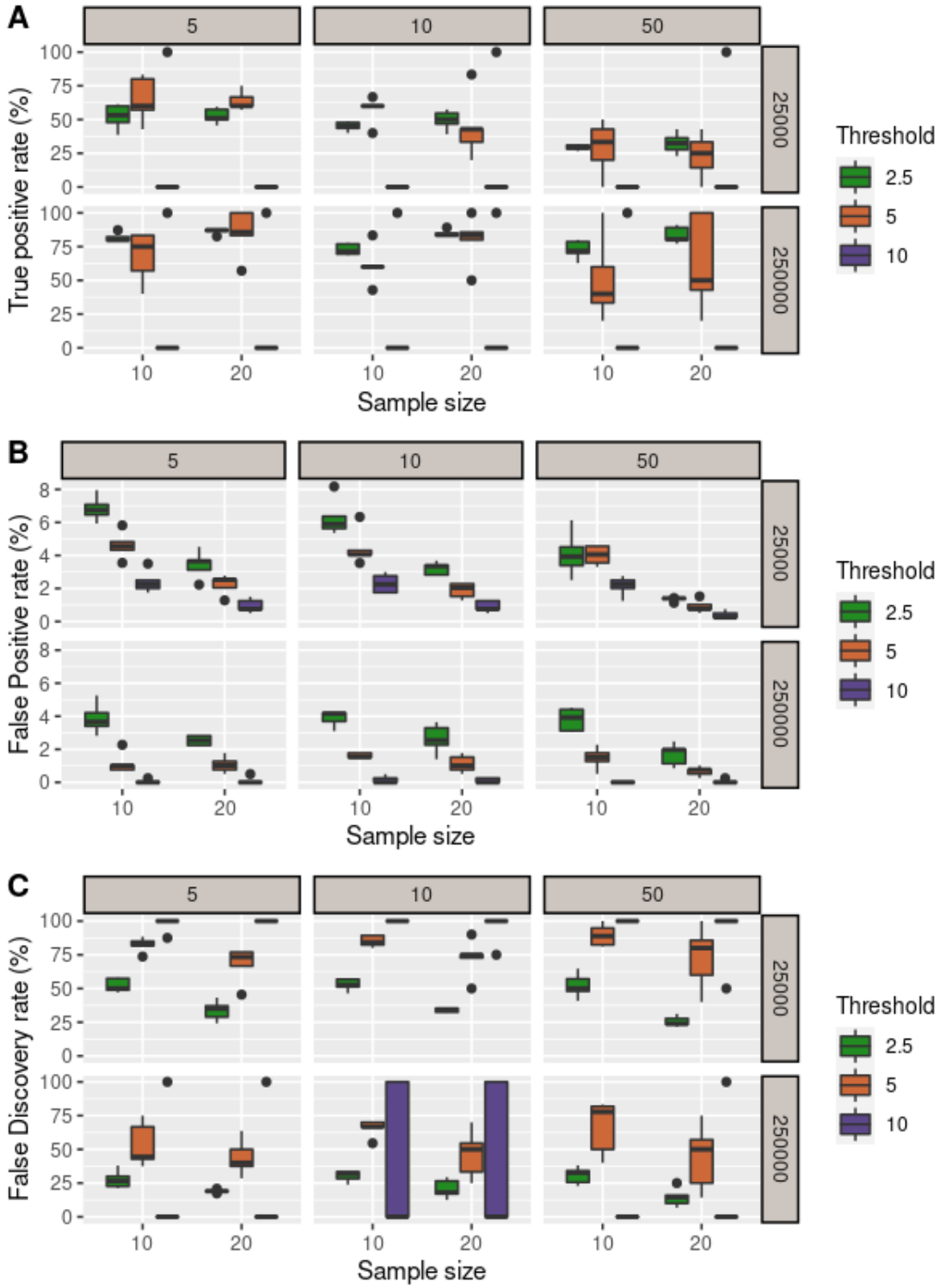
Supplementary Figure S4. Quality assessment of local recombination rates estimated by LDhelmet and averaged within 2.5kb windows across 10 replicates, for the 12 conditions tested in our analysis. The x axis shows the recombination rates of the mean simulated landscapes and the y axis the recombination rates of the mean inferred landscapes, both on a logarithmic scale. Each blue point corresponds to a local 2.5kb-window average calculated across 10 replicate populations obtained under identical simulation parameters. **A)** N_e = 25,000, SS = 10, BP = 5. **B)** N_e = 25,000, SS = 10, BP = 10. **C)** N_e = 25,000, SS = 10, BP = 50. **D)** N_e = 25,000, SS = 20, BP = 5. **E)** N_e = 25,000, SS = 20, BP = 10. **F)** N_e = 25,000, SS = 20, BP = 50.**G)** N_e = 250,000, SS = 10, BP = 5. **H)** N_e = 250,000, SS = 10, BP = 10. **I)** N_e = 250,000, SS = 10, BP = 50. **J)** N_e = 250,000, SS = 20, BP = 5. **K)** N_e = 250,000, SS = 20, BP = 10. **L)** N_e = 250,000, SS = 20, BP = 50.



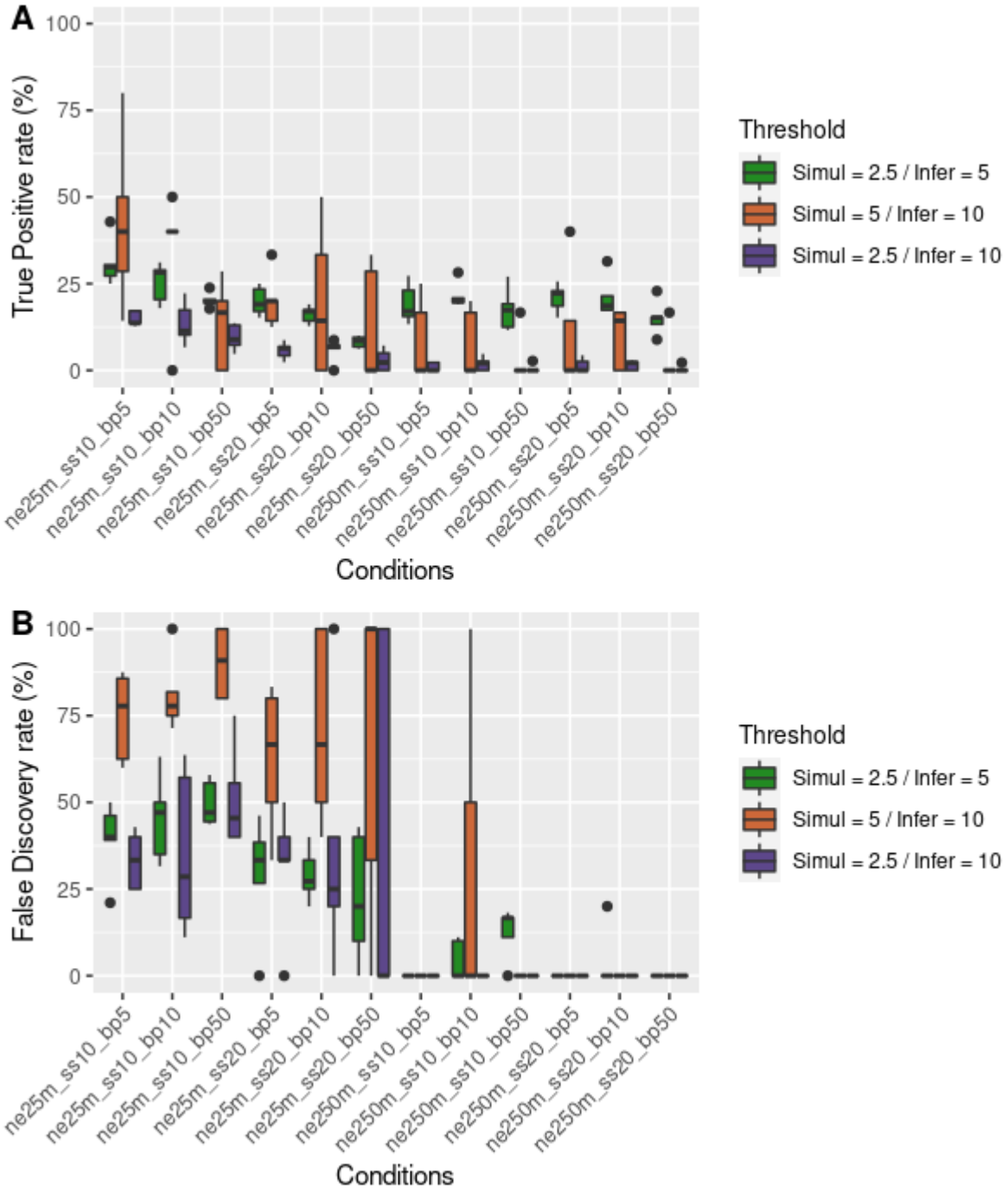
Supplementary Figure S5. Influence of the block penalty value on the recombination landscapes inferred by LDhelmet. The map recombination rate variation (represented in units ofρ/bp) is shown in blue for one of the five underlying landscapes. The corresponding recombination landscape inferred with a BP = 5, 10 and 50 are shown in yellow, orange and red respectively, for SS = 20 and N_e = 25,000 (**A**) and N_e = 250,000 (**B**).



Supplementary Figure S6. Mean number of hotspots detected from the simulated (A) and inferred (B) landscapes originating from the 5 different underlying landscapes. Results are represented for each of the 12 conditions tested, and for three hotspots detection threshold as shown by colours (*i.e.* 2.5, 5 and 10). The sample size parameter is shown on the x axis (*i.e.* 10 or 20), the upper panels correspond to conditions where N_e = 25,000, the lower panels correspond to conditions where N_e = 250,000, and the BP parameter values correspond to the vertical panels (*i.e.* from left to right : 5, 10 and 50). **A)** Mean number of simulated hotspots. **B)** Mean number of inferred hotspots.



Supplementary Figure S7. Hotspot detection from the 5 mean inferred landscapes originating from the 5 different underlying landscapes. Results are represented for each of the 12 conditions tested, and for three hotspot detection thresholds as shown by colours (*i.e.* 2.5, 5 and 10). The sample size parameter is shown on the x axis (*i.e.* 10 or 20), the upper panels correspond to conditions where N_e = 25,000, the lower panels correspond to conditions where N_e = 250,000, and the BP parameter values correspond to the vertical panels (*i.e.* from left to right : 5, 10 and 50). **A)** True positive rate (sensitivity). **B)** False positive rate. **C)** False discovery rate.



Supplementary Figure S8. Hotspot detection from the 5 mean inferred landscapes originating from 5 different underlying landscapes, for each of the 12 conditions tested (on the y axis), and for three combinations of "real hotspots definition / inferred hotspot detection threshold" applied to the mean simulated and the mean inferred landscape, respectively (*i.e.* 2.5/5, 5/10, 2.5/10, as shown by colours). **A)** True positive rate (sensitivity). **B)** False discovery rate.

Supplementary Table S2. Percentages of shared hotspots from simulated and inferred landscapes between populations with either different or identical underlying recombination landscapes. R^2 and mean proportion of shared hotspots are indicated for each type of comparison, including pairwise comparisons among simulated (left) or inferred (right) landscapes originating from either different underlying landscapes (top, following simulation framework from Figure 1A), or from the same underlying landscape (down, following simulation framework from Figure 1B).

		Mean proportion of shared simulated hotspots				Mean proportion of shared inferred hotspots		
	R^2 log(simuli~simulj)	Threshold = 5	Threshold = 2.5	Threshold = 10	R^2 log(inferi~inferj)	Threshold = 5	Threshold = 2.5	Threshold = 10
Different underlying landscapes								
Ne = 25000 SS = 10 BP = 5	0.012	1.715	8.065	0.00	0.033	5.385	13.99	3.215
Ne = 25000 SS = 10 BP = 10	0.015	1.835	7.275	0.00	0.043	4.945	11.28	0.00
Ne = 25000 SS = 10 BP = 50	0.014	1.55	7.595	0.00	0.039	5.075	5.895	2.51
Ne = 25000 SS = 20 BP = 5	0.014	2.00	9.045	0.00	0.032	3.95	9.735	1.835
Ne = 25000 SS = 20 BP = 10	0.013	1.55	8.505	0.00	0.042	1.125	7.46	0.00
Ne = 25000 SS = 20 BP = 50	0.015	1.43	8.55	0.00	0.084	0.00	4.28	0.00
Ne = 250000 SS = 10 BP = 5	0.017	2.085	8.42	0.00	0.013	0.00	11.245	0.00
Ne = 250000 SS = 10 BP = 10	0.016	2.00	8.18	0.00	0.013	3.055	11.125	0.00
Ne = 250000 SS = 10 BP = 50	0.015	2.00	7.11	0.00	0.012	0.97	9.86	0.00
Ne = 250000 SS = 20 BP = 5	0.017	1.715	7.675	0.00	0.014	1.17	9.55	0.00
Ne = 250000 SS = 20 BP = 10	0.013	1.55	7.12	0.00	0.014	1.18	8.695	0.00
Ne = 250000 SS = 20 BP = 50	0.016	1.67	8.075	0.00	0.015	1.34	7.32	0.00
Same underlying landscape								
Ne = 25000 SS = 10 BP = 5	0.7463	85.71	79.62	0	0.2139	30.65	28.55	20.95
Ne = 25000 SS = 10 BP = 10	0.7386	72.92	77.50	0	0.1429	10.9	12.9	0
Ne = 25000 SS = 10 BP = 50	0.7101	77.78	71.15	0	0.1173	10.7	21.7	0
Ne = 25000 SS = 20 BP = 5	0.7497	80.36	78.99	0	0.311	14.85	26.9	0
Ne = 25000 SS = 20 BP = 10	0.7908	78.75	72.61	0	0.3035	27.7	23.55	22.9
Ne = 25000 SS = 20 BP = 50	0.7622	65.00	76.09	0	0.2206	0	21.15	0
Ne = 250000 SS = 10 BP = 5	0.9466	80.36	90.42	0	0.6301	15.3	35.65	0
Ne = 250000 SS = 10 BP = 10	0.9346	85.71	85.81	0	0.6671	31.75	38.95	0
Ne = 250000 SS = 10 BP = 50	0.9391	81.25	90.33	0	0.6019	6.3	36	0
Ne = 250000 SS = 20 BP = 5	0.9554	80.36	93.79	0	0.7514	32.55	49.15	41.65
Ne = 250000 SS = 20 BP = 10	0.9562	92.86	92.57	0	0.7157	26.7	55.55	0
Ne = 250000 SS = 20 BP = 50	0.9488	87.50	82.38	0	0.7421	56.9	54.25	0