Patterns of nucleotide diversity under background selection and evolving recombination rates

Tom R. Booker*
*University of British Columbia

ABSTRACT *I'm just using the GENETICS template because it looks nice!*

KEYWORDS Evolutionary Genetics, Population Genetics

Introduction

Natural selection influences the frequencies of genetic variants that affect their carriers' fitness as well as neutral alleles linked to selected sites. The extent by which natural selection at one site influences linked sites depends on the rate of mutations that affect fitness and the rate of recombination. Of the mutations that influence fitness, the majority are likely deleterious with a comparatively small proportion being beneficial (REFS). When deleterious mutations occur they may removed through purifying selection and this can cause linked neutral variants may be lost along with them through a process referred to as background selection (BGS)(Charlesworth et al 1993). Alternatively, as beneficial alleles spread to high frequency, linked neutral variants may hitchhike along with them in a process known as a selective sweep. The effects of both BGS and selective sweeps on a particular neutral locus depend on the rate of recombination between the neutral site and the sites subject to selection (REFS). Indeed, theoretical models of background selection and selective sweeps have provided evolutionary biologists with a framework for understanding natural selection through the analysis of genome-wide patterns of nucleotide diversity.

The first empirical evidence that selection at linked sites influences genetic variation across the genome came from studies in *Drosophila*. Aguadé et al (1989) measured genetic variability in the *yellow-achaete-scute* regions located at the tip of the X-chromosome. The *yellow-achaete-scute* regions experience restricted crossing-over and Aguadé et al (1989) found that they harbour far less genetic variation than had been reported for more highly recombining regions of the genome. Aguadé et al (1989) suggested that selective sweeps (though that term was not coined until later), which reduce nucleotide diversity (π) to the greatest extent in regions of restricted recombination, potentially explained their findings. Begun and Aquadro (1992) then showed that there is a clear correlation between recombination rate and nucleotide diversity (π) using loci sampled from across the *D. melanogaster* genome. Furthermore, Begun and Aquadro (1992) showed that there was little evidence for a correlation of between-species divergence and recombination rate, which one might expect if recombination were itself mutagenic. By fitting a model of selective sweeps to the observed correlation between π and recombination rate, Wiehe and Stephan (1993) made inferences about the rate and strength of beneficial mutations. Around the same time, however, Charlesworth et al (1993) demonstrated that BGS could also potentially explain the correlation between π and the recombination rate. In the time since these papers were published, many studies have used the relationship between nucleotide diversity and recombination rate to make inferences about the strength of selection and the rate at which selected mutations occur (REF DUMP).

Interpreting patterns of genetic variability in terms of selection at linked sites relies on accurate estimates of recombination rates. Additionally, there is an implicit assumption that the landscape of recombination evolves

Manuscript compiled: Monday 23rd August, 2021

slowly and can be considered invariant over the timescale relevant for patterns of diversity to be influenced by selection. Empirical estimates of recombination rates can be obtained by examining the inheritance of genetic markers through known pedigrees, as in traditional genetic mapping, or by directly comparing an individual's genome to the the genomes of its gametes (REF). Both methods provide estimates of recombination in contemporary population, but if recombination rates evolve rapidly, empirical recombination rate estimates may not reflect patterns of nucleotide diversity that were influenced by the ancestral recombination rate landscape.

There is reason to think that recombination rate landscapes have evolved rapidly in some lineages. In the house mouse (*Mus musculus*), for example, there has likely been extensive evolution of recombination rate landscape in the last five million years. It is estimated that *Mus musculus* (2*n*=40) and *Mus pahari* (2*n*=48) began to diverge around 5 million years ago and in that time approximately 18 large syntenic breaks accumulated in the lineage leading to *M. musculus* (Thybert et al 2018). Due to the requirement of at least one cross-over per meiosis in mammals, an increase or a decrease in the length of a chromosome will necessarily influence the recombination rate. Thus karyotype evolution has likely influenced recombination rate landscapes at broad scales in *Mus musculus*. More recently, the *M. musculus* lineage began to diverge into the M. musculus species complex (REFS), and differences in the genome-wide rate of recombination have arisen amongst the members of this group (Dumont and Payseur XXXX). Finally, Recombination in mice is typically restricted to narrow windows of the genome (on the order of 10,000 basepairs) referred to as recombination hotspots (REFS). The location of recombination hotspots in mice, as in humans and several other mammals (REF), are determined by the binding of a zinc-finger protein encoded by the PRDM9 gene to specific DNA motifs (REFS). There is evidence that PRDM9 has undergone numerous bouts of positive selection (Kono et al 201X) and natural populations of *M. musculus spp.* harbour numerous PRDM9 alleles corresponding to varying suites of recombination hotspots (Smagulova et al 2015). Overall, there is clear evidence that recombination rates have evolved at broad and fine scales in mice.

In lineages that have experienced recent evolution of the recombination rate landscape the hallmark signature of selection at linked sites, a positive correlation between nucleotide diversity and recombination rate may be obscured. In this paper, we examine how patterns of neutral genetic variability under background selection respond to evolution of the recombination rate. We make use of a previously derived function that describes coalescence times after instantaneous population size change to model the effects of background selection on neutral genetic diversity after a change in the recombination rate. We demonstrate how evolution of the recombination rate at both broad and fine scales may influence the correlation between nucleotide diversity and recombination rate using simulations. Finally, we re-analyse results from Kartje et al (2020) on the correlation natural *Mus musculus domesticus* and find a pattern that suggests recombination rate evolution has obscured the patterns of selection at linked sites in that species.

Results

Evolution of the recombination rate causes a change in the effects of background selection that resembles a change in the effective population size (Figure 1).

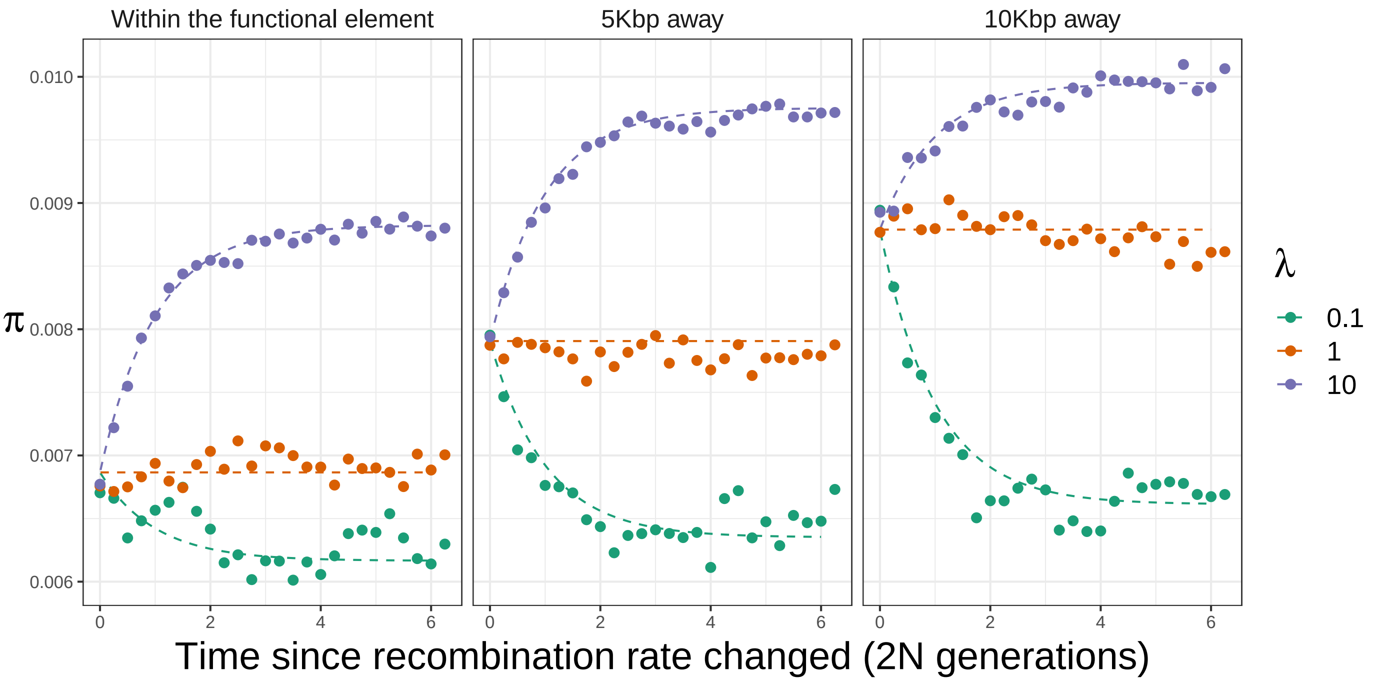

Figure 1 Nucleotide diversity over time after recombination rates change by a factor λ for neutral sites in or around a functional element. The dashed lines were calculated using Equation 2.

An instantaneous change in the recombination landscape may

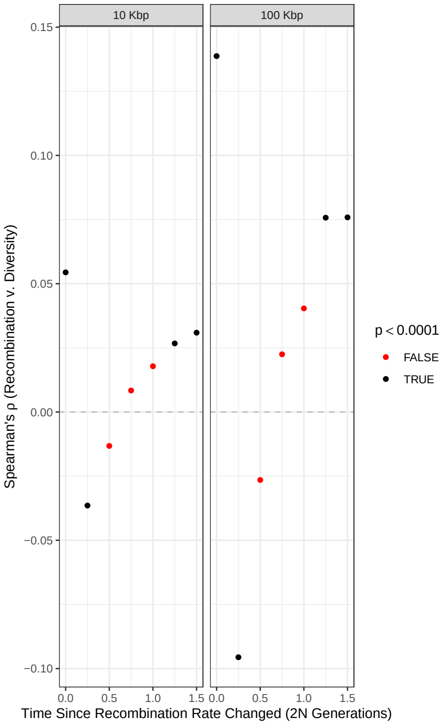

Figure 2 Spearman's correlation between nucleotide diversity (π) and recombination rate (r) over time after recombination rates evolve. Panel A shows results for 10 Kbp, 100 Kbp and 1Mbp analysis windows.

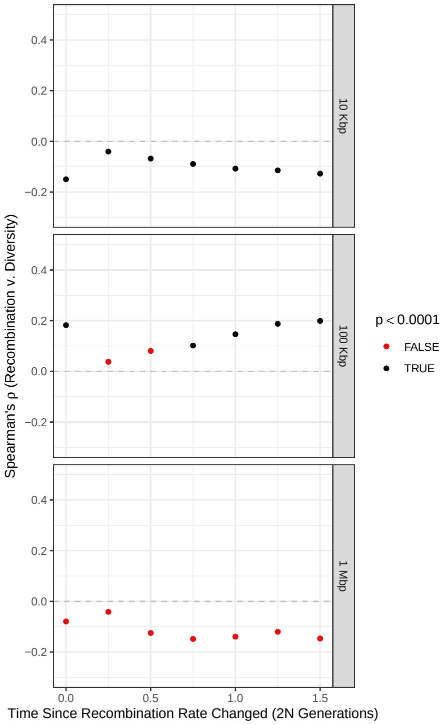

Figure 3 Spearman's correlation between nucleotide diversity (π) and recombination rate (r) over time after recombination rates evolve. Panel A shows results for 10 Kbp, 100 Kbp and 1Mbp analysis windows.

Discussion

Here is an erudite description of the results and what we can infer from them.

For the sake of simplicity, in the model we used and the simulations we performed, we assumed that recombination rates evolve instantaneously. While that is obviously an oversimplication, there is reason to expect that recombination rate changes may evolve rapidly. Chromosomal fusions may exhibit meiotic drive (REFS) and so may rapidly rise to fixation in a population

Cutter and Payseur (2012) noted that numerous species that have been examined do not exhibit a positive correlation between diversity and recombination rate. For example, wild and domesticated rice species (*Oryza rufipogon* and *Oryza sativa*, respectively) exhibit negative correlations between diversity and recombination rate (Flowers et al 2011).

		Whole Genome				Conserved Chromosomes				Non-Conserved Chromosomes			
Window	Population	$\bar{\pi}$	$\sigma(\pi)$	Spearman's ρ	p-value	$\bar{\pi}$	$\sigma(\pi)$	Spearman's ρ	p-value	$\bar{\pi}$	$\sigma(\pi)$	Spearman's ρ	p-value
5Kbp	Gough Island	0.002 22	0.002 67	0.007 67	4.28×10^{-5}	0.002 12	0.002 57	0.008 80	0.0102	0.002 27	0.002 71	0.004 86	0.0302
5Kbp	France	0.002 66	0.002 96	0.004 08	0.0295	0.002 59	0.002 86	0.0403	6.10×10^{-32}	0.002 70	0.003 00	−0.0107	1.76×10^{-6}
5Kbp	Germany	0.002 58	0.002 83	0.007 52	6.05×10^{-5}	0.002 52	0.002 79	0.0152	9.63×10^{-6}	0.002 61	0.002 85	0.003 86	0.0849
1Mbp	Gough Island	0.002 19	0.001 13	0.0536	0.009 46	0.002 07	0.001 10	0.0588	0.124	0.002 24	0.001 14	0.0437	0.0748
1Mbp	France	0.002 64	0.001 29	0.0450	0.0294	0.002 57	0.001 29	0.135	0.000 400	0.002 67	0.001 29	0.009 99	0.684
1Mbp	Germany	0.002 56	0.001 25	0.0535	0.009 53	0.002 49	0.001 27	0.0775	0.0428	0.002 59	0.001 23	0.0426	0.0828

Table 1 The correlation between nucleotide diversity (π) and recombination rate for three populations of house mice (*Mus musculus domesticus*) calculated from all autosomes, conserved chromosomes that exhibit no syntenic breaks between *M. musculus* and *M. pahari* and the non-conserved chromosomes. Conserved and non-conserved chromosomes were identified by Thybert et al (2018)

Cicconardi et al (2020) demonstrated a clear relationship between chromosome size and putatively neutral diversity in *Heliconius* butterflies. Between

We re-analysed the correlation between π and r in *Mus musculus domesticus* using previously analysed data from Kartje et al (2020) (Table 1).

Methods

Model

Background selection has been modelled as the reduction in effective population size (N_e) at a neutral site due to the removal of deleterious variants. The effects of background selection are often expressed as $B = \frac{N_e}{N_0}$, where N_e is the effective population size and N_0 is the expected population size under strict neutrality. In a non-recombining genome, B is proportional to the ratio of the deleterious mutation rate to the strength of selection acting on harmful mutations (Charlesworth et al 1993). For a neutral site present on a recombining chromosome, the effects of background selection depend on the density of functional sites (i.e. those that can mutate to generate deleterious alleles), the mutation rate, the strength of selection and the recombination rate (Hudson and Kaplan 1995; Nordborg et al 1996; Nordborg 1997). For a neutral locus ν linked to x functional sites, the reduction in N_e has been described with the following equation:

$$B_\nu = \frac{N_e}{N_0} = exp[-\sum_x \frac{u_x}{t(1+(1-t)r_{x,\nu}/t)^2}] \tag{1}$$

Where u_x is the deleterious mutation rate at functional site x, t is the heterozygous fitness effect of a deleterious mutation (i.e. $0.5s$ in the case of semi-dominance) and $r_{x,\nu}$ is the recombination map distance between the neutral locus and functional site x. In the above equation, deleterious mutations have fixed effects, but it is straightforward to incorporate a distribution of fitness effects (Nordborg et al 1996). The above equation holds when selection is sufficiently strong such that random drift does not overwhelm selection ($N_e s > 1$) (Good et al 2014).

When the recombination rate landscape evolves it may cause a change in the effects of BGS in particular genomic regions. We modelled the time it takes for coalescence times, or patterns of nucleotide diversity, to reflect background selection expected under a new recombination rate regime using expressions formulated to describe coalescence times after a population size change. For a neutral site ν, the combined effects of recombination, mutation and purifying selection cause there to be a reduction of $B_{\nu,1}$ to coalescence times. At time T_0 in the past (in $2N_e$ generations), the population underwent an instantaneous change in the recombination rate so ν now experiences a BGS effect of $B_{\nu,2}$. We modified an expression for coalescence times after an instantaneous population size change from Johri et al (2020), to obtain the following equation,

$$B_{\nu,\Delta r} = B_{\nu,2}(1 + (\frac{B_{\nu,1}}{B_{\nu,2}} - 1)e^{-T_0}) \tag{2}$$

Note that Equation X from Pool and Nielsen (2009) provided similar expressions to those given in Johri et al (2020).

Simulations

We simulated BGS under recombination rate evolution using two types of simulations in *SLiM* v3.2 (Haller et al 2018). In all cases, diploid populations of N = 1,000 individuals were simulated.

The first set of simulations was designed to examine how long it takes for patterns of neutral diversity under BGS to equilibrate after the recombination rate evolves. In these simulations, the genome was 25Kbp long with a 5Kbp functional element in the centre. Mutations occurred in the functional element at rate $\mu = 2.5 \times 10^{-6}$ and had semi-dominant fitness effects with a fixed selection coefficient of -0.05. We also simulated cases with varying fitness effects using a gamma distribution with mean ($\bar{s}$) of -0.1 and a shape parameter of 0.1. Recombination occurred at a uniform rate of $r = 2.5 \times 10^{-6}$ across the chromosome. After 15,000 generations, we simulated an instantaneous change in the recombination rate, multiplying r by λ, giving $r = \lambda 2.5 \times 10^{-6}$.

We simulated cases with λ = 0.1, 1.0 and 10.0. Simulated populations were sampled every 500 generations after the recombination rate changed and we performed 200 replicates for each set of parameters tested. Note that these simulations were not designed to be particularly realistic, but to provide clear cut patterns to test the theoretical predictions.

The second set of simulations was designed to examine how patterns of π versus r varied over time when recombination rates evolved at fine and/or broad scales. For these simulations, we modelled chromosomes that were 10Mbp long. Deleterious mutations occurred at random across the length of the sequence at a rate of 1×10^{-7} with semi-dominant fitness effects drawn from a gamma distribution with a mean ($\bar{s}$) of -0.1 and a shape parameter of 0.1. Populations evolved under background selection for 15,000 generations

We modelled the evolution of hotspots in the following way. At the beginning of a simulation, a Poisson number of hotspots was sampled with an expectation of 60, which was based on the average number of double-strand break hotspots observed by Smagulova et al (2015). Locations for 10,000bp hotspots were then sampled across the simulated chromosome. Recombination occurred at a uniform rate of $r = 2.08 \times 10^{-7}$ except in hotspots where it occurred at a rate of $r = 2.08 \times 10^{-5}$. These rates were chosen to give an overall recombination rate similar to that of a chromosome that recombined at a uniform rate of $r = 2.5 \times 10^{-6}$.

We modelled recombination rate evolution at broad scales in the following way. The genome-wide average recombination rate in *Mus musculus* is XXX cMMbp. The distribution of recombination rates is approximately normal with mean XXX cMMbp and variance YYY. We scaled recombination in these simulations to simulate a large natural population using a comparatively small number of individuals in *SLiM*. At the beginning of a simulation, we sampled 10 recombination rates from a normal distribution (mean scaled_XXX, variance scaled_YYY). As above, we generated a new recombination map after 15,000 generations of evolution, sampled the population then continued to sample the population every we generated a new set of recombination hotspots as above and sampled the population every 500 generations for a further 3,000 generations.

For all simulations, we used the tree sequence recording option in *SLiM* and neutral mutations were added to the resulting tree-sequences at a rate of 2.5×10^{-6} using PySLiM (version XX). Nucleotide diversity (π) was calculated in windows of varying size using sci-kit-allel (version X.Y, Citation).

Supplementary Material

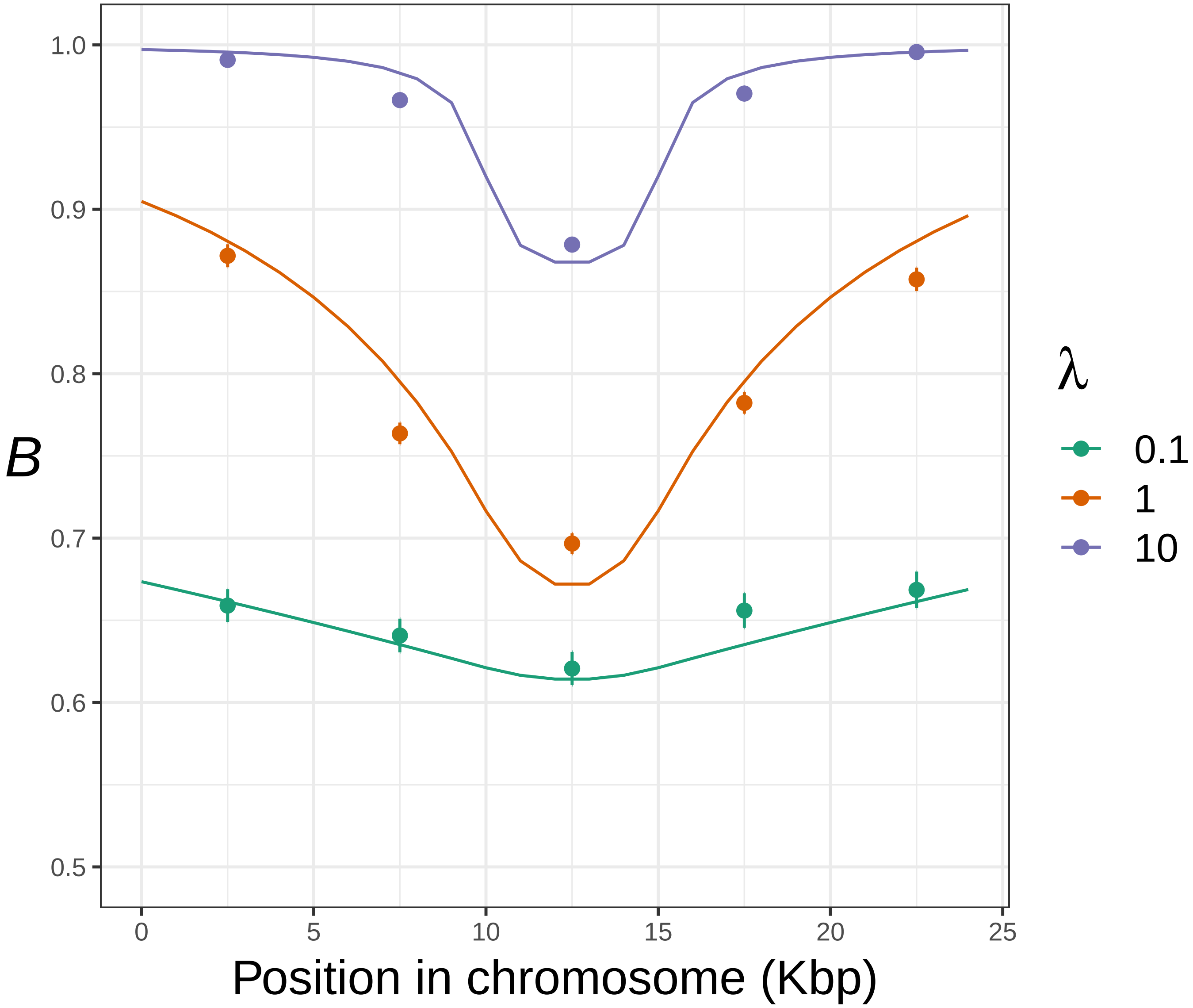

Figure S1 The effects of background selection across simulated chromosomes. B was calculated for simulated data by comparing observed π to the neutral expectation of $4N_e\mu = 0.01$. The lines show the theoretical expectation calculated using formulae from Nordborg et al (1996).