The breakdown of genomic ancestry blocks in hybrid lineages given a finite number of recombination sites

Running title: breakdown of ancestry blocks after hybridization

Thijs Janzen[1,2,*], Arne W. Nolte[1], Arne Traulsen[2]

[1]Carl von Ossietzky University, Carl-von-Ossietzky-Str. 9-11, 26111, Oldenburg, Germany
[2]Max-Planck-Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306, Plön, Germany
∗Carl von Ossietzky University, Carl-von-Ossietzky-Str. 9-11, 26111, Oldenburg, Germany, thijs.janzen@uni-oldenburg.de

Accepted Article

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/evo.13436

This article is protected by copyright. All rights reserved.

Accepted Article

Author Contributions:
TJ, AT and AWN designed the project, TJ and AT developed the mathematical framework, TJ performed the numerical analysis and individual based simulations. TJ wrote the first version of the manuscript upon which AWN and AT improved.

Acknowledgements
We thank S.J.E. Baird for helpful comments and for directing us towards the maximum packing density of junctions and the disparity between chromosomes with even and odd numbers of junctions. Furthermore, we thank N.H. Barton, Linda Odenthal-Hesse and Y. Pichugin for helpful comments and discussion. We acknowledge support through the Max Planck Society. AN was supported through funding from the ERC starting grant EVOLMAPPING. TJ is grateful for use of the computational cluster ADA of the Max Planck Institute for Evolutionary Biology, Pln. The authors declare that there is no confict of interest regarding the publication of this article.

Abstract
When a lineage originates from hybridization genomic blocks of contiguous ancestry from different ancestors are fragmented through genetic recombination. The resulting blocks are delineated by so called junctions, which accumulate with every generation that passes. Modeling the accumulation of ancestry block junctions can elucidate processes and timeframes of genomic admixture. Previous models have not addressed ancestry block dynamics for chromosomes that consist of a finite number of recombination sites. However, genomic data typically consist of informative markers that are interspersed with fragments for which no ancestry information is available. Hence, repeated recombination events may occur between markers, effectively removing existing junctions. Here, we present an analytical treatment of the dynamics of the mean number of junctions over time, taking into account the number of recombination sites per chromosome, population size, genetic map length and the frequency of the ancestral species in the founding hybrid swarm. We describe the expected number of junctions using equidistant molecular markers and estimate the number of junctions using random markers. This extended theory of junctions thus reflects properties of empirical data and can serve to study the genomic patterns following admixture.

This article is protected by copyright. All rights reserved.

Accepted Article

Introduction

Hybridization has long been recognized as a potential driver in the evolution of plants (Grant, 1981) and more recently as a process that generates biological diversity in animals (Abbott et al., 2013). Admixture among previously isolated populations may occur naturally, or result from many facets of human induced ecological change in recent historical time (Taylor et al., 2006; Vonlanthen et al., 2012; Bhat et al., 2014). It seems to be conspicuously associated with the colonization of new or perturbed habitats (Krehenwinkel and Tautz, 2013; Nolte et al., 2005). In these examples, recent admixture has been suspected to contribute to the ecological success of emerging lineages (Nolte and Tautz, 2010). Genetic admixture between differentiated lineages may even lead to hybrid speciation when the joint contribution of both parental species is instrumental in the rise of the new species, for example by creating direct barriers to reproduction with the parental species or by facilitating ecological isolation of the emerging hybrid lineage (Mallet, 2007; Nolte and Tautz, 2010; Abbott et al., 2013; Schumer et al., 2014). The genomic regions that convey a fitness advantage to a hybrid lineage can be expected to be subject to positive selection. On the other hand, we expect the purging of parental genetic variance that reduces the fitness of an admixed lineage (Buerkle et al., 2000; Barton, 2001). Although these studies predicted an initial lag phase during which an emerging hybrid lineage has to go through an evolutionary optimization, empirical studies suggest that hybrid speciation can occur rapidly, possibly within hundreds of generations (Nolte et al., 2005; Buerkle and Rieseberg, 2008; Lamichhaney et al., 2017).

Studies considering hybrid speciation are accumulating in recent years, but even well studied examples remain contentious (Schumer et al., 2014) and a multitude of evolutionary processes related to hybrids are difficult to generalize (Gompert and Buerkle, 2016). On the other hand, hybridization receives great interest from evolutionary biologists and conservationists alike. Hence, there is a need to develop methods that permit extensive comparisons among different study systems to identify shared evolutionary patterns. This includes models that can help to develop neutral evolutionary expectations

This article is protected by copyright. All rights reserved.

Accepted Article

for hybrid lineages (Stemshorn et al., 2011), which in turn permits to identify candidate genomic loci that may be subject to selection. Conventional molecular clock estimates are too coarse to be applied to any cases of rapid speciation, but lineages of hybrid origin hold the potential to estimate rather short time frames from the ancestry structure of admixed genomes (Ungerer et al., 1998; Buerkle and Rieseberg, 2008; Liang and Nielsen, 2014; McTavish and Hillis, 2014). Newly formed F1 hybrids contain complete chromosomes from either one of the ancestral species that constitute blocks (or tracts) of contiguous genomic ancestry. Genetic recombination leads to an exchange of genetic material between homologous chromosomes, which interrupts blocks of contiguous ancestry. As a consequence, the number of blocks accumulates while their size decreases from generation to generation. Modeling the accumulation of ancestry block junctions can elucidate processes and timeframes of genomic admixture. Since this process begins after the first generation of admixture, it holds the potential to study even the initial evolutionary steps that are of particular interest to study how a hybrid lineage evolves (Nolte and Tautz, 2010).

The study of genomic blocks dates back to Fisher, who recognized that genetic material is organized within contiguous haplotype blocks after a hybridization event (Fisher, 1949, 1954). He termed the delineations between these blocks junctions, and established that junctions have inheritance properties similar to those of point mutations: over time they are either lost from the population, or they become fixed. Fisher formulated the expected number of junctions given the number of generations passed since the initial hybridization event, for the case of sib-sib mating (Fisher, 1954). Following Fisher, the theory of junctions was quickly extended towards self-fertilization (Bennett, 1953), alternate parent-offspring mating (Fisher, 1959; Gale, 1964) and a population of randomly mating individuals (Stam, 1980). Apart from the number of junctions per chromosome, the distribution of block sizes has proven to be highly informative as well. Building upon the multilocus clines work of Barton (1983), which describes the block size distribution at equilibrium, Baird developed a robust framework describing the full dynamics of the block size distribution (Baird, 1995). Furthermore, using efficient simulation techniques, Baird showed the impact of selection on the distribution of haplotype blocks, and used this framework to infer the onset of hybridization in Helianthus Sunflowers (Ungerer et al., 1998).

Parallel to the development of the theory of junctions, theory has been developed to describe the genomic contribution of past migrants (Gravel,

This article is protected by copyright. All rights reserved.

Accepted Article

2012; Pool and Nielsen, 2009; Liang and Nielsen, 2014). Here, the focus is on the breakdown of large ancestry blocks (within this context often referred to as 'admixture tracts') introduced by migrants, either due to a single migratory event (Gravel, 2012) or due to ongoing migration (Pool and Nielsen, 2009). These studies illustrate that allele frequencies affect the formation of junctions. When only a few, rare blocks (depending on the rate of migration) are introduced into a foreign background the decay of these blocks is asymmetric as it depends on how often they may recombine with blocks that have a shared (here rare) ancestry. This line of research is especially applicable to the study of the history of humans, where migration between for instance European and African populations can be detected from genomic patterns (Hellenthal et al., 2014; Payseur and Rieseberg, 2016). Whereas the theory of junctions often focuses on a two-species mixture (but see (Baird et al., 2003)), the study of admixture tracts allows the application to be broadened to multiple migratory source populations.

The theory of junctions has focused on a starting population inspired by the genome of F1 hybrid individuals that contain equal proportions of the parental genetic material. Although the work of Fisher (Fisher, 1949, 1954) and subsequent extensions by Bennett (1953), Stam (1980), Chapman and Thompson (2002; 2003) and MacLeod et al. (2005) allows for deviations of this situation by changes to the initial heterozygosity H_0 at time $t = 0$, typical analysis has focused on the idealized situation of $H_0 = 0.5$. Such equal proportions are only expected if a hybrid lineage is founded by a sufficiently large population of F1 hybrids. Under natural conditions, however, a hybrid lineage of outcrossing organisms more likely emerges from a heterogeneous hybrid swarm, and the overall ancestry contribution of parental species to the founding hybrid swarm may differ (Edmands et al., 2005; Nolte and Tautz, 2010; Stemshorn et al., 2011; Schumer et al., 2016). Deviations from equal founding ancestry proportions and strong stochastic or fitness effects during the first few generations can sway the ancestry of an emerging hybrid lineage in favor of one of the ancestral species. Strong deviations from equality are likely to affect the formation of junctions as they decrease the frequency at which genotypes comprise material from both species. Moreover, the effect of drift is exacerbated when the ancestry contribution is strongly deviating from equal proportions. The impact of deviations from an even ancestry contribution of both ancestral species and the interaction with population size and drift remains understudied so far.

Paramount to applying the theory of junctions, is accurate detection of

This article is protected by copyright. All rights reserved.

Accepted Article

these junctions. MacLeod *et al.* 2005 showed that detection success was dependent on two crucial factors: the distribution of markers, and the density of markers (MacLeod et al., 2005). Macleod *et al.* compared markers that were equidistantly distributed with markers that were randomly distributed over the genome, whilst keeping the average density constant. Although equidistantly distributed markers did not detect 100% of all junctions, they detected 10-20% more junctions than randomly spaced markers. Furthermore, Macleod *et al.* found that the required density of markers to detect 90% of all junctions had to be at least 12.5 times the number of junctions. Because the expected number of junctions scales with the number of generations since the onset of hybridization, even for intermediate time-scales between 500 and 1000 generations, the number of markers required to accurately infer the number of junctions quickly escalates towards extremely high numbers. To circumvent this problem, Buerkle and Rieseberg (2008) used simulations to apply a correction to observed junction densities in order to infer the true junction density. Our aim is to extend the theory of junctions in order to better include the effect of a finite number of markers.

Using only a finite number of markers leaves gaps between markers that remain understudied. So far, the theory of junctions has assumed that recombination never occurs twice at the same location (e.g. there is an infinite number of recombination sites along the chromosome) (Fisher, 1954; Stam, 1980; Chapman and Thompson, 2003) and, accordingly, that all recombination events are detected. However, recent work has shown that such a distribution of the recombination rate might be the exception, rather than the rule, and that the recombination landscape is often organized into "warm" and "cold" areas, including hot spots (Gerton et al., 2000; Myers et al., 2005; Mackiewicz et al., 2013; Singhal et al., 2015; Arbeithuber et al., 2015; Smagulova et al., 2011). Repeated crossovers in the same area can impact the formation of new junctions, and can cause existing junctions to disappear. The same effect may manifest in genomic data even in the absence of recombination hotspots. Across the genome, ancestry informative markers are typically interspersed with fragments for which no ancestry information is available. As a result, recombination may occur repeatedly between markers. Hence, inclusion of recurrent recombination into the theory of junctions could add to our understanding of recombination hot spots and accommodates effects of incomplete information in studies using molecular markers..

This warrants methods that can be applied broadly in empirical studies on the consequences of admixture. Here, we present an extension to the

This article is protected by copyright. All rights reserved.

theory of junctions including a finite number of recombination sites, and for hybrid swarms with an arbitrary contribution of either parental species. We first extend the theory of junctions towards a finite number of recombination sites in infinite populations, and then extend this framework towards finite populations. Then, using a novel approach, we derive a generalized theory of junctions, which only depends on the number of junctions obtained at equilibrium (e.g. at $t = \infty$). Lastly, using individual based simulations, we demonstrate the validity of our framework, and explore the implications of marker distributions on junction detection.

Analytical model

Our aim is to derive the expected number of junctions per chromosome, in an isolated hybrid lineage depending on the time since the onset of hybridization. A scenario in which a single mating event and the resulting F1 hybrids are the sole founders of a hybrid lineage seems to be too constrained to represent the breadth of results from empirical studies. Conversely, we assume that a hybrid lineage emerges from a hybrid swarm. Individuals representing parental species or backcrossed individuals may become part of the hybrid swarm and bias the overall ancestry proportions of the founding population. As such, the hybrid swarm approach allows us to study situations in which there are deviations from a 50-50 distribution of ancestral genomic material at the onset of a hybrid lineage. Note that the special case of a founding population of only F1 individuals constitutes a special case of the more permissive scenario studied here. We assume full knowledge of ancestry along the genome. For simplicity, we ignore selection and drift and study the dynamics of junctions in neutrality. We assume that a Poisson number mean C crossover events occurs per chromosome per meiosis, which corresponds to the assumption that chromosomes are C Morgan long. The recombination rate is assumed to be uniform across the chromosome. Each individual is diploid. Orthologous chromosome pairs are interchangeable and paralogous chromosome pairs are inherited independently, which allows us to track junctions within only one chromosome pair, rather than all pairs simultaneously. Furthermore, we assume that populations are in Hardy-Weinberg Equilibrium.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

An infinite number of recombination sites

We first assume infinite population size and an infinite number of recombination sites along the chromosome. We denote the frequency of genomic material of parental species P by p, and the frequency of genomic material of the other parental species Q by q, where $p = 1 - q$. The initial average heterogenicity across the genome is then defined by: $H_0 = 2pq$, where H_t is the heterogenicity at time t. Heterogenicity here refers to the mean proportion of the genome heterozygous by source, e.g. stemming from different parental species (sensu (Fisher, 1954, 1959; Stam, 1980)). If we would substitute all genomic content from species P by allele A, and all genomic content from species Q by allele a, it follows that the mean heterogenicity is equivalent to the mean heterozygosity across the genome. In our following derivations it turns out that we can use known expressions for the change in mean heterozygosity, which makes it important to realize that here, the mean heterozygosity is equivalent to the mean heterogenicity.

During crossover, a junction is only formed if crossover takes place at a site that is heterozygous for the parental genomic content. Thus, in an infinite population with an infinite number of recombination sites along the chromosome, the average number of junctions per chromosome, after t generations, is given by (Chapman and Thompson, 2002; MacLeod et al., 2005; Buerkle and Rieseberg, 2008)

$$J_t = \sum_{i=0}^{t-1} H_i C = H_0 C t, \tag{1}$$

where $H_i = H_0$ is the proportion of the heterozygous genomic material (following (Fisher, 1954; MacLeod et al., 2005; Buerkle and Rieseberg, 2008)) in an infinite population, where the proportion of heterozyous genomic material does not change over time. In a finite population, the average heterozygosity changes over time (Crow and Kimura, 1970) as

$$H_t = H_0 \prod_{j=0}^{t-1} \left(1 - \frac{1}{2N_j}\right). \tag{2}$$

Assuming a constant population size over time, $N_t = N$ for all t, and substituting H_i in Eq. (1) by Eq. (2), we obtain an expression for the average

This article is protected by copyright. All rights reserved.

number of junctions at time t, in a finite population (Chapman and Thompson, 2002)

$$J_t = H_0 C \sum_{i=0}^{t-1} \left(1 - \frac{1}{2N}\right)^i$$
$$= 2H_0CN - 2H_0CN \left(1 - \frac{1}{2N}\right)^t. \quad (3)$$

In the limit of $N \to \infty$ we recover Equation (1). For $t \to \infty$, we have (MacLeod et al., 2005)

$$J_\infty = 2H_0CN. \quad (4)$$

Thus, for finite N the expected number of blocks in the descendant hybrid lineage converges to a finite number determined by the population size, the size of the chromosome in Morgan, and the initial heterozygosity (which in turn depends on the frequency of the ancestral species in the hybrid swarm) .

A finite number of recombination spots

In the previous section, we have assumed that recombination never occurs twice at the same spot. In reality, a chromosome can not be indefinitely divided into smaller parts. Furthermore, the presence of recombination hot spots (Gerton et al., 2000; Myers et al., 2005; Mackiewicz et al., 2013; Singhal et al., 2015; Arbeithuber et al., 2015; Smagulova et al., 2011) indicates that recombination in fact often does occur multiple times at the same site. We therefore proceed to study the change in number of junctions in a chromosome consisting of $R+1$ different genomic segments, where each segment represents a minimal genomic element that cannot be broken down further due to recombination. The most general interpretation would be a genomic area delineated by two genetic markers. Considering a chromosome of $R+1$ genomic segments, there are R possible junction sites (we do not distinguish the process of junction formation at different recombination positions, we validate this assumption using individual based simulations, see the section "Individual Based Simulations"). We assume that the $R+1$ genomic segments are of equal size, and as a result, the R junction sites are uniformly spaced across the chromosome (in Section "A finite number of markers" we relax this assumption). Given that at time t, there are J_t junctions, the

Accepted Article

This article is protected by copyright. All rights reserved.

probability that a recombination event takes place at an existing junction is then given by

$$\alpha_t = \frac{J_t}{R}.$$

We focus on the dynamics of one of the two produced chromosomes. Because the choice of chromosome is random, averaging across a large number of recombinations ensures that we cover all possible outcomes. During recombination, we can then distinguish four possible events, taking into account the location of recombination on both chromosomes (Figure 1):

(A) **With probability α^2, recombination takes place on an existing junction on both chromosomes.** In this case, there are two possible outcomes, depending on the transitions that the two junctions represent. Either there is no change in the number of junctions (when the transitions of the two junctions are identical), or a decrease in the number of junctions (when the transitions of the two junctions are of opposing type). The probability of either event happening is $\frac{1}{2}$, yielding an average change in the number of junctions when crossover takes place on an existing junction on both chromosomes of $-\frac{1}{2}$.

(B) **With probability $\alpha(1-\alpha)$, recombination takes place on an existing junction on one chromosome, and within a block on the other chromosome.** There are two possibilities: either the block on the other chromosome is of the same type as the genomic material before the existing junction, or it is of the other type. If it is of the same type, the existing junction disappears, and the number of junctions decreases by one. If it is of the other type, the existing junction remains and the number of junctions does not change. The probability of either event happening is $\frac{1}{2}$, and hence we expect the number of junctions on average to change by $-\frac{1}{2}$.

(C) **With probability $(1-\alpha)\alpha$, recombination takes place on within a block on one chromosome, and on an existing junction on the other chromosome.** The outcome is exactly the opposite of case (B). If the genetic material after the junction on the second chromosome is of the same type as the block on the first chromosome, no new junction is formed and the number of junctions stays the same. If the genetic material after the junction on the second chromosome is of a different type than that of the block on the first chromosome, a new junction is

Accepted Article

This article is protected by copyright. All rights reserved.

formed and the number of junctions increases by one. The probability of either event happening is $\frac{1}{2}$, and hence we expect the number of junctions on average to change by $\frac{1}{2}$.

(D) **With probability $(1-\alpha)^2$, recombination takes place within a block on both chromosomes.** In this case, matters proceed as described for the continuous chromosome: with probability H_0 we observe an increase in the number of junctions. However, since we are dealing with a finite number of recombination positions along the chromosome, the frequency of recombination sites of a genomic type is no longer directly related to H_0. If there would be no blocks, i.e. if the genomic material would be distributed in an uncorrelated way, pR potential recombination sites are of type P, that is, they are within a block of type P. Similarly, qR potential recombination sites are within a block of type Q. As new junctions are formed, the number of potential recombination sites that are still within a block decreases. With the formation of a new block, on average both a recombination site within a block of type P and a recombination site within a block of type Q are lost, such that on average, after the formation of a new junction, the number of recombination sites of type P decreases by $\frac{1}{2}J_t$. Thus the number of recombination sites within a block of type P is $pR - \frac{J_t}{2}$. Similarly, the number of recombination sites within a block of type Q is $qR - \frac{J_t}{2}$. The probability of selecting a recombination site within a block of type P is then given by the number of recombination sites of type P divided by the total number of recombination sites. Let us denote the probability of selecting a recombination site of type P by p^*, which is then given by:

$$p_t^* = \frac{pR - \frac{1}{2}J_t}{pR - \frac{1}{2}J_t + qR - \frac{1}{2}J_t} = \frac{pR - \frac{1}{2}J_t}{R - J_t}. \tag{5}$$

Concluding, for scenario D, we observe that the number of junctions increases on average by $2p_t^*q_t^*$ (where $q^* = 1 - p^*$).

Accepted Article

This article is protected by copyright. All rights reserved.

Combining the scenarios (A)-(D), and noticing that $2pq = H_0$, we can formulate the total expected change in number of junctions

$$\begin{aligned} J_{t+1} &= J_t + C\left(-\frac{1}{2}\alpha_t^2 + \frac{1}{2}\alpha_t(1-\alpha_t) - \frac{1}{2}\alpha_t(1-\alpha_t) + 2p_t^*q_t^*(1-\alpha_t)^2\right) \\ &= J_t + C\left(2p_t^*q_t^*(1-\alpha_t)^2 - \frac{1}{2}\alpha_t^2\right) \\ &= J_t + H_0C - C\frac{J_t}{R} \end{aligned} \tag{6}$$

The solution of recursion Equation (6) is given by

$$J_t = H_0R - H_0R\left(1 - \frac{C}{R}\right)^t \tag{7}$$

which is reminiscent of Equation (3). The exponentially decaying term leads to convergence at $t \to \infty$, where we obtain

$$J_\infty = H_0R. \tag{8}$$

Note that the number of junctions obtained at $t \to \infty$ does not depend on the size of the chromosome in Morgans C, but the speed of convergence does. A Taylor expansion of Eq. (7) at $t = 0$ shows that initially, the number of blocks increases linearly

$$J_t \approx -H_0R\ln\left(1 - \frac{C}{R}\right)t.$$

Once the number of junctions has reached the limit H_0R, new junctions will still be formed. Nevertheless, the average number of junctions does not increase further, because the population has reached the maximum packing density of junctions. Another approach to understanding the link between genome size and the maximum packing density of junctions is the following: As $t \to \infty$, all alleles are in linkage equilibrium, and we can consider the probability of observing either allele P or Q as independent along the chromosome. Then, assuming that each allele has a probability p of being P, the sequence consists of $R+1$ independent Bernouilli trials, and the number of consecutive P sites (and Q sites) is given by a Negative Binomial distribution parameterized with $r = 1$ (the number of successes before 1 failure) and p. The full haplotype sequence then consists of alternating sequences of

Accepted Article

This article is protected by copyright. All rights reserved.

NB(1,p) sites of type P and NB(1,q) sites of type Q. The mean length of junctions intervals is then given by:

$$\frac{E[NB(1,p)+1]}{2} + \frac{E[NB(1,q)+1]}{2} = \frac{1}{2pq}$$

Then, the average number of junctions packed in a chromosome of $R+1$ length, is given by $J_\infty = 2pqR = H_0R$, which is identical to our previous result in Equation (8). Hence, whereas initially accumulation of junctions is hampered by repeated recombination in the same site, at equilibrium it is the maximum packing density of junctions, rather than the dynamic equilibrium between formation and removal of junctions, that determines the maximum number of junctions across the population.

A finite number of recombination spots in a finite population

Arguably the most realistic scenario involves a finite number of recombination sites, within a finite population. For a finite chromosome in a finite population, accumulation of junctions is limited by two processes: decay of the heterozygous proportion of the genome due to drift in a finite population, and the probability of junction removal due to recombination occuring at a previous recombination site. We have described limitation of junction accumulation due to repeated recombination at the same site as a recurrence equation (cf. Equation (6)). We can similarly express the limitation of junction accumulation due to the decrease in the proportion of the genome heterozygous due to finite population size (Eq. (3)) as

$$J_{t+1} = J_t + H_0C - \frac{J_t}{2N}. \tag{9}$$

In Equation (6) the accumulation of junctions over time is slowed down by the term $C\frac{J_t}{R}$, representing the slowdown of junction accumulation due to repeated recombination at the samesite. For a finite population we observe a similar pattern, where the accumulation of junctions is slowed down by the term $\frac{J_t}{2N}$, which represents the decay of the portion of genome heterozygous. Assuming that these two effects are independent, and only focusing on mean junction dynamics (ignoring drift in p) (we will show with individual based simulations that these assumptions are good approximations), the combined

Accepted Article

This article is protected by copyright. All rights reserved.

effect of a finite population and of a finite number of recombination sites, is then given by

$$J_{t+1} = J_t + H_0 C - \frac{J_t}{2N} - C\frac{J_t}{R}. \tag{10}$$

The solution to equation (10) is given by

$$J_t = H_0 C \frac{2NR}{2NC+R} - H_0 C \frac{2NR}{2NC+R}\left(1 - \frac{1}{2N} - \frac{C}{R}\right)^t, \tag{11}$$

the exponentially decaying term leads to a convergence at $t \to \infty$, where we obtain

$$J_\infty = H_0 C \frac{2NR}{2NC+R}. \tag{12}$$

For $R \to \infty$, Equation (12) simplifies to Equation (4), and for $N \to \infty$ Equation (12) simplifies to Equation (8).

Generalized Junction Dynamics

The general pattern of the accumulation of junctions over time is highly similar across the different scenarios we have studied here: after an initial period of a strong increase in the number of junctions, the increase in the number of junctions slows down and eventually approaches a maximum. Furthermore, comparing equations (3), (7) and (11) we observe that there are some generalities. Generally speaking, we can express the number of junctions at time t as a function of the maximum number obtained at $t \to \infty$. We find (a full derivation can be found in the Appendix)

$$J_t = K - K\left(1 - \frac{H_0 C}{K}\right)^t. \tag{13}$$

Where K is the maximum number of blocks obtained at $t \to \infty$. the derivation is graphically summarized in Figure 2: although for different values of N and R, junction accumulation curves differ strongly (Figure 2 A), we find that after normalizing the curves by the number of junctions obtained at $t \to \infty$ (Figure 2 B), the curves partially collapse on each other. After rescaling the time axis by the initial slope of each curve, all curves collapse on each other (Figure 2 C), and are described by Equation (13). In the most general case, K is then given by Equation (12).

Accepted Article

This article is protected by copyright. All rights reserved.

The limit of accuracy

Given knowledge of the number of junctions and the population size, one can use information about the number of junctions to infer the onset of hybridization, which is given by (cf. Eq. (11))

$$t = \frac{\log\left(1 - \frac{J}{K}\right)}{\log\left(1 - \frac{1}{2N} - \frac{C}{R}\right)} \tag{14}$$

In contrast to using a molecular clock, information about the number of junctions provides information on a relatively short timescale. However, because the number of junctions plateaus over time, there is a limit in the accuracy of this method. The maximum accurate time τ_{MAT} for which one can still infer the onset of hybridization using equation (14), is given by (the full derivation can be found in the Supplementary Material):

$$\tau_{\text{MAT}} = \frac{\log\left(K^{-1}\right)}{\log\left(1 - \frac{1}{2N} - \frac{C}{R}\right)}, \tag{15}$$

where K is given by Equation (12). Figure 3 shows that τ_{MAT} scales roughly exponential with both population size and chromosome size.

Individual based simulations

To verify our analytical framework, and test the impact of marker distributions, we test our findings using an individual based model. We use a Wright-Fisher process, extended with recombination, assuming a constant population size N, diploid individuals, and a uniform recombination rate across the genome. During initialization of the model, N individuals are generated, where each individual can have either two parents of type P (with probability p^2), two parents of type Q (with probability q^2) or one parent of type P and one parent of type Q (with probability $2pq = H_0$). In every consecutive time step, N new individuals are produced, where each individual is the product of a reproduction event between two individuals from the previous generation. Parental individuals are drawn with replacement, such that one individual could reproduce multiple times, but will on average reproduce one time. We assume that in a mating event, both parents produce a large number of haploid gametes from which two gametes (one from each parent)

Accepted Article

This article is protected by copyright. All rights reserved.

are chosen to form the new offspring. During production of the gametes, the number of recombination sites is drawn from a Poisson distribution with a mean of C.

An infinite number of recombination sites

To model an infinite number of recombination sites, we represent each chromosome as a continuous line of arbitrary length, that can be subdivided into an infinite number of smaller lines. We only keep track of junctions delineating the end of a block (following Baird (1995); Ungerer et al. (1998)). For each junction, we record the position along the chromosome and whether the transition is from genetic material of type $P \rightarrow Q$ or $Q \rightarrow P$. We observe that over time, the number of junctions reaches a maximum value, but only if the population is finite (Figure 4), in line with our mathematical predictions. In the first few generations the accumulation of junctions follows that of an infinite population (dashed line in left panel in Figure 4), but rapidly simulation results start deviating from the infinite population dynamics. When one of the two ancestral species is overrepresented in the initial hybrid swarm (e.g. $H_0 = 0.1$), the maximum number of junctions is lower (as expected, following Eq. (4), and is reached within a shorter timespan.

A finite number of recombination sites

To represent a chromosome consisting of a finite number of genomic elements, we model the chromosome as a bitstring, where a 0 indicates a chromosomal segment of type P, and a 1 indicates a chromosomal segment of type Q. To approximate an infinite population we use a populationsize of 100,000 (simulation results using a population size of 100,000 are very close to our predictions assuming an infinite population size). The mean number of junctions in the stochastic simulations closely follow our analytical estimates (Figure 4, middle column), for all chromosome lengths considered here. Again we observe that for strongly skewed ancestral proportions ($H_0 = 0.1$), the maximum number of junctions is lower (as expected following Equation (8)), and the maximum is reached within a shorter timeframe. When we simulate using a finite population size and a finite number of genomic elements (Figure 4, right column), we observe similar patterns, where the number of junctions approaches a maximum value, albeit that this maximum value is lower than for an infinite population. Furthermore we observe that mean results of sim-

Accepted Article

This article is protected by copyright. All rights reserved.

ulations are highly similar to our analytical expectations following equation (11). We have shown here only results for $R \leq 2NC$, which ensures that limitation in the accumulation of junctions is the result of both the effect of a finite chromosome size, and of a finite population size (see Equation (29) in the Appendix). For $R \gg 2NC$, our results reduce to the infinite chromosome case.

The distribution of junctions in the population

So far, we have focused on the mean number of junctions within the population. Apart from the mean, the distribution of junctions within the population has some interesting properties as well. Most interestingly, the number of chromosomes with an even and with an odd number of junctions is not distributed equally. We find that for increasing deviations from $H_0 = 0.5$, the frequency distribution of junctions starts to resemble more and more a "stegosaurus" pattern, where the high peaks are associated with chromosomes with an even number of junctions, and the low peaks (the back of the stegosaurus) are associated with chromosomes with an odd number of junctions (See Figure 5 A). Furthermore, we find that the degree of "stegosaurusness", e.g. the relative frequency of chromosomes with an even number of junctions, is directly proportional to p (Figure 5 B). The number of even and uneven junctions could therefore potentially be used as an independent estimator of p. However, empirically estimating the exact number of junctions per chromosome is hard, and errors in estimating the number of junctions smoothens out the distribution, making the relative frequency of even chromosomes an unreliable predictor of p. Furthermore, due to this smoothing, for empirical applications it seems better to focus on the mean of the distribution (as described in the previous and next section), rather than focus on the peculiarities of this distribution. Lastly, although the disparity between chromosomes containing even and uneven numbers of junctions could introduce errors when estimating the mean number of junctions at t, we do not find such errors (See Figure 5 C), most likely because we focus here on the mean, rather than variation round the mean. This shows that although from a mathematical viewpoint interesting patterns appear, for practical purposes it is possible to ignore these.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

A finite number of markers

A finite number of recombination sites can also be interpreted as a number of markers uniformly spaced across the genome. In reality, however, markers are seldom placed uniformly spaced across the genome. In this section we numerically explore the impact of having randomly spaced markers along the genome. We simulated markers by simulating a chromosome consisting of an infinite number of recombination sites (as described above), and then superimposing a fixed (random) marker distribution. For each marker position, we assessed the ancestral state by checking the transition direction at the junctions surrounding the marker position.

We performed individual based simulations for $N = 100, 500$ and 1000, $H_0 = 0.5, C = 1$, and explored marker numbers $R = 50, 100, 500, 1000$ and 5000. Per parameter combination we simulated 100 random marker distributions, and per marker distribution, 100 replicate simulations. We find (Figure 6 A & D, dots) that the number of detected junctions is lower than expected using K based on evenly spaced markers (Equation (12))(Figure 6, A,D, dashed line). Because K is the number of junctions obtained in the limit $t \to \infty$, we can substitute the maximum number of junctions obtained at the end of the simulations (e.g. a numerical estimate for K), and plot the expected number of junctions over time following the generalized framework (Equation (13)) (Figure 6, A & D, dashed lines). We find that the simulation data follows the expected number of junctions reasonably well. Nevertheless, it appears that in the initial generations, the detected number of junctions is lower than that expected under this adjusted framework (Figure 6 B & E). Indeed, we find (Figure 6, B & E) that the detected number of junctions divided by the number of junctions expected using the value of K estimated from the simulations, tends to one as t tends to large values (as expected), but is lower in the initial stages of hybridization, and can be as low as 0.8. Nevertheless, we find that the observed values of K (e.g., the number of junctions at the end of the simulation), and those expected for the respective R values, correlate strongly (Figure 6, right panel). For $N = 100$ we find a slope of 0.97 ($R^2 > 0.99$), for $N = 500$ we find a slope of 0.88 ($R^2 > 0.99$, figure not shown) and for $N = 1000$ we find a slope of 0.84 ($R^2 > 0.99$). Hence, we find that the difference between the expected and observed value of K increases with increasing N. This seems to suggest an interaction between population size and the number of randomly distributed markers that we cannot account for. If the effect of randomly distributed markers follows

This article is protected by copyright. All rights reserved.

generalized underlying dynamics, we expect to be able to rescale obtained junction accumulation curves in a similar fashion as shown in Figure 2 and in Section "Generalized Junction Dynamics", e.g. by first rescaling the number of junctions by the obtained maximum, and then rescaling time by the initial slope. We find that after rescaling, the curves do not line up (results in the Supplementary Material), suggesting that there are effects affecting the number of detected junctions other than a fixed penalty of junction detection. Nevertheless, as R increases, the approximation of an even marker distribution becomes increasingly accurate, which is demonstrated by the close fit of $R = 5000$, $N = 100$, as shown in Figure 6. This might also reflect the decaying impact of a finite number of markers, which suggests that the impact of a finite number of markers is less important if $R \gg 2N$ (derivation in the Appendix).

Discussion

Although the theory of junctions dates back to Fischer Fisher (1949) and has been developed towards modern analysis of hybrid zones (Barton, 1983; Baird, 1995) empirical studies on the evolution of junctions in admixed lineages are still scarce. This may be partly owed to the fact that genomic data and relevant study systems are only starting to become available. Moreover, existing methods were challenging to use. Whereas previous work on the theory of junctions assumes a chromosome that consists of an infinite number of recombination sites, we take into account a finite number of recombination sites, and a non-zero probability of recurrent recombination events (Equation (7)). This accommodates study systems where the recombination landscape is not homogenous and where the number of markers that are informative of ancestry is limited. We developed a novel framework that describes the accumulation of junctions over time as a generalized function only dependent on the number of junctions obtained at $t = \infty$ (Equation (13)). Furthermore, we have derived general expressions that provide the upper limit for inferring the age of hybridization, given the number of molecular markers used (R) and the population size (N). Lastly, we have used this general framework to study the effect of randomly distributed markers as compared to evenly distributed markers.

Our framework can serve as a neutral model of the accumulation of ancestry junctions in studies that seek to analyze the early stages of hybrid

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

speciation and recent admixture, which are particularly informative of evolutionary processes (Nolte and Tautz, 2010; Gompert and Buerkle, 2016). The junction framework that we have presented provides a neutral expectation but it considers only certain effects of drift. Drift causes allele frequencies to change over time. Initially these changes are coupled among linked markers, but they become more and more independent through consecutive rounds of recombination. As a result, within genome variation in allele frequencies increases over time (Stemshorn et al., 2011) which might impact junction formation. For simplicity, we have not taken into account the effect of inter and intra-chromosomal variation in allele frequencies on junction formation, and have only modeled the effects of population size on drift, as described by Crow and Kimura Crow and Kimura (1970). However, effects of recombination on drift were explicitly modeled in the individual based simulations. It is reassuring that the mean dynamics of the simulations agree well with our mathematical model (Figure 6). Nevertheless, we acknowledge that for a complete and full description of neutral junction dynamics, the effect of within genome variation in allele frequencies on junction formation should be taken into account.

Here we summarize properties of our method and we discuss processes that may cause deviations from the model that may be relevant to interpret results. Our extension of the theory of junctions accurately describes the expected number of detected junctions, provided that the markers are regularly spaced across the chromosome. Unfortunately, regular marker spacing can be difficult to obtain in reality. We have therefore simulated the effect of randomly spaced markers, whilst keeping the overall marker density across the chromosome fixed. We found that the amount of junctions that remained undetected varied over time, and that applying a fixed ratio to correct the expected number of junctions did not resolve this issue. In the absence of an explicit description of the number of detected junctions over time, we have shown how to obtain a numerical approximation by using simulations to estimate the value of K within our generalized framework. Although this approximation still tends to overestimate the number of junctions at intermediate timescales, estimates are much closer than predictions from the evenly spaced markers model. The overestimation of the approximation seems to be the result of an interaction between the number of markers, the number of junctions and the population size. This is in line with previous findings by MacLeod et al. (2005), who found that the ratio between markers and junctions (e.g. R/J_t) was the main factor driving the detection probability.

This article is protected by copyright. All rights reserved.

Accepted Article

We would like to note, that the mismatch between our predictions and the observed number of junctions due to the difference between evenly and randomly spaced markers reduces significantly as the number of markers (R) increases (see Figure 6). Note also, that although it has become increasingly feasible to analyze large numbers of markers, alleles in empirical data are often not fully ancestry informative. The ancestry of genomic blocks can still be determined with a degree of uncertainty (see for instance Corbett-Detig and Nielsen (2017)), but problems arising from ancestry uncertainty are outside the scope of the current work.

Processes that cause deviations from the neutral model fall into two general categories: either processes acting upon the underlying genomic content, or processes that affect population dynamics. When a genomic region from only one parent is under selection this reduces heterozygosity and therefore recombination events in this region, as individuals recombined in this region are selected against (Kimura, 1956; Lewontin and Hull, 1967). Hence, selection can slow down the formation of new junctions. Future work could focus on the minimal level of selection to offset neutral junction dynamics, or whether deviations from neutral junction dynamics can be used to identify genomic areas that are under selection. However, caution should be taken as the signal of selection may be lost due to overshadowing effects of drift due to limited population size, or changes in population demography (see also below).

Chromosomes share the same evolutionary history but constitute separate genetic elements. Hence, it would be conceivable that a comparison of the number of junctions could reveal interesting differences at the level of chromosomes. Under neutral dynamics, the number of junctions per chromosome is expected to be approximately Poisson distributed. Although we have a firm understanding of the expected mean of this distribution, an understanding of the expected variation around the mean has so far only been accessible through simulation (Chapman and Thompson, 2002, 2003). Using our simulations, we have also explored how variation in the number of junctions changes over time (see Supplementary Material). It appears that the variation does not follow straightforward dynamics, especially if the population is finite in size. Furthermore, the variation appears unrelated to the mean, suggesting that a Poisson approximation is invalid (see also (Chapman and Thompson, 2002, 2003)). We observe that variation increases initially until a tipping point is reached, where the increase in variation due to the creation of new junctions is offset by the reduction in variation due to fixation

This article is protected by copyright. All rights reserved.

as a result of finite population size. After this point, variation decreases until all junctions are fixed in the population. The exact mathematical description of the expected variance of the number of junctions among chromosomes currently lies outside our grasp.

A non-uniform recombination rate across the physical chromosome could also cause deviations from neutral junction dynamics. Empirical work has shown that recombination rates are often not equal across the chromosome, but may be increased towards the peripheral ends of the chromosome (Lukaszewski and Curtis, 1993; Pan et al., 2012; Roesti et al., 2012).). As long as the number of recombination sites is infinite in the chromosome, the exact shape of the recombination landscape has no effect on junction dynamics. For a chromosome with a finite number of recombination sites, however, junction dynamics change (see the Supplementary Material for a demonstration of increased recombination towards the peripheral ends). If some sites have an increased probability of recombination compared to others, recombination is more likely to occur at a site that has already experienced a recombination event before. As a result, the formation of new junctions is slower than expected under uniform recombination. Interestingly, the maximum number of junctions that can be reached remains unaffected.

Population level processes are also expected to affect junction dynamics. Firstly, deviations from having a constant-population size over time are expected to cause deviations from our universal junction framework. A natural extension of this study would be to include either exponentially or logistically growing populations in order to mimic real life dynamics more closely. In exponentially growing populations, the average heterozygosity does not change (Crow and Kimura, 1970), which results in dynamics that resemble an infinite population. Similarly, for logistically growing populations, during the initial growth phase, junction dynamics are expected to closely resemble junction dynamics in an infinite population. We do have to take into account drift effects even though a population (hybrid lineage) is growing exponentially when this involves a wave front invasion of an open habitat or space (Hallatschek et al., 2007). This could be particularly relevant to cases of hybrid speciation that involve the segregation of a hybrid lineage into a habitat that is not available to the parental species (Mallet, 2007; Nolte and Tautz, 2010). How drift and the rate of growth interact and influence junction dynamics remains the subject of future study. Population subdivision, founder effects, a bottleneck or a permanent decrease in population size could speed up fixation of junctions in the population through drift. Accordingly,

Accepted Article

This article is protected by copyright. All rights reserved.

the average heterozygosity decreases faster than expected, and the accumulation of new junctions is slowed down. Furthermore, the maximum number of junctions decreases as well (following Equation 12) although previously accumulated 'excess' junctions might not be lost after a population change. Thus, depending on the speed and timing of the decrease, individuals in the final population potentially display a larger number of junctions than expected from the current population size, retaining junctions fixed in the population before the population decreased in size.

An important population level process that affects junction dynamics is given by secondary introgression of genetic material from parental populations after the hybrid lineage has emerged. This can be expected to occur in young systems studied in the context of hybrid speciation (Stemshorn et al., 2011; Trier et al., 2014) and introduces parental genomic blocks into the population that have not yet recombined. It leads to an apparent reduction of the number of junctions (Pool and Nielsen, 2009) effectively turning back time in an analysis of junctions if not taken into account. Secondary introgression introduces haplotype blocks that are disproportionally large, compared to the standing haplotype block size distribution and these blocks are expected to be randomly distributed across the genome as opposed to larger blocks that were fixed through selection (Baird, 1995; Ungerer et al., 1998). The haplotype block size distribution therefore provides information that can help to assess whether secondary introgression leaves a strong footprint in an admixed lineage (Barton, 1983; Baird, 1995; Ungerer et al., 1998; Pool and Nielsen, 2009).

We expect that there are more processes that can affect the accumulation of junctions, including, but not limited to, sib mating, interference between recombination events, chromosomal inversion polymorphisms, mutation, meiotic drive, segregation distortion and heterochiasmy. All processes mentioned above cause a slowdown in the accumulation of junctions. In contrast, the formation of new junctions could speed up under balancing selection, when the combination of alleles from both parents provides a selective advantage. The resulting overdominance favors a heterozygous genotype (Maruyama and Nei, 1981), in turn favoring the formation of new junctions. Likewise, positive epistasis among linked loci could favor a combination of alleles of different parents. As a result, recombination between these loci would be selected for, speeding up junction formation. With the exception of overdominance or positive epistasis, this suggests that the additions to the theory of junctions that we have presented here provides an upper speed limit to junction dy-

Accepted Article

This article is protected by copyright. All rights reserved.

namics. Thus, estimates of the onset of hybridization using our framework reflect the youngest age of the hybrids in question. Our framework will help to identify patterns of admixed genome evolution and provide insights in the early evolutionary history of hybrid lineages.

Code availability

Code used for the individual based simulations, and code to evaluate relevant equations has been made available in the cran-R package 'junctions' (`https://CRAN.R-project.org/package=junctions`), and has been included in the Supplementary Information as stand-alone R code.

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

References

Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, R. K. Butlin, U. Dieckmann, F. Eroukhmanoff, A. Grill, S. H. Cahan, J. S. Hermansen, G. Hewitt, A. G. Hudson, C. Jiggins, J. Jones, B. Keller, T. Marczewski, J. Mallet, P. Martinez-Rodriguez, M. Möst, S. Mullen, R. Nichols, A. W. Nolte, C. Parisod, K. Pfennig, A. M. Rice, M. G. Ritchie, B. Seifert, C. M. Smadja, R. Stelkens, J. M. Szymura, R. Väinölä, J. B. W. Wolf, and D. Zinner, 2013. Hybridization and speciation. Journal of Evolutionary Biology 26:229–246.

Arbeithuber, B., A. J. Betancourt, T. Ebner, and I. Tiemann-Boege, 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proceedings of the National Academy of Sciences 112:2109–2114.

Baird, S., 1995. A simulation study of multilocus clines. Evolution 49:1038–1045.

Baird, S., N. Barton, and A. Etheridge, 2003. The distribution of surviving blocks of an ancestral genome. Theoretical Population Biology 64:451–471.

Barton, N., 2001. The role of hybridization in evolution. Molecular Ecology 10:551–568.

Barton, N. H., 1983. Multilocus clines. Evolution 37:454–471.

Bennett, J., 1953. Junctions in inbreeding. Genetica 26:392–406.

Bhat, S., P.-A. Amundsen, R. Knudsen, K. Ø. Gjelland, S.-E. Fevolden, L. Bernatchez, and K. Præbel, 2014. Speciation reversal in european whitefish (coregonus lavaretus (l.)) caused by competitor invasion. PLoS ONE 9:e91208.

Buerkle, C. A., R. J. Morris, M. A. Asmussen, and L. H. Rieseberg, 2000. The likelihood of homoploid hybrid speciation. Heredity 84:441–451.

Buerkle, C. A. and L. H. Rieseberg, 2008. The rate of genome stabilization in homoploid hybrid species. Evolution 62:266–275.

This article is protected by copyright. All rights reserved.

Accepted Article

Chapman, N. and E. Thompson, 2003. A model for the length of tracts of identity by descent in finite random mating populations. Theoretical Population Biology 64:141–150.

Chapman, N. H. and E. A. Thompson, 2002. The effect of population history on the lengths of ancestral chromosome segments. Genetics 162:449–458.

Corbett-Detig, R. and R. Nielsen, 2017. A hidden markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. PLoS Genetics 13:e1006529.

Crow, J. F. and M. Kimura, 1970. An Introduction to Population Genetics Theory. Harper and Row, New York.

Edmands, S., H. Feaman, J. Harrison, and C. Timmerman, 2005. Genetic consequences of many generations of hybridization between divergent copepod populations. Journal of Heredity 96:114–123.

Fisher, R. A., 1949. The Theory of Inbreeding. Oliver and Boyd.

———, 1954. A fuller theory of "junctions" in inbreeding. Heredity 8:187–197.

———, 1959. An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. Heredity 13:179–186.

Gale, J., 1964. Some applications of the theory of junctions. Biometrics Pp. 85–117.

Gerton, J. L., J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, and T. D. Petes, 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast saccharomyces cerevisiae. Proceedings of the National Academy of Sciences 97:11383–11390.

Gompert, Z. and C. A. Buerkle, 2016. What, if anything, are hybrids: enduring truths and challenges associated with population structure and gene flow. Evolutionary applications 9:909–923.

Grant, V., 1981. Plant speciation. Columbia University Press.

This article is protected by copyright. All rights reserved.

Accepted Article

Gravel, S., 2012. Population genetics models of local ancestry. Genetics 191:607–619.

Hallatschek, O., P. Hersen, S. Ramanathan, and D. R. Nelson, 2007. Genetic drift at expanding frontiers promotes gene segregation. Proceedings of the National Academy of Sciences 104:19926–19930.

Hellenthal, G., G. B. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers, 2014. A genetic atlas of human admixture history. Science 343:747–751.

Kimura, M., 1956. A model of a genetic system which leads to closer linkage by natural selection. Evolution Pp. 278–287.

Krehenwinkel, H. and D. Tautz, 2013. Northern range expansion of european populations of the wasp spider argiope bruennichi is associated with global warming–correlated genetic admixture and population-specific temperature adaptations. Molecular Ecology 22:2232–2248.

Lamichhaney, S., F. Han, M. T. Webster, L. Andersson, B. R. Grant, and P. R. Grant, 2017. Rapid hybrid speciation in darwin's finches. Science P. eaao4593.

Lewontin, R. and P. Hull, 1967. The interaction of selection and linkage iii synergistic effect of blocks of genes. Der Züchter 37:93–98.

Liang, M. and R. Nielsen, 2014. The lengths of admixture tracts. Genetics 197:953–967.

Lukaszewski, A. and C. Curtis, 1993. Physical distribution of recombination in b-genome chromosomes of tetraploid wheat. Theoretical and Applied Genetics 86:121–127.

Mackiewicz, D., P. M. C. de Oliveira, S. M. de Oliveira, and S. Cebrat, 2013. Distribution of recombination hotspots in the human genome–a comparison of computer simulations with real data. PloS ONE 8:e65272.

MacLeod, A., C. Haley, J. Woolliams, and P. Stam, 2005. Marker densities and the mapping of ancestral junctions. Genetical research 85:69–79.

Mallet, J., 2007. Hybrid speciation. Nature 446:279–283.

This article is protected by copyright. All rights reserved.

Accepted Article

Maruyama, T. and M. Nei, 1981. Genetic variability maintained by mutation and overdominant selection in finite populations. Genetics 98:441–459.

McTavish, E. J. and D. M. Hillis, 2014. A genomic approach for distinguishing between recent and ancient admixture as applied to cattle. Journal of Heredity 105:445–456.

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324.

Nolte, A. W., J. Freyhof, K. C. Stemshorn, and D. Tautz, 2005. An invasive lineage of sculpins, cottus sp. (pisces, teleostei) in the rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. Proceedings of the Royal Society B 272:2379–2387.

Nolte, A. W. and D. Tautz, 2010. Understanding the onset of hybrid speciation. Trends in Genetics 26:54–58.

Pan, Q., F. Ali, X. Yang, J. Li, and J. Yan, 2012. Exploring the genetic characteristics of two recombinant inbred line populations via high-density snp markers in maize. PLoS ONE 7.

Payseur, B. A. and L. H. Rieseberg, 2016. A genomic perspective on hybridization and speciation. Molecular Ecology .

Pool, J. E. and R. Nielsen, 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181:711–719.

Roesti, M., A. P. Hendry, W. Salzburger, and D. Berner, 2012. Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. Molecular Ecology 21:2852–2862.

Schumer, M., R. Cui, D. L. Powell, G. G. Rosenthal, and P. Andolfatto, 2016. Ancient hybridization and genomic stabilization in a swordtail fish. Molecular Ecology .

Schumer, M., G. G. Rosenthal, and P. Andolfatto, 2014. How common is homoploid hybrid speciation? Evolution 68:1553–1560.

This article is protected by copyright. All rights reserved.

Accepted Article

Singhal, S., E. M. Leffler, K. Sannareddy, I. Turner, O. Venn, D. M. Hooper, A. I. Strand, Q. Li, B. Raney, C. N. Balakrishnan, S. C. Griffith, G. McVean, and M. Przeworski, 2015. Stable recombination hotspots in birds. Science 350:928–932.

Smagulova, F., I. V. Gregoretti, K. Brick, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova, 2011. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature 472:375–378.

Stam, P., 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genetical Research 35:131–155.

Stemshorn, K. C., F. A. Reed, A. W. Nolte, and D. Tautz, 2011. Rapid formation of distinct hybrid lineages after secondary contact of two fish species (cottus sp.). Molecular Ecology 20:1475–1491.

Taylor, E., J. Boughman, M. Groenenboom, M. Sniatynski, D. Schluter, and J. Gow, 2006. Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (gasterosteus aculeatus) species pair. Molecular Ecology 15:343–355.

Trier, C. N., J. S. Hermansen, G.-P. Sætre, and R. I. Bailey, 2014. Evidence for mito-nuclear and sex-linked reproductive barriers between the hybrid italian sparrow and its parent species. PLoS Genetics 10:e1004075.

Ungerer, M. C., S. J. Baird, J. Pan, and L. H. Rieseberg, 1998. Rapid hybrid speciation in wild sunflowers. Proceedings of the National Academy of Sciences 95:11757–11762.

Vonlanthen, P., D. Bittner, A. G. Hudson, K. A. Young, R. Müller, B. Lundsgaard-Hansen, D. Roy, S. Di Piazza, C. R. Largiadèr, and O. Seehausen, 2012. Eutrophication causes speciation reversal in whitefish adaptive radiations. Nature 482:357–362.

This article is protected by copyright. All rights reserved.

Figure captions

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

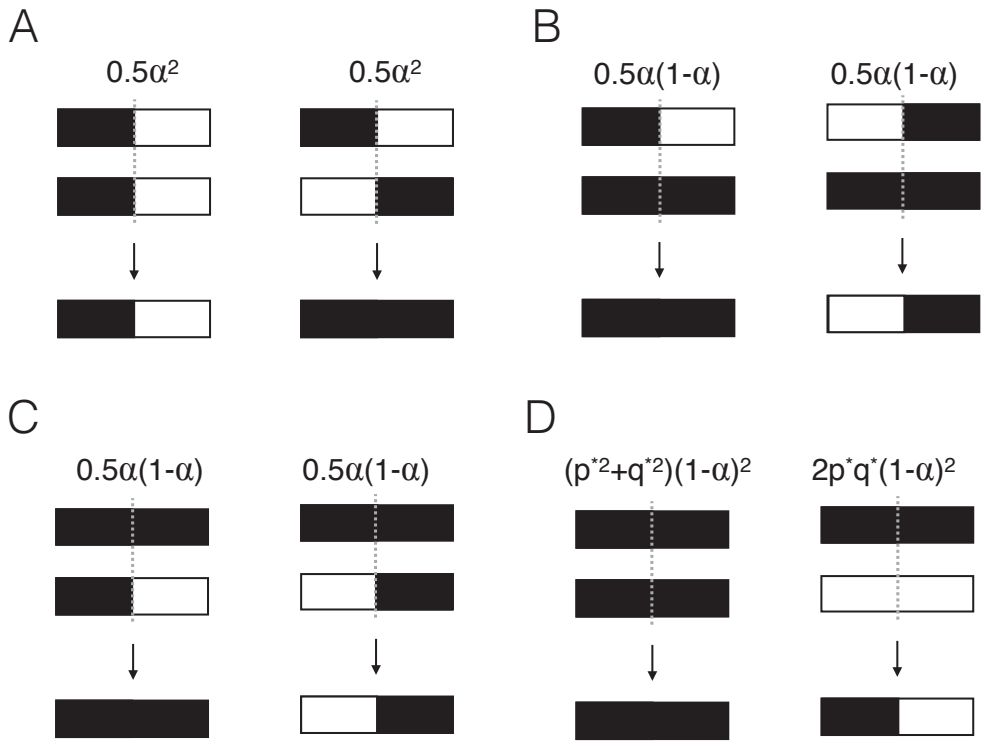

Figure 1: Change in number of junctions depending on the genomic match between blocks. Full chromosomes are shown here, but the same rationale applies to subsets of a chromosome. Top rows within each panel indicate the two parental chromosomes, bottom row indicates one of two possible resulting chromosomes after meiosis, where recombination takes place at the dashed grey line. Genomic material of type P is indicated in black, genomic material of type Q is indicated in white. With probability α^2 recombination takes place on an existing junction on both chromosomes **A**, with probability $\alpha(1-\alpha)$ recombination takes place on an existing junction on one chromosome, and within a block on the other chromosome **B**, with probability $(1-\alpha)\alpha$ recombination takes place on within a block on one chromosome, and on an existing junction on the other chromosome **C**, with probability $(1-\alpha)^2$ recombination takes place within a block on both chromosomes **D**.

This article is protected by copyright. All rights reserved.

Accepted Article

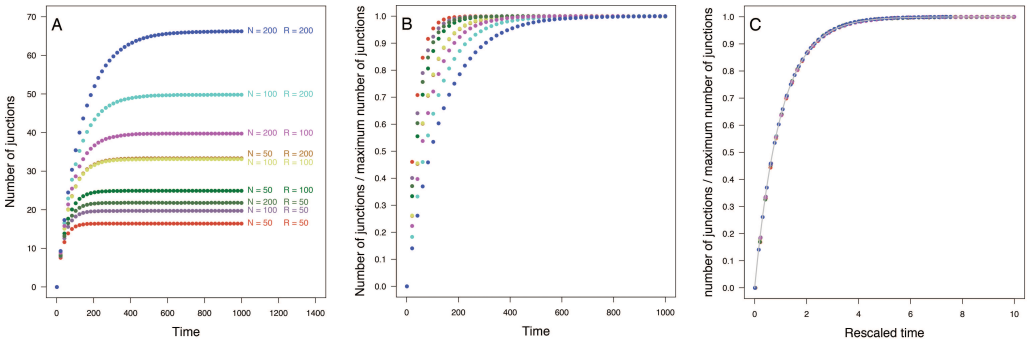

Figure 2: Graphical example of the construction of universal junction dynamics using results from individual based simulations. **A**: mean number of junctions for $H_0 = 0.5$, $N = [50, 100, 200]$ and $R = [50, 100, 200]$, number of replicates = 10,000. **B**: The mean number of junctions for the same parameter combinations, after rescaling the number of blocks relative to the maximum number of junctions K. **C**: The rescaled number of junctions vs rescaled time, by rescaling time according to β in Equation (21). After rescaling both the number of junctions according to K, and time according to β, all curves for different values of N and R reduce to a single, universal, curve, which follows Equation (22).

This article is protected by copyright. All rights reserved.

Accepted Article

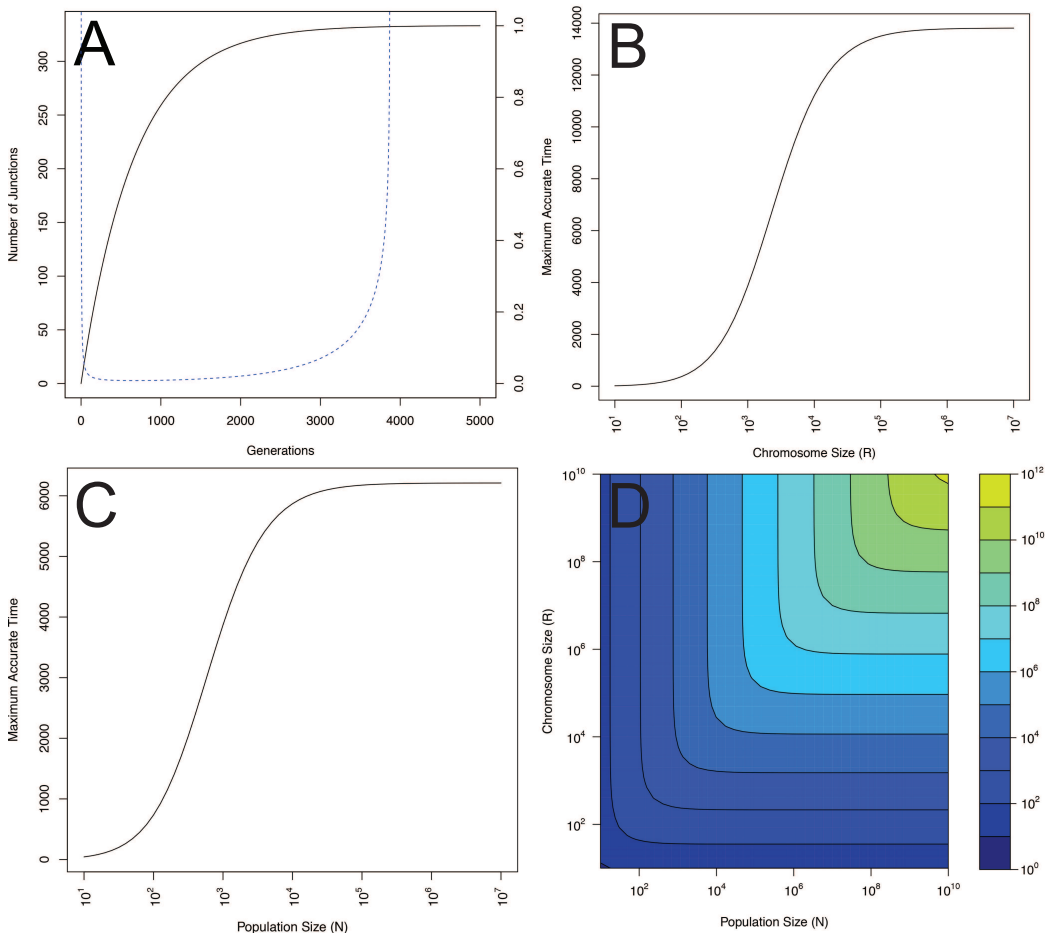

Figure 3: **A:** Number of junctions over time for $N = 1000, R = 1000$ (black line), and the associated relative error dt/t (Eq. 4 in the Supplementary Material) (dashed blue line). Y-axis at the left hand side corresponds to the number of junctions, y-axis on the right hand side corresponds to the associated relative error. As the number of junctions approaches K, the relative error approaches infinity. **B:** Maximum accurate time (Equation (15)) for $N = 1000$, for an increasing number of markers. As the number of markers increases towards high numbers (10^7), the maximum accurate time approaches 14000. **C:** Maximum accurate time (Equation (15)) for $R = 1000$, for increasing population size. As population size increases towards large values, the maximum accurate time plateaus around 6000 generations. **D:** Maximum Accurate Time (τ_{MAT}) for N and R, please note the logarithmic scale. Results show that the accurate range increases exponentially with increasing N and R. Results show that the accurate range increases exponentially with increasing N and R. For all four plots, $H_0 = 0.5, C = 1$.

This article is protected by copyright. All rights reserved.

Accepted Article

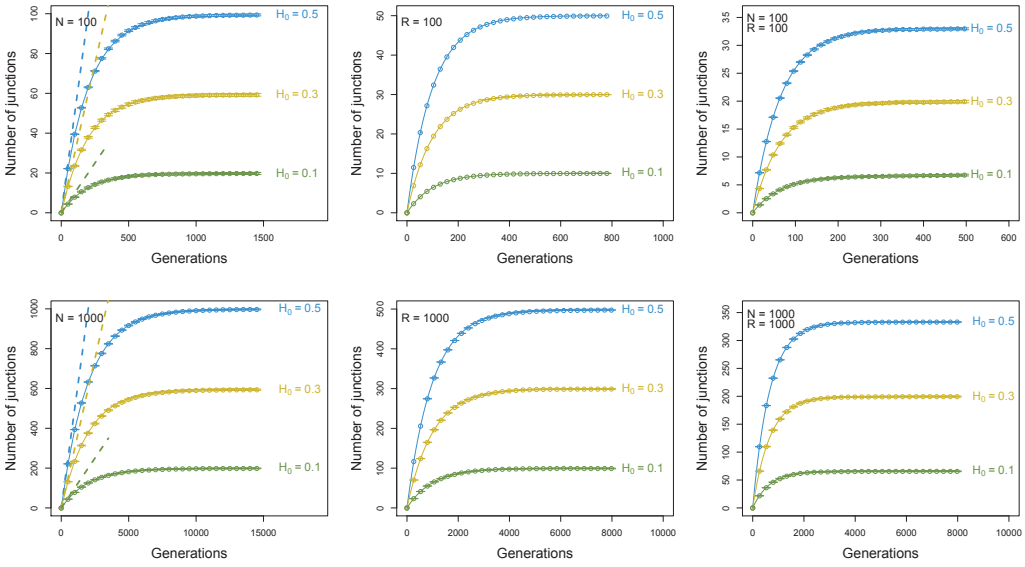

Figure 4: Number of junctions over time for individual based simulations, and analytical predictions. **Left column:** Results assuming a chromosome with an infinite number of junction sites, $R = \infty$, and a population size (N) of 100 or 1000 individuals (circles), the analytical prediction for an infinite population size (dashed line), or the analytical prediction for a finite population size (Equation (3)), solid line. **Middle column:** Results assuming a chromosome with a finite number (R) of junction sites, where R is 100 or 1000, and a population size of 100,000 (circles), or the analytical prediction according to Equation (7) (solid line). **Right column:** Results for a chromosome with a finite number of junction sites of length R of 100 or 1000, and a finite population of size N is 100 or 1000 (circles), or the analytical prediction according to Equation (11). Error bars indicate the standard error of the mean across 1,000 replicates. Shown are results for different initial frequencies of heterozysity H_0. For all results shown, $C = 1$.

This article is protected by copyright. All rights reserved.

Accepted Article

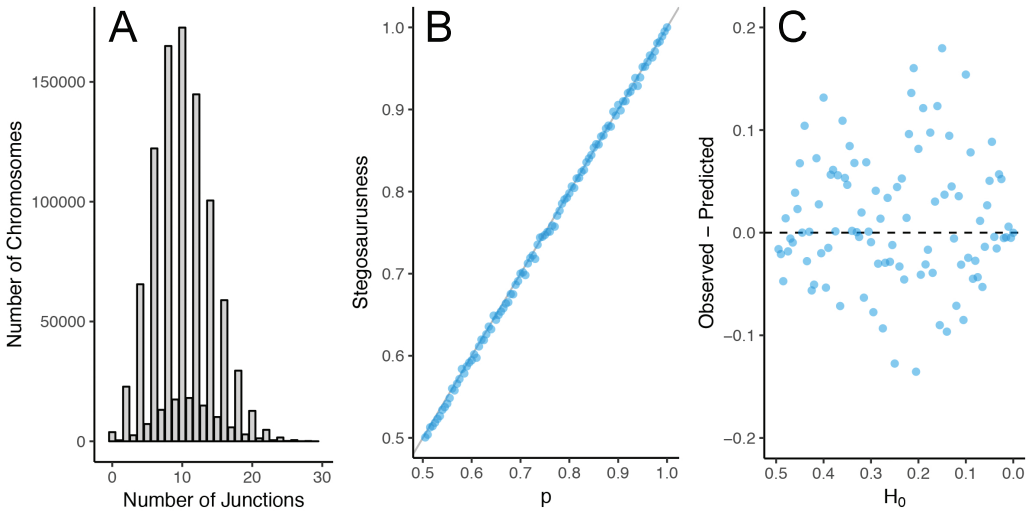

Figure 5: **A**: Example distribution of number of chromosomes with n junctions. Results are shown for $N = 500,000$, $C = 1$, $H_0 = 0.1$, $R = 100$ at $t = 500$. Results show the 'stegosaurus' pattern, where chromosomes with an even number of junctions are more frequent than chromosomes with an odd number of junctions. **B**: Linear relationship between the degree of 'stegosaurusness' and the frequency of the most common ancestor in the hybrid swarm, p. Results are shown for $N = 500,000$, $C = 1$, $R = 100$ at $t = 500$. **C**: Inaccuracy in the estimate of the maximum number of blocks K, depending on the initial heterozygosity $H_0 = 2p(1-p)$. Although the degree of stegosaurusness increases with decreasing H_0, error in the estimate of K, and overall estimate in the accumulation of junctions, does not increase. Points are the mean of 10 replicates.

This article is protected by copyright. All rights reserved.

Accepted Article

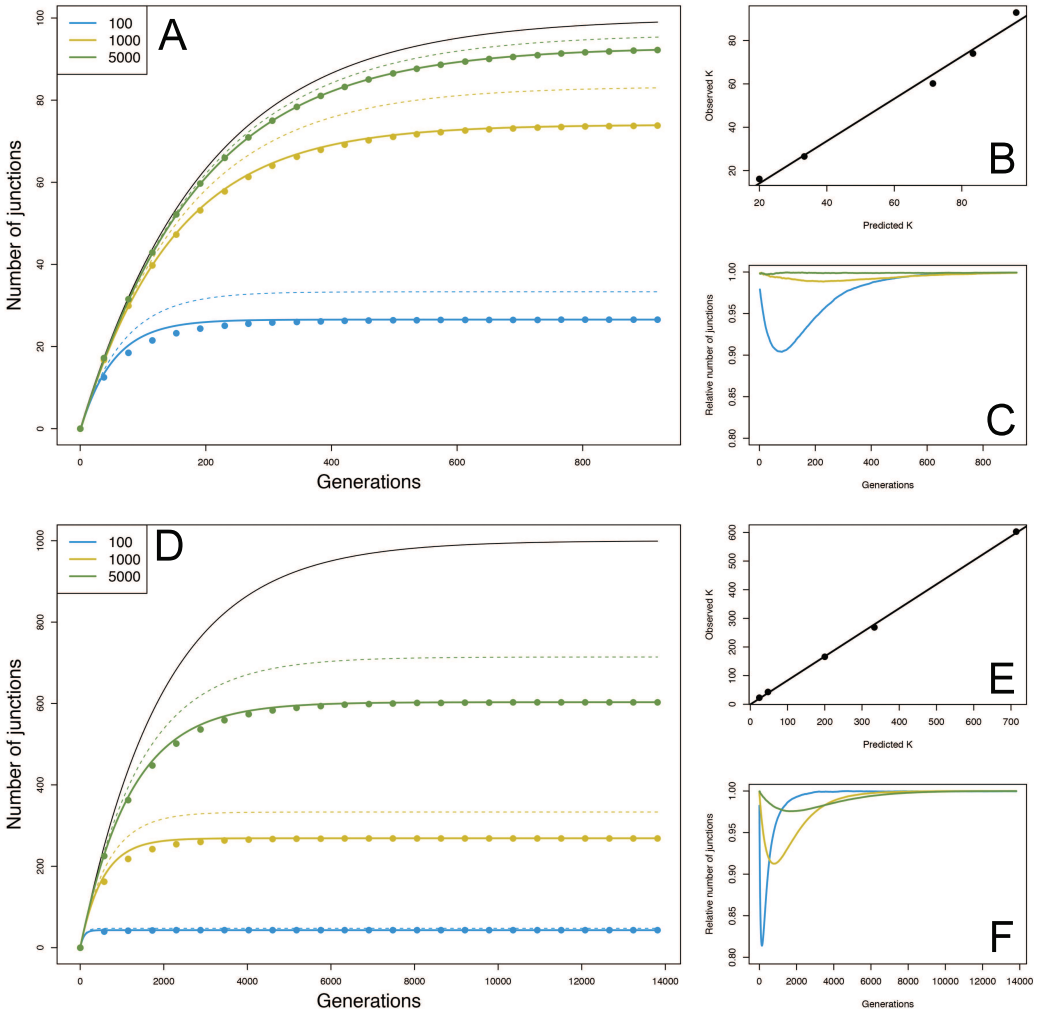

Figure 6: Shown are results for two different population sizes: $N = 100$ (A-C) and $N = 1000$ (D-F) **A,C:** The number of junctions over time ($H_0 = 0.5$, $C = 1$). The black line indicates the true number of junctions, assuming an infinite number of recombination sites. The dashed lines indicate the expected number of junctions for a chromosome consisting of a finite number of recombination sites, where $R = 100, 1000$ and 5000 respectively. Dots indicate the mean number of junctions from simulations using randomly spaced markers across the chromosome. The solid line indicates the expected number of junctions, using the generalized framework with K equal to the maximum mean number of junctions from simulations using random markers. **B, E:** Correlation between the observed value of K, and the expected value for K, with $R = 50, 100, 500, 1000$ and 5000, and $N = 100$ (B), and $N = 1000$ (E). R^2 values are for both population sizes above 0.99. Slopes are 0.97, and 0.84 for $N = 100$ and $N = 1000$ respectively. **C, F:** The ratio between the mean number of junctions with $R = 100, 1000$ and 5000 in simulations using randomly spaced markers, and the expected number of junctions using the universal framework with K equal to the maximum mean number of junctions from simulations using random markers.

This article is protected by copyright. All rights reserved.

Accepted Article

Appendix

Deriving the unifying framework

Generally, we can infer that the maximum number of junctions is dependent on p, N, R and C, as do the time dynamics required to reach this maximum. Hence, we can describe the number of junctions relative to the number of junctions at $t = \infty$, which we define here as K. During meiosis, the average number of junctions in the most ideal case can increase with γ junctions. There might be limitations, dependent on p, N, R or the number of already existing junctions. As such, we make the ansatz

$$\frac{dJ}{dt} = \gamma - \lambda J \tag{16}$$

where γ is the maximum growth rate and λ encompasses all factors limiting the formation of new junctions. This includes, but is not limited to, factors induced by a finite population size N and by a finite number of recombination spots R. Defining $\tau = \lambda t$, we can rescale time in Equation (16) such that

$$\frac{dJ}{d\tau} = \frac{\gamma}{\lambda} - J. \tag{17}$$

Measuring the number of junctions in terms of their asymptotic values, $\tilde{J} = \lambda\frac{J}{\gamma}$, we obtain

$$\frac{d\tilde{J}}{d\tau} = 1 - \tilde{J} \tag{18}$$

The solution of Equation (18) is of the form

$$\tilde{J}(\tau) = 1 - e^{-\tau}. \tag{19}$$

(Please note that we refer to the number of junctions in continuous time as $J(\tau)$ and the number of junctions in discrete time as J_t). We can find the scaling of time by solving

$$\tilde{J}(1) - \tilde{J}(0) = K(1 - e^{-\beta t}) - K(1 - e^{0}) \tag{20}$$

By definition $\tilde{J}(0)$ is zero, which allows us to calculate β

$$\beta = -\ln\left(1 - \frac{J(1)}{K}\right) \tag{21}$$

This article is protected by copyright. All rights reserved.

Returning to discrete time, we can formulate the universal dynamics as

$$J_t = K\left(1 - \exp\left(\ln\left(1 - \frac{J(1)}{K}\right)t\right)\right)$$
$$= K - K\left(1 - \frac{J(1)}{K}\right)^t \tag{22}$$

For $t \to \infty$, this converges to K, and for $t = 0$ this is equal to zero. Equation (22) provides us with a general scalable equation where all junction dynamics are described in terms of K and $J(1)$.

To implement the unifying equation (22), we only need to know K and $J(1)$. Since $J(0) = J_0 = 0$, we obtain from Eq. (10) $J(1) = H_0C$. Thus, regardless whether the chromosome contains a finite or infinite number of recombination sites, and regardless of whether the population is finite or infinite, we find that the initial change is always H_0C, and that $J(1) = H_0C$. This also makes intuitive sense: in the first generation, none of the factors that limit recombination as a result of finite population size, or finite chromosome size come into play yet. When the population is finite, the formation of new blocks is limited by recombination taking place at a recombination spot where in a previous generation recombination has already taken place. However, in the first generation, all chromosomes are non-recombined, and finite population effects have no effect yet. When the number of recombination sites is finite, the formation of new blocks is limited by the maximum packing density of junctions. However, in the first generation, no junctions are apperent yet, and this effect is negligible. Concluding, we can formulate our general haplotype framework as (Equation (13) in the main text):

$$J_t = K - K\left(1 - \frac{H_0C}{K}\right)^t. \tag{23}$$

Importance of finite chromosome length

Our universal framework introduces the possibility to take into account a finite number of recombination sites. Previous work has inferred junction dynamics for chromosomes with an infinite number of recombination sites. An important question to be answered is for which values of R our universal framework approaches previous results.

Accepted Article

This article is protected by copyright. All rights reserved.

For small $\frac{H_0}{K}$, we can approximate our unifying framework (Equation (23) in the Appendix, and Equation (13) in the maintext) by

$$J_t \approx H_0 C t - \frac{H^2 C^2}{2K}(t^2 - t). \tag{24}$$

Where the last term describes the limitation in the number of junctions dependent on K. For $K \to \infty$, we recover $J_t = H_0 C t$. Substituting the upper limit for finite R, $K = H_0 R$ in Equation (24), we obtain

$$J_t \approx H_0 C t - H_0 C \left(\frac{C}{2R}\right)(t^2 - t). \tag{25}$$

We expect that the impact of R is negligible compared to the linear term in t if the second term of Equation (25) is much smaller than the first term. This implies that

$$R \gg C\frac{t-1}{2}, \tag{26}$$

which for $t \gg 1$ is approximately $t \ll 2R/C$ – the approximation of an infinite R is good as long as the time in generation is much shorter than twice the length of the chromosome (in genetic elements), divided by the size in Morgan.

When both R and N are finite, we can substitute the appropriate K (Equation (12)) into Equation (24), and obtain

$$J_t \approx H_0 C t - H_0 C \left(\frac{C}{2R} + \frac{1}{4N}\right)(t^2 - t). \tag{27}$$

Comparing this with the approximation for a finite population (but an infinite number of recombination spots), where $K = 2H_0 C N$, we see that

$$J_t \approx H_0 C t - H_0 C \left(\frac{1}{4N}\right)(t^2 - t). \tag{28}$$

Differences induced by a finite R (e.g. differences between Equations (27) and (28)) are then due to the term $\frac{C}{2R}$ in Equation (27), and the impact of finite R thus decreases rapidly with increasing R. Furthermore, the impact of R on the number of junctions, compared to the impact of N, is neglible if:

$$R \gg 2NC \tag{29}$$

Accepted Article

This article is protected by copyright. All rights reserved.