Genet. Res., Camb. (2005), **85**, *pp.* 69–79. *With 7 figures.* © 2005 Cambridge University Press
doi:10.1017/S0016672305007329 Printed in the United Kingdom

Marker densities and the mapping of ancestral junctions

A. K. MACLEOD[1]*, C. S. HALEY[1], J. A. WOOLLIAMS[1] AND with an Appendix by P. STAM[2]
[1] *Roslin Institute (Edinburgh), Roslin, Midlothian, EH25 9PS, UK*
[2] *Department of Plant Science, Wageningen University, P.O. Box 386, 6700 AJ Wageningen, The Netherlands*

(*Received 2 June 2004 and in revised form 8 September and 9 December 2004*)

Summary

In any partially inbred population, 'junctions' are the loci that form boundaries between segments of ancestral chromosomes. Here we show that the expected number of junctions per Morgan in such a population is linearly related to the inbreeding coefficient of the population, with a maximum in a completely inbred population corresponding to the prediction given by Stam (1980). We further show that high-density marker maps (fully informative markers with average densities of up to 200 per cM) will fail to detect a significant proportion of the junctions present in highly inbred populations. The number of junctions detected is lower than that which would be expected if junctions were distributed randomly along the chromosome, and we show that junctions are not, in fact randomly spaced. This non-random spacing of junctions significantly increases the number of markers that is required to detect 90% of the junctions present on any chromosome: a marker count of at least 12 times the number of junctions present will be needed to detect this proportion.

1. Introduction

Several recent studies examining densely spaced genetic markers in human populations have presented evidence that some regions of the genome are of limited haplotype diversity (Daly *et al.*, 2001; Reich *et al.*, 2001; Gabriel *et al.*, 2002). These regions have been termed 'haplotype blocks', and are separated by regions of higher diversity. Such blocks have been reported in several regions of the human genome, including the MHC region of chromosome 5 (Daly *et al.*, 2001), across the whole of chromosomes 19 (Patil *et al.*, 2001) and 21 (Phillips *et al.*, 2003), and at randomly selected regions throughout the genome (Gabriel *et al.*, 2002). It is argued (Johnson *et al.*, 2001) that if such blocks were present throughout the genome, they would facilitate whole-genome association studies to identify quantitative trait loci (QTL), as fewer markers would need to be genotyped to locate a putative QTL within a haplotype block. This has led to the development of a project to characterize the diversity of haplotypes in the human genome: the human HAPMAP project (International HapMap Consortium 2003).

However, the phenomenon of haplotype blocks is not completely understood. In the MHC region of chromosome 5, the regions which separate haplotype blocks correspond to recombination hotspots (Daly *et al.*, 2001), but there is evidence that such hotspots are not necessary for haplotype blocks to form, as blocks can be observed resulting from simulation studies of random recombination and genetic drift only (Zhang *et al.*, 2003). The problem of separating these two causes, the former an attribute of the genome, the latter an attribute of the population, is compounded by the lack of a precise, universally applicable definition of a haplotype block. Blocks are often defined as regions where inter-marker linkage disequilibrium exceeds a certain threshold (e.g. Jeffreys *et al.*, 2001) or regions where a few haplotypes can account for most of the observed marker genotypes (e.g. Patil *et al.*, 2001), but these definitions refer only to marker genotypes, without referring to the underlying patterns of ancestral segments.

To distinguish between these two mechanisms of block formation, it is necessary to quantify what may

* Corresponding author. e-mail: andy.macleod@bbsrc.ac.uk.

be expected from the two processes. One theory that relates genetic drift to the lengths of ancestral segments is Fisher's theory of junctions (Fisher, 1954). Fisher describes the theory of junctions as a 'strand theory' rather than a 'point theory', one that looks at the inheritance of whole tracts of chromosome rather than individual loci. In Fisher's theory, a junction is formed where a recombination event occurs between two chromosomes that are not identical by descent (IBD) at the location of the crossover, where IBD is defined relative to a base generation. Junctions can be treated as point mutations, and followed to fixation or loss within the population over time. The genome of any individual in a partially inbred population will consist of alternating IBD and non-IBD sections, which are separated by 'external' junctions (Stam 1980). In a completely inbred population, all individuals will be IBD across the whole genome, with each chromosome consisting of several segments of different lengths, each derived from a distinct chromosome in the base population, and separated by 'internal' junctions (Stam, 1980).

Stam (1980), and Chapman & Thompson (2003) expand Fisher's theory by deriving expressions for the expectation and variance of the length of an IBD tract (the length of the section of chromosome between two external junctions), as opposed to the probability that an individual is IBD at a specific locus. The papers of Stam and Chapman & Thompson relate to random mating populations excluding and including selfing, respectively. Stam (1980) showed that starting with a fixed outbred population of size N, with a genome of length L Morgan distributed over n pairs of homologous chromosomes, and allowing the population to mate randomly with the exception of selfing, until the population was entirely inbred, the number of distinct segments observed in an individual in the inbred population (S_∞) would be:

$$S_\infty = 2(N+1)L+n.$$

The expected number of junctions (excluding chromosome ends) would thus be:

$$J_\infty = 2(N+1)L \tag{1}$$

corresponding to $2(N+1)$ junctions per Morgan in a completely inbred population (Stam, 1980).

Any inference made on a population's history using the theory of junctions will necessarily be made using genetic markers. Whilst the results of Stam (1980) and Chapman & Thompson (2003) give a precise record of the make-up of each chromosome, this would not be available if markers alone were examined. Indeed Stam states (p. 143), with reference to his results, that using spaced markers 'may result in an underestimation of the number of junctions'. This paper reviews the results of Stam and examines the influence of marker maps of varying density on the observation of junctions. This is accomplished by simulation, with precise tracking of junction location, and by superimposing marker maps and assessing the existence of junctions from these maps alone. The distributions of segment lengths are also examined with reference to both theory and simulation.

2. Methods

(i) *Concepts and definitions*

A 'segment' is defined here as the region of a chromosome between two junctions, whereas a 'bracket' is defined as the region between two markers. Segments are regions inherited unbroken from a single chromosome in the ancestral population; brackets may appear to be unbroken if the markers at either end derive from the same ancestral chromosome, but may contain any number of junctions.

Following Fisher, junctions are one of two major types, depending on the state of the genome on either side of the junction on a pair of homologous chromosomes. An external junction forms the boundary between IBD and non-IBD tracts. Internal junctions are either IBD on both sides of the junction or non-IBD on both sides of the junction, and are labelled type I and type II, respectively (Stam, 1980). These junction types are illustrated in Fig. 1.

There are 14 junctions in total across the two chromosomes in Fig. 1, but only 8 are detected from the marker map alone. This map thus underestimates the actual number of junctions present.

(ii) *Simulated populations*

A series of populations was simulated using FORTRAN 90 software. Each individual consisted of a single pair of homologous chromosomes of length 1 Morgan. Each chromosome was unique in the first generation, i.e. the population was not inbred. Generation $t+1$ was generated by random mating of individuals in generation t, with self-fertilization excluded, to be consistent with the simulations of Stam (1980).

For each individual in generation $t+1$, the first parent was selected at random from generation t and one gamete generated, and the second parent selected from the pool of remaining individuals. The number of recombination events occurring at the formation of each gamete was sampled from a Poisson distribution with parameter 1, and the location of each recombination was placed at random along the chromosome, as indicated by a random number drawn from U[0,1].

Marker maps were either based on m equidistant markers or obtained by generating a random array of

https://doi.org/10.1017/S0016672305007329 Published online by Cambridge University Press

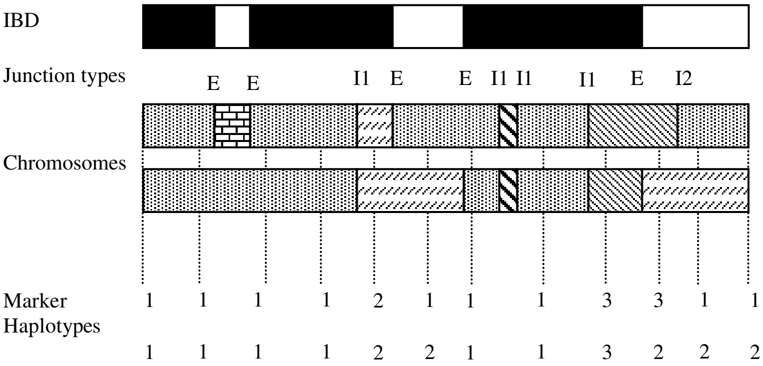

Fig. 1. An example of internal and external junctions matched to the pattern of IBD along gametes. Junctions types are shown above the chromosomes: I, internal (types 1 and 2); E, external, with the pattern of IBD across the chromosome shown above (black, IBD, white, non-IBD). Below the chromosomes are shown the marker haplotypes for a set of markers spaced across the chromosome.

m marker locations. In the latter case, markers were placed at the two ends of the chromosome, but otherwise at positions indicated by a random number drawn from U[0,1]. The number of possible alleles at each locus was $2N$, the number of chromosomes in the founder population, with the marker haplotypes defined in the founders as $\{j, j, \ldots, j\}$ for the jth ancestral chromosome, where $j=1, \ldots, 2N$. Thereafter, inheritance followed directly from the sampling of gametes and recombinations described previously.

Simulations were run for all combinations of population sizes of $N=2$, 5, 10, 25 and 50 with $B=10$, 20, 50, 100, 200 marker brackets, formed by $m=B+1$ markers including a marker at each end of the chromosome. Each combination of parameters was simulated over 1000 replicates, with each replicate proceeding until the population was completely inbred.

(iii) *Observations*

For each replicate, in each generation, t, we measured: (i) the number of junctions present on each chromosome, and J_t, the average junctions per Morgan over the $2N$ chromosomes; (ii) the genotypes of all markers on each chromosome, and (iii) the number of junctions per chromosome, inferred from the marker genotypes. To calculate the inferred number of junctions from the marker map, brackets were classified as IBD, if the markers at either end derived from the same ancestral chromosome, or non-IBD otherwise. The number of non-IBD brackets was assumed to equate to the number of junctions. When each population was completely inbred, the length of each chromosome segment was recorded, as well as the average number of junctions per Morgan. Results were obtained by averaging over replicates, and standard errors were calculated for the variation between replicates. Inbreeding coefficients in each generation were calculated as $F_t=(1-H_t)$, where $H_t=K_{t-1}$, and $K_t=\frac{1}{2N}H_{t-1}+\left(1-\frac{1}{N}\right)K_{t-1}$. H_t is the probability that two homologous loci drawn at random from one individual in generation t are non-IBD, and K_t the probability that two homologous loci drawn at random from two distinct individuals in generation t are non-IBD (Stam, 1980).

(iv) *Predictions*

(a) *Segment lengths*

Assuming that the location of junctions that become fixed in the final generation represents a random sample of J_∞ points along the chromosome, the distribution of the $J_\infty+1$ segment lengths will be given by a β distribution with parameters $[1, J_\infty]$ (Waddington *et al.*, 2000). The lengths obtained from simulation were compared with this distribution, by taking the first segment length for each simulated population. Under the assumptions given above, the distributional property remains valid despite the chosen segment being situated at one chromosome end.

The distribution of segments obtained for a specified population size, N, and number of markers, m, are conditional on the realized value of J_∞. Therefore unconditional expectations for the mean and variance of first segment lengths, L, were obtained from the mean and variances of $\beta[1, J_\infty]$ and the mean and variance of J_∞ over replicates.

$$E[L]=E_{replicates}[1/(J_\infty+1)] \tag{2}$$

$$\begin{aligned}\mathrm{Var}[L]=&\mathrm{Var}_{replicates}[1/(J_\infty+1)]\\&+E_{replicates}[J_\infty(J_\infty+1)^{-2}(J_\infty+2)^{-1}].\end{aligned} \tag{3}$$

(b) *Detection of junctions*

Separate predictions were made for randomly spaced markers and for equidistant markers assuming that

https://doi.org/10.1017/S0016672305007329 Published online by Cambridge University Press

Table 1. *Predicted and observed numbers of junctions in the completely inbred populations*

Population size (N)	Predicted junctions	Observed junctions	Standard deviation	Range	Inter-quartile range
2	6	6·06 (0·094)	2·999	[0, 17]	[4, 8]
5	12	11·90 (0·134)	4·249	[1, 27]	[9, 14]
10	22	22·06 (0·205)	6·487	[3, 47]	[17, 26]
25	52	51·90 (0·313)	9·911	[17, 98]	[45, 58]
50	102	102·31 (0·458)	14·482	[59, 150]	[92, 112]

Predicted values follow equation (1), observed values are averaged over 1000 replicates, with standard errors in parentheses.

the junctions were randomly, and independently, scattered over the chromosome.

Randomly scattered markers. On a chromosome with $B+1$ markers, one placed at each end of the chromosome but otherwise randomly distributed, and J_∞ junctions, the number of junctions we predict the marker map will detect is taken to equal to the number of brackets that contain one or more junctions. This is equivalent to the number of runs of one or more junctions in a randomly drawn sequence of J_∞ junctions and $B-1$ markers (i.e. the number of groups of one or more junctions separated from each other by one or more markers). Thus, ignoring ends, a sequence with 4 internal markers (m) and 3 junctions (j) given by *jmmjjmm* would be assumed to detect only 2 junctions. Note that this would be expected to further underestimate the number of junctions when runs of length 2 or more result in the flanking markers coming from the same ancestral chromosome, resulting in no junctions being detected in that bracket. However, this error may be expected to diminish as N becomes large. Using the results on theory of runs from Feller (1967), for a chromosome with J_∞ junctions, and $B+1$ markers (including one at each end), let J_D represent the junctions detected. Then:

$$E[J_D]=J_\infty B(J_\infty+B-1)^{-1} \tag{4}$$

$$\mathrm{Var}[J_D]\approx E[J_D]^2(J_\infty+B-1)^{-1}. \tag{5}$$

Note that the expected number of runs of length 2 or more provides an estimate of one component of the risk of missing junctions. This can be shown to be $E[J_D](J_\infty-1)(J_\infty+B-2)^{-1}$. By defining variables $X_t=1$ if the sequence $t-1$, t, $t+1$ is *jjm* and 0 otherwise, this is calculated as $\sum_{t=2}^{J_\infty+B+1}E(X_t)$.

Equi-distant markers. Although detection of junctions for equidistant markers still relies upon interpreting sequences such as *jmmjjmm*, the markers are no longer spaced randomly, and the problem is equivalent to that of the classical occupancy problem described by Feller (1967): for a chromosome with a uniformly spaced marker map, the problem becomes one of distributing J_∞ junctions in B brackets of equal length, rather than looking at the order of two non-uniformly distributed sequences of markers and junctions. With J_∞ junctions in B brackets, Feller (1967) shows that as J_∞ and B become large, the distribution of unoccupied brackets tends to Poisson with mean $B\mathrm{e}^{-\gamma}$, where $\gamma=J_\infty/B$. Thus:

$$E[J_D]=B(1-\mathrm{e}^{-\gamma}) \text{ and } \mathrm{Var}[J_D]=B\mathrm{e}^{-\gamma}. \tag{6}$$

The expected single occupancy is also approximated by Poisson with mean $\gamma B\mathrm{e}^{-\gamma}$, so the expected multiple occupancy is $B(1-(1+\gamma)\mathrm{e}^{-\gamma})$.

Predicted proportions were compared to the actual proportion of junctions detected for randomly spaced marker maps of a given (average) density. For both random and equidistant markers the moments given above, conditional on J_∞, were made conditional on N using the approach described for segment lengths (see equations 2 and 3).

Empirical predictions. Empirical predictions were made for J_D, the number of junctions detected on an inbred chromosome, for randomly distributed markers, based upon 1000 records comprising 40 replicates from each of the 25 combinations of brackets and population size described above. The prediction was made using generalized linear models, fitted using Genstat, with Poisson errors and a reciprocal link with a linear model based upon the expectations from the 'theory of runs' model described above:

$$E[J_D^{-1}]=a+bJ_\infty^{-1}+cB^{-1}+d(BJ_\infty)^{-1}. \tag{7}$$

3. Results

(i) *Validity of Stam's predictions*

Table 1 shows the average number of junctions present per Morgan in the final generation of the simulations, when the populations have become completely inbred, and compares the average values to those predicted by equation (1). Stam's predicted values, and the observed averages over 1000 replicates, are shown for inbred populations of size $N=2$, 5, 10, 25 and 50. The observed averages do not differ significantly from Stam's predictions, validating Stam's

https://doi.org/10.1017/S0016672305007329 Published online by Cambridge University Press

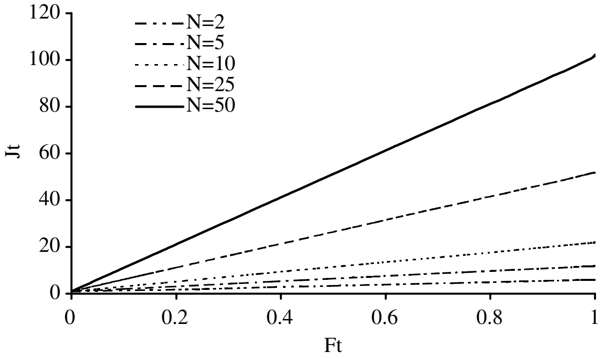

Fig. 2. Average junctions per Morgan (J_t) against inbreeding coefficient (F_t). The relationship between J_t and F_t is $J_t=(J_\infty-1)F_t+1$, and is derived analytically in Appendix.

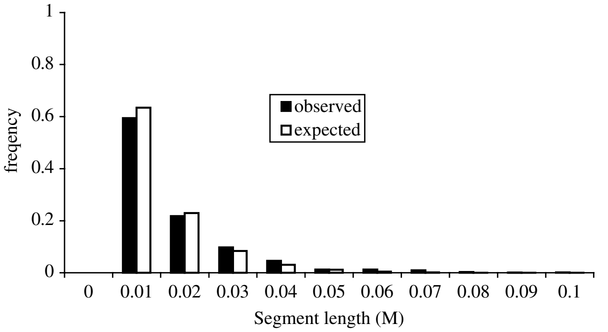

Fig. 3. Distribution of lengths of the first segment in inbred populations of size $N=50$. Expected distribution is a mixture of β distributions. See text for details.

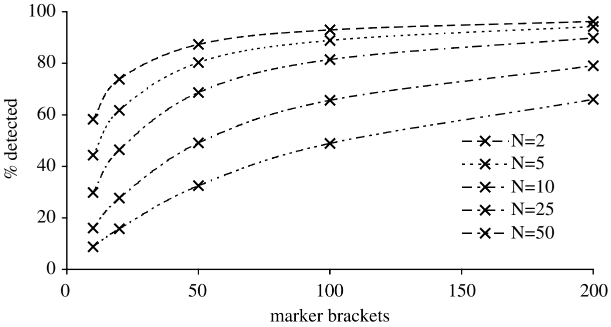

Fig. 4. Percentage of junctions recorded in inbred populations of fixed size N, with equidistant markers forming $B=10$, 20, 50, 100, 200 brackets. Junctions are recorded where adjacent markers originate from distinct ancestral chromosomes (see Section 2).

approach for the simulated populations examined here. Further simulations in larger populations confirmed the expected results. For example, a population of size $N=1000$, has an expectation of 2002 junctions per Morgan at inbreeding, hence 20·02 junctions per centimorgan. Simulations at $N=1000$ with a 1 cM chromosome gave an average junction count of 20·47 junctions over 100 replicates with a standard error of 0·629, suggesting that Stam's expectation is valid for larger populations.

Initially J_t increased rapidly over time, towards J_∞, and Fig. 2 shows that this increase was linearly related to the inbreeding coefficient F_t, with $J_t=F_t(J_\infty-1)+1$. This result can be confirmed by equating J_t to ΣH_j from generation $j=1$ to $j=t-1$, and is equivalent to equation 5 in Chapman & Thompson (2002). This result can also be derived analytically as illustrated in Appendix. Whilst the expected number of recombination events present on each chromosome will increase linearly with time (1 per generation per Morgan), fewer recombinations will result in junction formation in later generations, since they will occur between loci that are already IBD.

(ii) *Segment lengths*

Figure 3 shows the distributions of the lengths of the first segment in inbred chromosomes for populations of size $N=50$, and the expected distribution based on a mixture of β distributions, calculated as:

$$f(x)=\sum_{J_\infty} P(J_\infty)g(x;1,J_\infty)$$

where $g(x;1,J_\infty)$ is the β density function with parameters 1, J_∞, and $P(J_\infty)$ are the observed frequencies of J_∞ in the 1000 replicates. It is clear that the observed distribution is skewed towards the more extreme segment lengths (note that mean segment length is the same for observed and expected distributions), the actual distribution being more dispersed than expected, with extreme segment lengths over-represented, which can be confirmed by comparison with equation (3). Using a mixture of β distributions accounts for any variation in J_∞ over the replicates, so any deviation in the segment length frequencies will be due to the deviation of the distribution of segment lengths from a β distribution, as defined by equations (2) and (3). It may therefore be inferred that the distribution of segment lengths over these replicates does not follow the expected β distributions, suggesting that the junction locations are not at random along the chromosome, but tend to cluster in various locations.

(iii) *Assessment of junctions using markers*

The detection of junctions using equidistantly spaced markers became less efficient as N increased, and B decreased, as shown in Fig. 4. As marker density increased, the percentage of junctions detected increased towards 100%, but at a slower rate for the larger population sizes. The lowest percentage of junctions detected in the final generation for all combinations of N and B was 9·7% for $N=50$, $B=10$ (markers spaced every 10 cM). Even with marker spacing of 0·5 cM ($B=200$), the average percentage of

https://doi.org/10.1017/S0016672305007329 Published online by Cambridge University Press

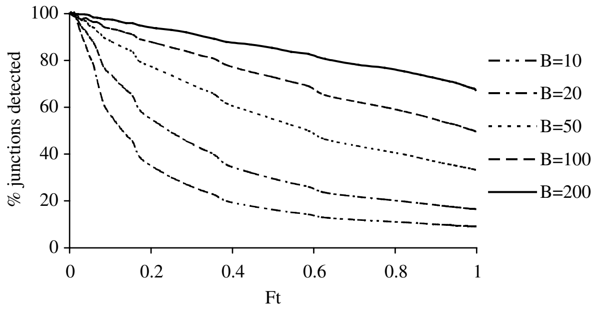

Fig. 5. Percentages of junctions detected in a population of size $N = 50$, proceeding to inbreeding, from various densities of marker maps. Markers are positioned equidistant along the chromosome and at both ends with junctions detected as described in Section 2.

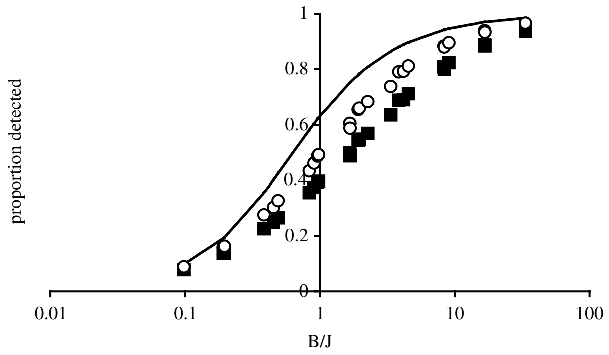

Fig. 6. The mean proportion of junctions detected for each simulated combination of B and N, for equidistant markers (open symbols) and randomly distributed markers (filled symbols), plotted against the bracket to junction ratio i.e. $1/\gamma$. The line represents the expected proportion for randomly distributed, fully identifiable junctions with equidistant markers i.e. $E[J_D/J_\infty] = \gamma^{-1}(1 - e^{-\gamma})$. Each point represents the mean of 5000 replicates.

junctions detected in an inbred population was as low as 66·9% for the parameters studied here.

The fraction of junctions present and detected at time t declined with time (see Fig. 5). Whilst the percentage detected declined approximately linearly with F for $B = 200$, for lower values of B, this decrease was non-linear, with the efficiency of detection decreasing very rapidly over the initial stages of inbreeding.

The fraction of junctions detected when the populations were fully inbred is shown in Fig. 6 for both marker distributions. For randomly distributed markers it predicts that marker maps require to be 20-fold denser than junctions to achieve a 90% detection rate, 5-fold more dense than may have been anticipated with naïve assumptions which result in the continuous line in Fig. 6: $E[J_D/J_\infty] = \gamma^{-1}(1 - e^{-\gamma})$. Even with informative equidistant needed the markers needed to be 2-fold denser than estimated from naïve assumptions. When marker density is random and of the same order as junction density only 40% of junctions may be detected.

The Human HapMap project (International HapMap Consortium, 2003) states as one of its initial aims to characterize one SNP every 5kb across the genome. Taking 1 Morgan ≈ 1000 kb, this is equivalent to 200 markers per centiMorgan. Looking at this density of markers in simulations with $N = 1000$, $L = 1$ cM, the expected number of junctions per chromosome in an inbred population is 20·02. One hundred simulations under these conditions gave an average of 20·47 junctions per centiMorgan. A randomly spaced marker map with $B = 200$ gives a value for γ of 0·1024, and an average of 88·22% of the junctions were detected over these 100 replicates. Compare this figure with the value for $N = 10$, $B = 200$ in Fig. 4. With an expected value of 22 junctions, and 201 markers distributed randomly across the 1·0 M chromosome, the value for γ is 0·11 and the percentage of junctions detected is 90·25%, confirming that junction detection is a function of γ, the ratio of junctions per Morgan to brackets per Morgan.

(iv) *Predictions of effectiveness of junction detection*

Table 2 shows the efficiency of junction detection for randomly distributed and equidistant markers of varying densities, for both $N = 2$ and $N = 50$ when the population was fully inbred with J_∞ junctions. In the cases examined, the number of junctions observed when using randomly distributed markers was lower than for equidistant markers. The expected relative efficiency of random to equidistant markers was $J_\infty\ (B + J_\infty - 1)^{-1}(1 - e^{-\gamma})^{-1}$ and for $B \gg J_\infty \gg 1$, this ratio will be approximately $(1 + \gamma)^{-1}$. However, in the most extreme case presented here ($B = 200$, $N = 50$, $\gamma = 0{\cdot}51$) the random markers did marginally better than predicted by this asymptotic ratio, with a relative efficiency compared to equidistant of 88%.

The number of junctions observed using the markers was much smaller than predicted (e.g. see Fig. 6). This difference was small when $N = 2$, but substantial when $N = 50$. In the latter case, only approximately 82% of the expected number, based on the predictions, were detected when $B = 200$; note this corresponds to approximately 68% and 56% of the actual number of junctions present for equidistant and random spacing, respectively.

There are a number of possible reasons for this discrepancy that were tested and will be exemplified by $B = 200$, $N = 50$, with an expected J_∞ of 102 and $\gamma = 0{\cdot}51$. Firstly, the predictions may be poor given the assumptions made; however, simple simulation of classical random occupancy, placing junctions at random across a number of brackets, suggested an error of less than 0·1% for the parameters investigated here, and cannot account for difference between J_D and $E[J_D]$ given by equation (6) for equidistant markers. Secondly, when multiple junctions occur

https://doi.org/10.1017/S0016672305007329 Published online by Cambridge University Press

Table 2. *Observed* (O) *and expected* (E) *junctions in inbred populations of size* $N=2$ *and* $N=50$, *when markers are either randomly distributed over the chromosome or placed at equidistant intervals*

	$N=2$, $E[J_\infty]=6$, Observed = 5·97				$N=50$, $E[J_\infty]=102$, Observed = 102·05			
	Random		Equidistant		Random		Equidistant	
B	E	O	E	O	E	O	E	O
10	3·74	2·91	4·25	3·51	9·18	8·00	10·00	9·01
20	4·56	3·79	4·99	4·40	16·82	14·07	19·85	16·60
50	5·30	4·77	5·55	5·25	33·64	27·01	43·25	33·31
100	5·61	5·28	5·75	5·57	50·52	40·54	63·60	50·27
200	5·78	5·59	5·86	5·77	67·51	55·94	79·64	67·33

Observed values are means from 5000 replicates. Expectations assume random distribution of junctions as described in the Materials and Methods. Standard errors of O vary between 0·02 and 0·04 for $N=2$ and 0·02 and 0·12 for $N=50$.

Table 3. *Regression coefficients for terms in generalised linear models used to predict* J_D

	Model Terms				
Model	Constant ($\times 10^3$)	$1/J_\infty$	$1/B$	$1/(J_\infty B)$	Mean Deviance (d.f.)
I	2·6 (0·3)	0·972 (0·017)	1·163 (0·025)	4·238 (0·799)	0·338 (996)
II	1·5 (0·3)	1·035 (0·014)	1·256 (0·019)	–	0·347 (997)
III	–	1·081 (0·011)	1·307 (0·017)	–	0·357 (998)
IV	–	1	1·394 (0·013)	–	0·376 (999)

The analytical model is described in Section 2 and equation (7). Model: I, full model; II, dropping term $1/(J_\infty B)$; III, dropping 'Constant'; and IV constraining coefficient for $1/J_\infty$ to equal 1.

within a bracket, there is a finite probability that the recombination events lead to both the markers that define the bracket originating from the same founder gamete and no junction is identified. Further analysis of the simulation results suggests that this accounts for approximately two-thirds of the shortfall between J_D and $E[J_D]$. Approximately 8 brackets that contained junctions were missed for this reason. This is greater than might have been expected from an *ad hoc* correction of $1/2N$, based on sampling from founder generations. Nevertheless, this leaves a shortfall (statistically very significant) of approximately 4 brackets between J_D and $E[J_D]$, which may be explained by greater multiplicity of junctions within brackets than predicted by random occupancy.

For randomly distributed markers, similar conclusions can be drawn (note that equation (4) is exact and not approximate). Thus the lower than expected detection can be explained by a combination of fewer than expected brackets containing junctions, due to an increased multiplicity of junctions within brackets, in addition to failure to identify junctions where they are present within brackets which appear to be inherited unbroken from a single founder chromosome.

Empirical predictions using generalized linear models in Genstat are shown in Table 3. In all four models, all terms displayed were statistically significant; however, the progression from I to IV was carried out with a view to parsimonious prediction. However, the dropping of terms changed the mean deviance only slightly. In models I to III, the terms $1/B$ and $1/J_\infty$ were the primary terms in predicting J_D; however, these models predict that not all junctions will be found even as B becomes very large. The term $(BJ_\infty)^{-1}$ has rapidly diminishing influence as B and J_∞ become large. Model IV was fitted to provide a predictive model in which $J_D/J_\infty \to 1$ as $B \to \infty$. Thus with model IV, $J_D/J_\infty \sim B/(B+1{\cdot}39J_\infty)$ or $1/(1+1{\cdot}39\gamma)$, predicting that for $J_\infty=202$, corresponding to $N \sim 100$, approximately 2500 randomly distributed markers would be required to detect 90% of the junctions, i.e. 12·5-fold denser markers than junctions. However, it should be recognized that progressing from model I to IV becomes increasingly optimistic in the relative marker density required. Nevertheless 20 replicates of ($N=100$, $B=2500$, $\gamma=0{\cdot}0808$) on a chromosome of length 1·0M resulted in a mean detection rate for junctions of 86% (SE 1%), in good agreement with model IV. Note that the value of γ for these simulations (202/2500) is close to the value described in Section (iii) above, where 200 markers were placed along 1 cM in a population of size $N=1000$, and a similar percentage

https://doi.org/10.1017/S0016672305007329 Published online by Cambridge University Press

of junctions was detected. These results suggest that the number of junctions detected is a function of the ratio of junctions to brackets in a particular section of chromosome, and is independent of the length of that particular segment.

This model can be applied to intermediate generations, by replacing J_∞ with J_t. We can predict $E[J_t]$ for given values of N and F_t, and using model IV, predict the number of these junctions we would expect to detect for a given marker map.

4. Discussion

This study has shown that the prediction by Stam (1980) of the expected number of internal junctions in an inbred population is accurate (there are no external junctions in inbred populations), and that furthermore the increase in such junctions over time is linearly related to the inbreeding coefficient. However, the distribution of the locations of junctions present in the inbred population is not random over the chromosome. This lack of randomness has considerable impact upon the effectiveness of detecting junctions when using a net of markers.

Any pair of homologous chromosomes in a partially inbred population will consist of IBD and non-IBD regions, where IBD is measured relative to some founder population. These regions will be separated by external junctions and will contain within them internal junctions, where a junction is a point on a chromosome in the current population where segments derived from two distinct founder chromosomes meet as a result of a recombination event at some point in the past. Stam (1980) derived an expression for the number of (internal) junctions that would be expected on a chromosome in a completely inbred population. Here we have shown that the number of junctions per chromosome in a partially inbred population is linearly related to the inbreeding coefficient, reaching Stam's expectation when the population becomes completely inbred. The expected number of junctions per Morgan, J_t, in a randomly mating population of size N at time t, with an inbreeding coefficient F_t is $J_t = F_t\ (J_\infty - 1) + 1$, where J_∞ is Stam's expectation for the number of junctions per Morgan in a completely inbred population (equation 1).

The evidence for non-randomness in the junction locations came from two sources: the distribution of segment lengths and variations in junction density. In a completely random dispersal of junctions across the chromosome, the distribution of segment lengths will be described by a β distribution with one of its two parameters determined by the number of junctions present (Waddington *et al.*, 2000). The observed distribution of the first segment length was found to be over-dispersed with an excess of longer segments, and since the mean was correct, an excess of shorter segments. Again with random dispersal of junctions over the chromosome, accurate predictions are available for the distribution of runs and occupancy of marker brackets. In all parameterizations studied, with random or equidistant markers, the extent of multiple occupancy by junctions of marker brackets was in excess of expectation. For $B = 200$ equal-sized marker brackets and $N = 50$, the excess multiple occupancy was approximately 10% of the expected number.

A mechanism for this clustering and over-dispersion can be advanced. There are two possibilities for non-randomness: (i) junctions have different survival probabilities; and (ii) junctions do not occur at random positions. In our neutral models the first can be dismissed since the survival probability of a newly formed junction (together with an arbitrarily small segment of chromosome either side of the junction) will be $1/2N$, independent of position: the size of the intact region will depend on N but nevertheless a region will exist. However, both a mechanism and supporting evidence can be found for non-random occurrence of junctions. As described by Stam (1980), junctions occur over time, decreasing in rate of appearance as heterozygosity decreases. At an intermediate stage of heterozygosity: (i) some regions will be fixed and IBD, and junctions cannot occur within such regions; (ii) crossovers will still occur, but are of no significance to the segregation of the variation since junctions can only occur where there is segregation among founder alleles. Regions that become fixed early will have been subject to fewer crossovers before fixation, and consequently would be expected to form longer IBD segments. Junctions formed at this intermediate stage that ultimately survive can only occur in increasingly smaller regions, thereby forming clusters with shorter segment lengths. Based on this hypothesis it would be predicted that markers that are fixed early belong to longer segments, and this is confirmed in Fig. 7.

It can be concluded that the processes of genetic drift and uniform recombination will result in a mechanism that will tend to concentrate the segregating genetic variation remaining from any given generation into localized regions. Whatever specific definition of 'haplotype block' one assumes, the underlying paradigm is the same: that in a given population of chromosomes, there are large regions of low haplotype diversity that are separated by smaller regions of higher diversity. Our results show that such an outcome can be observed in a population with uniform recombination rates, and give more detail than the explanations provided by Zhang *et al.* (2003) on the impact of drift and recombination.

Most results concerned with the effectiveness of detecting junctions using markers presented in this

https://doi.org/10.1017/S0016672305007329 Published online by Cambridge University Press

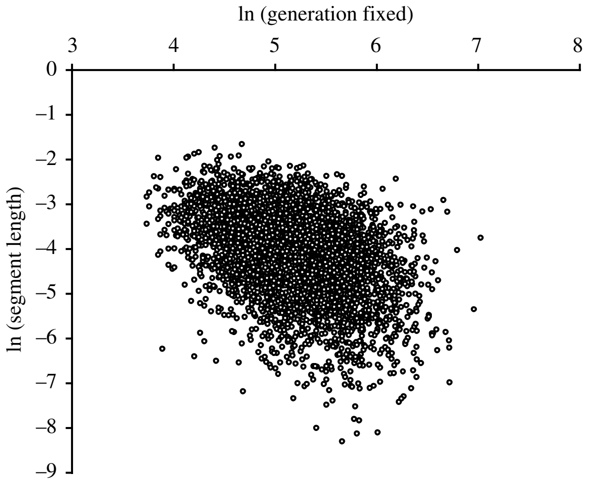

Fig. 7. The relationship between the logarithm of segment length containing a fixed, but randomly chosen locus and the logarithm of the time to fixation for the locus. The locus was positioned 15 cM from an end, and observed correlation from 5000 replicates was $-0{\cdot}42$.

study are derived from inbred lines, but they have both immediate and general applicability. Completely inbred lines are rare in nature, but they occur in laboratory populations, for example in recombinant inbred lines. Although most natural populations are partially inbred, with F_t values intermediate between 0 and 1, we have shown that for a randomly mating population on fixed size, excluding self-fertilization, that the average number of junctions per chromosome, J_t, is linearly related to the inbreeding coefficient at generation t. The number of junctions we can expect to detect for a given marker map can be calculated by substituting J_t for J_∞ in Section (iv) of the Results. For populations with different histories, the approach of Chapman & Thompson (2002) can be followed to calculate $E[J_t]=\sum_{j=0}^{t-1}H_j$ and the substitution made for J_∞.

For a marker bracket to be recorded as containing a junction, the markers at either end must derive from distinct ancestral chromosomes. In actual populations, this may be complicated, as markers in the current generation may be identical by state without being identical by descent, depending on the number and frequency of alleles in the founder population. In the simulated populations investigated here, each marker allele can be assigned unambiguously to a single chromosome in the founder population, and a junction inferred where consecutive markers derive from distinct ancestral chromosome. This inference will lead to an underestimation of the actual number of junctions for two reasons: (i) if $J_t > B$, then at least one or more brackets will contain multiple junctions, some of which will not be detected from the marker map; and (ii) a bracket containing two or more junctions may, by chance, have markers from the same ancestral chromosome at either end, and none of the junctions therein will be recorded. This second discrepancy can be overcome if there are sufficient markers to ensure that each junction is the only one in a bracket. However, as the population size and/or the inbreeding coefficient of the population increases, J_t will increase, and the number of markers needed to detect all of these junctions may be prohibitive.

The results have demonstrated the difficulty of detecting junctions in practice, since they can only be identified by using marker maps. The circumstances considered here have developed a theoretical upper bound, shown in Fig. 6, in which the markers are assumed to be equidistant and to be fully informative, in which the detection rate α is given by $\gamma^{-1}(1-e^{-\gamma})$ where $\gamma = J/B$ where expected J at any time t is given by $E[J_t]=F_t[2N+1]+1$ junctions per Morgan for the populations investigated here. This predicts that for approximately 90% detection rate, i.e. $\alpha = 0{\cdot}9$, $B \sim 5J$. However, our data have shown that this is considerably optimistic. Models based on our simulated data suggest that for randomly distributed markers $\alpha \sim 1/(1+1{\cdot}4\gamma)$, equivalently $B=1{\cdot}4J_t\alpha/(1-\alpha)$, and a 90% detection rate requires $B \sim 12{\cdot}5J_t$. Even these assumptions are likely to be conservative due to the strong assumption of perfectly informative markers, although this will be partially offset by the ability to design a marker map that moves towards equidistance.

The impact of this is the scale of marker maps needed to get a complete picture of a junction map. Consider the simulated population used by Zhang *et al.* (2003) where a population of $N_e = 10\,000$ was inbred for 10 000 generations. In this case $F_t \sim 0{\cdot}4$, giving $E[J_t] \sim 8000$ junctions per morgan, hence approximately 13 junctions over the region of 0·17 cM investigated. This suggests that at least 170 markers would be needed over this region to detect 90% of the junctions. The study by Zhang *et al.* (2003) used 10 markers in this region and did not have the objective of defining junctions; nevertheless it demonstrates the difficulty of reliable extrapolation from such data. In practice the density of markers required to detect junctions will influence the reliability of tracking ancestral chromosome segments and thus potentially of locating functional polymorphisms that they contain.

Appendix. Relationship between inbreeding coefficient and the mean number of junctions per Morgan

Following the argument of Stam (1980), the fate of a new junction is the same as that of a neutral mutation. It may become fixed or lost in the end, but its expected frequency remains constant (1 copy). The rate at which new junctions forms is H_t, and the expected

https://doi.org/10.1017/S0016672305007329 Published online by Cambridge University Press

number of junctions in any generation is thus

$$J_t = \sum_{j=0}^{t-1} H_j.$$

The linear relationship between $F_t(=1-H_t)$ and J_t can most easily be seen by allowing self-fertilization (we will return to the exclusion of selfing later).

$$H_{t+1} = \left(1 - \frac{1}{2N}\right) H_t$$

$$H_t = \lambda^t H_0 \ (\lambda = 1 - 1/2\mathrm{N})$$

or, since $H_0 = 1$,

$$H_t = \lambda^t$$

The expected number of junctions, J_t, thus equals

$$\begin{aligned} J_t &= \sum_{j=0}^{t-1} H_j \\ &= \sum_{j=0}^{t-1} \lambda^j \\ &= \frac{1-\lambda^t}{1-\lambda} \\ &= \frac{1}{1-\lambda}(1-H_t) \\ &= 2N(1-H_t) \\ &= 2NF_t, \end{aligned}$$

a simple linear relationship.

Now consider the case where selfing is excluded. Again following Stam,

$$H_t = A\lambda_1^t + B\lambda_3^t,$$

where

$$\lambda_1 = \frac{N-1+\sqrt{N^2+1}}{2N},$$

$$\lambda_3 = \frac{N-1-\sqrt{N^2+1}}{2N},$$

$$A = \frac{1-\lambda_3}{\lambda_1 - \lambda_3},$$

$$B = -\frac{1-\lambda_1}{\lambda_1-\lambda_3}.$$

[Note: checking Stam's equation for the expected limiting number of junctions we see that

$$J_\infty \sum_{t=0}^{\infty} H_t = A\frac{1}{1-\lambda_1} + B\frac{1}{1-\lambda_3},$$

which simplifies to $2(N+1)$.]

Now consider an intermediate generation.

$$\begin{aligned} J_t &= \sum_{j=0}^{t-1} H_j = A\sum_{j=0}^{t-1} \lambda_1^j + B\sum_{j=0}^{t=1} \lambda_3^j \\ &= A\frac{1-\lambda_1^t}{1-\lambda_1} + B\frac{1-\lambda_3^t}{1-\lambda_3} \\ &= \frac{A}{1-\lambda_1} + \frac{B}{1-\lambda_3} - \left(\frac{1}{1-\lambda_1}\right)A\lambda_1^t - \left(\frac{1}{1-\lambda_3}\right)B\lambda_3^t. \end{aligned}$$

From the above note, the first two terms sum to $2(N+1)$. Since λ_3 is small and negative, after a few generations $A\lambda_1^t$ is the dominant term in the expression for H_t, so H_t can be approximated by $A\lambda_1^t$ without serious error. After a few generations, we may write

$$\begin{aligned} J_t &\approx 2(N+1) - \frac{1}{1-\lambda_1}A\lambda_1^t \\ &\approx 2(N+1) - \frac{1}{1-\lambda_1}H_t \\ &= 2(N+1) - \frac{2N}{N+1-\sqrt{N^2+1}}H_t \\ &\approx 2(N+1) - (2N+1)H_t \\ &= (2N+1)(1-H_t)+1 \end{aligned}$$

or

$$J_t = (J_\infty - 1)F_t + 1.$$

This work is funded by the BBSRC. We are grateful to Mr Dave Waddington for discussions on some of the statistical aspects of the analysis, to Professor W. G. Hill and two anonymous referees for comments on earlier versions of the manuscript, and to Professor P. Stam for further useful comments and providing the derivation in the Appendix.

References

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

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

Daly, M. J., Rioux, J. D., Schaffner, S. E., Hudson, T. J. & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics* **29**, 229–232.

Feller, W. (1967). *An Introduction to Probability Theory and its Applications*. 3rd edition. New York: Wiley.

Fisher, R. A. (1954). A fuller theory of junctions in inbreeding. *Heredity* **8**, 187–197.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**, 789–796.

Jeffreys, A. J., Kauppi, L. & Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* **29**, 217–222.

Johnson, G. C. L., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C. J., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C. L., Clayton, D. G. & Todd, J. A. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics* **29**, 233–237.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T. N., Norris, M. C., Sheehan, J. B., Shen, N. P., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. A. & Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723.

Phillips, M. S., Lawrence, R., Sachidanandam, R., Morris, A. P., Balding, D. J., Donaldson, M. A., Studebaker, J. F., Ankener, W. M., Alfisi, S. V., Kuo, F. S., Camisa, A. L., Pazorov, V., Scott, K. E., Carey, B. J., Faith, J., Katari, G., Bhatti, H. A., Cyr, J. M., Derohannessian, V., Elosua, C., Forman, A. M., Grecco, N. M., Hock, C. R., Kuebler, J. M., Lathrop, J. A., Mockler, M. A., Nachtman, E. P., Restine, S. L., Varde, S. A., Hozza, M. J., Gelfand, C. A., Broxholme, J., Abecasis, G. R., Boyce-Jacino, M. T. & Cardon, L. R. (2003). Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genetics* **33**, 382–387.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R. & Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature* **411**, 199–204.

Stam, P. (1980). The distribution of the genome identical by descent in finite random mating populations. *Genetical Research* **35**, 131–135.

Waddington, D., Springbett, A. J. & Burt, D. W. (2000). A chromosome based model for estimating the number of conserved segments between pairs of species from comparative genetic maps. *Genetics* **154**, 323–332.

Zhang, K., Akey, J. M., Wang, N., Xiong, M., Chakraborty, R. & Jin, L. (2003). Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. *Human Genetics* **113**, 51–59.

https://doi.org/10.1017/S0016672305007329 Published online by Cambridge University Press