Dependence in the survival of ancestral genomes

Elizabeth Thompson*
University of Washington, Seattle, USA- eathomp@uw.edu

Abstract

Study of the descent of genome in defined pedigrees underlies many genetic analyses, including the survival of founder DNA in the complex pedigrees of managed endangered species. It has long been known that, across a chromosome, descent of genome through the m meioses of a defined pedigree may be represented as a random walk on the vertices of an m-dimensional hypercube. At any single genome location, survival of a specified founder genome must decrease the probability of survival of others, the highest negative correlations in survival being between genomes in a single diploid founder, and next within a founder couple. Across a chromosome the reverse is true. The survival of an ancestral DNA segment from a founder greatly increases the probability of survival of a segment from an adjacent founder genome, where adjacency is in terms of the vertices of the hypercube. Results have practical application in studying the diversity of founder genomes present in key current individuals (for example, in a clone), in studying the survival of introgressed genomes, and the effect of both breeding choices and natural selection for or against such genomes on the survival of other founder genomes.

Keywords: genome survival; random walk model; endangered species; ancestral diversity.

1. Introduction

Under the process of meiosis, segments of genome are copied from generation to generation. The genome of a current individual consists of segments of ancestral DNA from ancestors at a specified time-depth, or from founders of a defined pedigree. At a single genome location (hence as a genome-wide expectation) methods for studying the extinction of genes in a defined pedigree are well established (Thompson, 1983). Likewise, the framework for studying the descent of genome segments was developed by Donnelly (1983). However, the combination of complex pedigree and genome-wide analysis has not been prioritized. However, the potential availability of genome-wide genetic data, and the potential to restore extinct genetic variation by cloning of historical species members have re-awakened interest in methods for analysis of survival of genome. A particular case of interest is that of the Przewalski horse (*Equus przewalskii*), a species descended by a complex pedigree from just 13 founder members, 12 of whom lived over 100 years ago. In 2020, a clone of an animal born in 1975 was successfully produced.

2. Methods and models

Within a defined pedigree of a diploid organism the inheritance of genome at a location x is most easily specified by binary *meiosis indicators*. The meioses are the transmission of DNA from parent to sperm or egg cell, and each non-founder individual in the pedigree results from two meioses. In a defined pedigree we may label the m meioses i, $i = 1, ..., m$ and

$$\begin{aligned} S_i(x) &= 1 \text{ if in meiosis } i \text{ the parent's paternal DNA is transmitted} \\ S_i(x) &= 0 \text{ if in meiosis } i \text{ the parent's maternal DNA is transmitted} \end{aligned}$$

Then Mendel's First Law (Mendel, 1866) states that for each x and each i

$$P(S_i(x) = 0) \; = \; P(S_i(x) = 1) \; = \; 1/2$$

and that meioses i are independent. Further, in the absence of genetic interference, switches between the two binary states of $S_i(x)$ occur as a Poisson process rate 1 per Morgan (defining this unit of genetic distance) so that

$$P(S_i(x+y) \neq S_i(x)) \quad = \quad (1/2)(1-\exp(-2y)) \qquad \text{(Haldane, 1919).}$$

Donnelly (1983) represented the above process as a continuous-time random walk on the vertices of an m-dimensional hypercube: each vertex is a binary m-vector, and each component of the vector switches independently at rate 1, so the total rate of leaving a vertex is m, and the time (i.e. length of genome) until a change is exponential with mean $1/m$. This enabled him to derive many results on the survival and extinction of genome, but the key one for our purpose here is that the probability that in a stretch of haploid genome length L Morgans, some part descends in a direct line to a k^{th}-generation descendant is $Q_k \approx (1 - \exp(-kL/2^k))$.

3. Ancestry of genome in an individual

Consider first a current individual, *ego*, and the representation of the genomes of its k^{th}-generation ancestors in *ego*'s genome. For now, we ignore that some of these ancestors may be the same individual, and since *ego*'s paternal genome is an independent replicate of the process leading to the maternal genome we consider only the *ego*'s maternal genome (Figure 1).

At k generations, *ego* has 2^{k-1} diploid maternal ancestors, or 2^k haploid ancestral genomes. The 2^k lineages are k-digit binary vectors; the point probability of each descent is $1/2^k$. In a genome length L, the total expected length from a given haploid ancestral genome is $L/2^k$, the probability an ancestral genomes is not represented is $(1 - Q_k)$ or approximately $\exp(-kL/2^k)$. The number of segments is approximately Poisson with mean $kL/2^k$, with each segment length exponential with mean length $1/k$ (Thompson, 2013).

Whereas, at a single genome location, conditioning on the survival of any subset of founder genomes can only decrease the survival of other founder genomes, across the genome the reverse is true. Survival of a segment of genome from a founder will increase the survival probability of an adjacent segment from founders adjacent in the ancestry. To be more specific, consider the survival of the two founder genomes within a k^{th}-generation ancestor (Figure 1). Neither genome survives, if there is a no segment surviving from the genome they contributed to their offspring at generation $(k-1)$. The ratio of the probability that neither survives relative to the product of the probabilities for each is thus

$$\frac{(1-Q_{k-1})}{(1-Q_k)^2} \; - \; \frac{\exp(-(k-1)L/2^{k-1}}{\exp(-2kL/2^k)} \quad = \quad \exp(L/2^k) \; > \; 1$$

Likewise $Q_{k-1} > Q_k^2$ showing the positive dependence in survival and in extinction between the two haploid genomes within a founder individual.

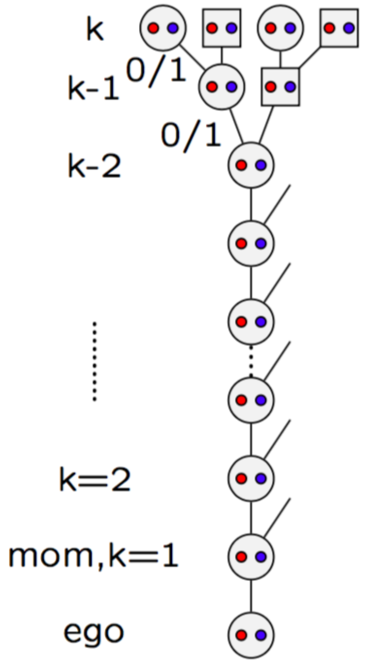

Figure 1: The maternal k^{th} generation maternal ancestry of an individual

This argument extends also to the four genomes of a founder couple at generation k. Segments from all four will survive in *ego* if a segment survives from the descendant haploid genome at generation $(k-2)$ and then is there is at least one recombination within the segment in the meiosis from generation $(k-1)$ to $(k-2)$, and additionally within each of the two subsegments transmitted from generation k to $(k-1)$. Segments descending over k generations have exponential lengths mean $1/k$ Morgans, and in any meiosis the probability of no recombination in a segment of length y is $\exp(-y)$. Integrating over the genome segment lengths and switch locations to obtain the probability of the 3 required switch events we obtain probability $1/(k^2(k-1))$. This result makes sense: there must be three independent switch events, one occurring in a segment at level $(k-1)$ (probability $1/(k-1)$) and two in segments at level k (probability $1/k$). Since $Q_{k-2}/(k^2(k-1)) > Q_k^4$ we again have a pattern of positive dependence in the survival of genome from adjacent ancestors.

Recall that recombination switches in an ancestral lineage length k correspond to transitions on the vertices of the k-dimensional hypercube. Thus adjacency in this space does not correspond to the usual concept of pedigree adjacency. A recombination is equally likely to happen at any meiosis in the lineage. However, the two genomes within a founder do differ by a single meiosis (at generation k), while the four genomes in a founder couple in an maternal lineage corresponding to vertices 00...000, 00...001, 00...010, and 00...011 and so differ by at most 2.

Conditional on survival of a given genome, say w.l.o.g 00000...000, one may consider the conditional probability of survival of other genomes at a given distance from this one. Figure 2 shows these probabilities over a continuous genome of length L Morgans, with L ranging from 1 to 33 Morgans (the approximate length of the human autosomal genome).

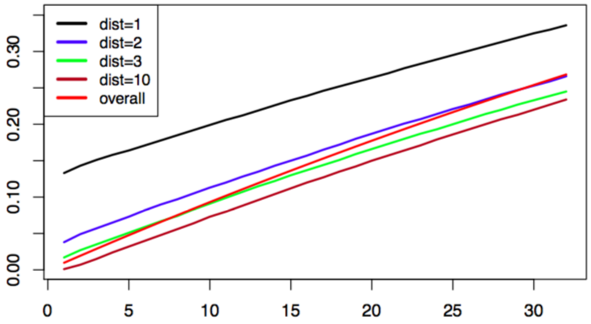

Figure 2: Conditional on the survival of a specified genome, the survival probabilities of other genomes at sppecified meosis count distance over a genome length L Morgans

4. Number of represented ancestral genomes and the effect of chromosome size

The probability that a haploid founder genome at generation k is represented in a length L genome in *ego* is $Q_k \approx (1-\exp(-kL/2^k))$, Donnelly (1983). Since *ego* has 2^{k-1} diploid maternal ancestors, the expected number of haploid genomes represented is $2^k Q_k \approx kL$ for large k. This is consistent with previous results, as the chance of more than one ancestral segment surviving from a given generation-k genome is close to 0, and each surviving segment has expected length $1/k$. However, as seen above (Figure 2) the represented founder genomes will be clustered in the space of hypercube vertices. In reality the autosomal genome is broken into some number of chromosomes, of varying length. This has a significant impact on the total number of founder genomes represented, since at the start of each new chromosome, the random walk on the hypercube will restart at a randomly selected vertex. An approximation due to G.Coop that accommodates the number of chromosomes C reflecting these random restarts is $Q_k = 1-\exp(-(C+kL)/2^k)$, with k adjusted here to reflect our count of maternal founder haploid genomes.

In particular we considered the case $k=10$, and the number of founder genomes (out of $2^k = 1,024$ that are represented in the maternal genome of *ego* (Figure 1). We studied the number of new founder genomes contributing to *ego*'s maternal genome, on each successive chromosome. Initially, a new genome is encountered with each new chromosome, since the chance that a random start is at an already visited vertex is small. There is an additional small effect, in that founder genomes close to the new start are also unlikely to have been previously visited. However, with each successive chromosome, as the set of visited hypercube vertices become less sparse, there is less effect of each random restart. Further, we found that, conditional on the mean number of new founder genomes encountered on a chromosome in *ego*, the distribution of the number is very close to Poisson.

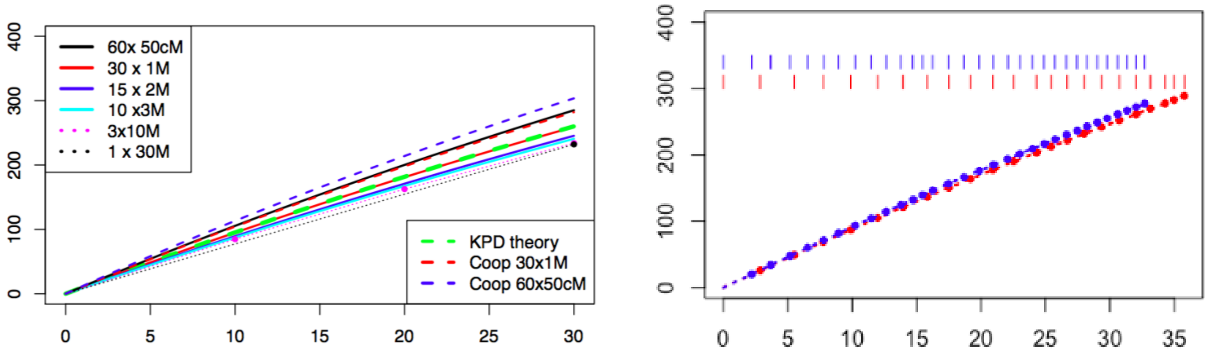

Figure 3: The effect of chromosome length on the expected number of 10^{th} generation ancestral haploid genomes represented in the maternal genome of a descendant. Left: counts at varying partitions of a genome of length 30 Morgans. Right: the results for chromosomes of varying length representing the human and the Przewalski horse genomes.

We considered the case of a genome length 30 Morgans broken into chromosomes of varying length. The one that most close mimics the human genome would be 30 chromosomes each of length 1 Morgan. In expectation fewer than 300 ancestral genomes ($\approx$ 30%) will be represented in a typical-length genome 10 generations later. As reasoned above, and shown in Figure 3 (Left) more chromosomes for the same total length results in the representation of more ancestral genomes, but the effect is not large. Both theoretical approximations for a continuous genome overestimate founder representation. Indeed, Donnelly (1983) clarifies why the two approximations in his formulae, ignoring chromosomes and ignoring clustering of hits on a given founder genome in the random walk process, largely compensate, making his approximation very accurate for genomes comparable to that of humans, here represented by the 30 chromosomes of length 1M. The additional correction due to G.Coop overcompensates, leading to a curve following that for 60 Chromosomes of length 0.5M. (The biological process of meiosis requires that every chromosome has genetic length at least 0.5M).

The right panel of Figure 3 shows the expected 10-generation founder genome counts for the human and the Przewalski horse pedigree. Since there is no genetic distance map of the Przewalski genome, translation from base-pair counts to Morgans was made using a non-interference genetic map (Haldane, 1919), but conditioning on a chiasma in the meiosis tetrad resulting in all chromosomes having length at least 0.5M. The horse genome is about 10% shorter than the human one, but has more short chromosomes, with 8 pairs of autosomes having genetic length barely above 0.5M. The effects of this are seen in Figure 3 (Right)). The longer human genome (shown in red) results in more founder genomes being represented in *ego*, but at a given total length the shorter horse chromosomes (in blue) result in slightly more ancestors being represented.

5. Gene extinction and survival in the Przewalski Horse

The Przewalski Horse provides a unique case study for analysis of genome extinction and survival in a managed population of known pedigree. For a summary of the earlier history of the species up to 1988, see Geyer et al. (1989). Although the world-wide species now numbers over 2,000, the entire population descends from just 13 founders, 12 of whom lived in the early 1900's and the remaining one, a wild-caught mare, contributing to the population from 1960. Additionally one of the early founders is known to have been a Mongolian domestic horse, while another is suspected of being an F_1 hybrid.

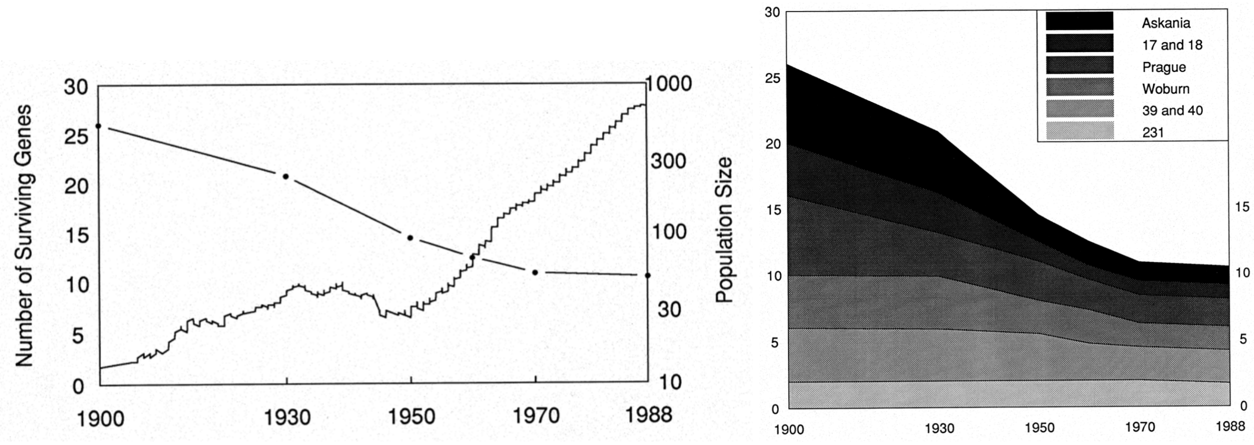

Figure 4: The history of the Przewalki Horse world population to 1900 to 1988. Left: the expected pointwise survival of the 26 original founder genomes, and the population on a log scale, Right: the survival of genome of the different founder groups. Figures from Geyer et al. (1989)

The genomes of *Equus przewalskii* and all other horse breeds (*Equus caballus*) are very similar, the major difference being a centromeric fusion of two Przewalski chromosomes: Przewalski horses have a total of 66 chromosomes (32 autosomal pairs, and 2 sex chromosomes) while all other horses have 64. Thus fertile hybrids and introgression from other horses into the Przewalski population is possible. Of course, there are many genetic variants present in only one of the two species, and very significant allele frequency differences for other variants. Although introgression and other factors have perhaps reduced potential differentiation, the picture is very broadly comparable to the difference between humans and chimpanzees: the large human chromosome-2 is the result of a centromeric fusion of two chtomosomes ancestral to chimpanzees.

Geyer et al. (1988,1989) carried out extensive analysis of single-locus gene survival and extinction in the Przewalski horse pedigree from the founders to 1988, using, among other methods, algorithms for computation gene extinction probabilities for joint sets of founders, developed earlier by Thompson (1983). Single-locus computations of course provide also genome-wide expectations. Geyer and Thompson (1988) studied also dependence in survival between founder genomes: at any single locus survival of some subset of founder genomes can only decrease the survival probabilities of other founder genomes.

Figure 4 shows some results from Geyer et al. (1989). The total population size (shown on a log scale) was very small until 1950. Although it increased markedly from 1950 to 1970, there was little attempt at genetic management, and founder genetic material continues to be lost. After 1970 numbers increased, and genetic management strategies were implemented with very little continuing loss. The righthand-panel shows the expected genome loss in different founder subgroups, showing high variation. One group that has survived poorly are the pair (#17,#18): #18 is the suspected hybrid. Proportionately, the Prague group has also lost significant founder genome: this group contains the known domestic mare.

6. Genome of a Przewalski Horse clone

In August 2020, a clone of a Przewalski horse (# 615) who lived from 1975-1998 was produced from a frozen tissue cell, born to a surrogate domestic mare, and now thrives in San Diego (https://sandiegozoowildlifealliance.org/pr/kurtandholly). This animal and the genetic variation he represents has re-awakened interest in analysis of surviving genome in the Przewalski Horse. Additionally, the availability of genomic data informing the ancestry of segments of genome in individuals, broaden the scope for developing strategies to maintain genetic variation in the species, and indeed to restore lost or poorly represented variation by selective choices of animals for future cloning.

The pedigree of #615 is shown in Figure 5. The horse was brought from Europe to USA, but was not extensively bred in USA due to his known domestic horse ancestry. However, the horse represents seven original Przewalski founders, and the five other than the known domestic horse (#229) and suspected hybrid (#18) are founders not well represented in the USA population, and constituting a significant portion of surviving Przewalski horse genetic variation (Figure 4). It is therefore of particular interest to consider the genome of #615, both in relation to the founders, and in relation to the current population.

Note first there is nothing intrinsically abnormal in integrating a cloned animal into the current population: it is simply as if #615 had survived to the current day. On the other hand, as a horse at a generation about midway between the founders and the current day, the clone will have founder genome segments on average about twice as long as those in other current individuals.

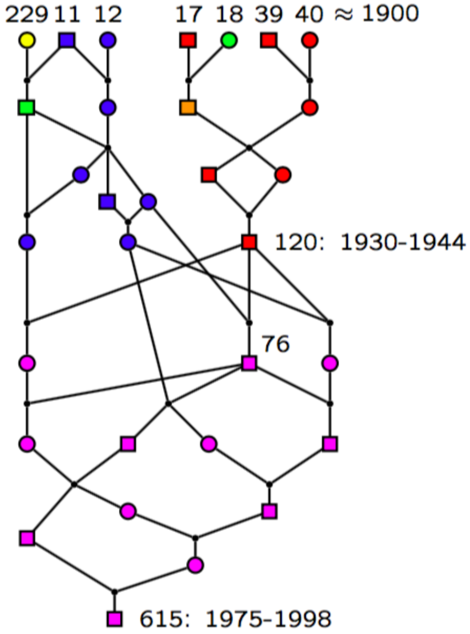

Figure 5: The pedigree of the Przewalksi Horse clone

Due to the symmetry and restricted descent from the four early founders (#17, #18, #39, #40), in analyses of the founder genome segments of #615 we consider only 8 genomes, the pairs of genomes in each of the ”Prague group” (#11, #12, #229; Figure 4), and in #120. The clone has inbreeding coefficient 0.3687, and kinships to these four ancestors of 0.11562, 0.0723, 0.0840 and 0.234 respectively: #120 himself has inbreeding coefficient 0.25 in his descent from the other four founders. There are 22 meioses in descent to #615 from the 1930 level of #120, and 14 additional meioses in the earlier descent from the Prague founders. But each of these 2^{36} binary descent paths leads to one of the 8 founder genomes considered. The pattern of clustering of each founder genome in this high-dimensional space will affect the patterns of founder segments present in #615.

On the left part of Table 1, the expected genome contrbutions of the founders to the maternal and the paternal genome of the clone are given both as genome length and segment number. While #120

mat/pat of founder	% genome		% segments
	maternal	paternal	
# 11	31.3	31.3	33.7
# 12	14.8	14.1	17.1
# 229	16.4	17.1	19.5
# 120	37.5	37.5	29.7

	11m	11p	12m	12p	229m	229p	120m	120p
11m	0	0.26	0.10	0.10	0.12	0.12	0.15	0.15
11p	0.26	0	0.10	0.10	0.12	0.12	0.15	0.15
12m	0.20	0.20	0	0.17	0.085	0.085	0.13	0.13
12p	0.20	0.20	0.17	0	0.085	0.085	0.13	0.13
229m	0.20	0.20	0.075	0.075	0	0.17	0.14	0.14
229p	0.20	0.20	0.075	0.075	0.17	0	0.14	0.14
120m	0.17	0.17	0.075	0.075	0.09	0.09	0	0.33
120p	0.17	0.17	0.075	0.075	0.09	0.09	0.33	0

Table 1: Left: Summary of founder genome represented in the maternal and paternal genome of the clone. Right: Overall transition probabilities between founder genomes across the genome of the clone.

has the largest contribution (although of course this is divided among his 4 founder ancestors), #11 contributes more separate segments. The domestic horse #229 contributes slightly more than #12, both in terms of genome length and segments, but interestingly #12 contributes relatively more to the clone's maternal genome, and #229 to the paternal, wheras #11 and #120 contribute equally to both the clone's maternal and paternal genomes. On average there are 6.4 segments per Morgan in the clone's maternal genome, and 5.7 segments per Morgan in the clone's paternal genome. By contrast, there are about 12-14 founder genome segments per Morgan in other current Przewalski horses.

The right part of Table 1 shows overall average transition probabilities between founder genomes represented across the genome of the clone. Note this process is not Markov, as the state space of founder genomes is a very collapsed version of the underlying space of meiosis transitions. However, te matrix shows the relatively high probabilities of transition between the two genomes of each founder, and the slightly higher probabilities of transitions from #12 and #229 to #11 rather than to #120. In considering #120, because of the increased probabilities of adjacent chromosome segments from genealogically adjacent founders, segments from #17, and to a lesser extent #39 and #40 are likely to carry with them segments from the hybrid #18, while segments from founder #11 and to a lesser extent #12 are likely to bring introgression from the domestic horse #229 (Figure 5). Conversely, introgressed domestic horse segments, which may be detected from the genetic variants they carry. are likely adjacent to relatively long Przewalski segments from otherwise under-represented Przewalski founders.

The analyses in this paper are prior probabilities based only on pedigree information. When suffient genome-wide genetic data are available, it will be of considerable interest to examine the realized genomes of the clone and available reletives, particularly with regard to introgressed domestic horse genome segments from #229 or #18.

Acknowledgment

I am grateful to Dr Oliver Ryder of the San Diego Zoo Wildlife Alliance for introducing me to the Przewalski horse pedigree and studies, for collaboration over many years, and for his continuing interest in this work.

References

Donnelly, K. P. (1983) The probability that related individuals share some section of genome identical by descent. Theor. Pop. Biol. 23: 3463.

Geyer C.J. and Thompson E.A. (1988). Gene survival in the Asian Wild Horse (Equus przewalskii): I. Dependence in gene survival in the Calgary Breeding Group pedigree. Zoo Biology7, 313-327.

Geyer C. J., Thompson E. A. and Ryder O. A. (1989) Gene survival in the Asian Wild Horse (Equus przewalskii): II. Gene survival in the whole population, in subgroups, and through history. Zoo Biology, 8, 313-329.

Haldane, J. B. S. (1919) The combination of linkage values and the calculation of distances between the loci of linked factors, Journal of Genetics, 8, 229–309.

Mendel, G., (1866) *Experiments in Plant Hybridisation: English translation and commentary by R. A. Fisher (1965).* Ed, J.H.Bennett, Oliver and Boyd, Edinburgh, UK.

Thompson (1983) Gene extinction and allelic origins in complex genealogies. Proc Roy.Soc.(Lond.) B 219, 241-251.

Thompson, E. A. (2013) Identity by descent: Variation in meiosis, across genomes, and in populations. Gen 194: 301-326.