Genetic importance and genomic descent

E.A.Thompson,
Department of Statistics, GN-22
University of Washington,
Seattle, WA 98195.

Abstract

Considerations of the conservation of the gene pool of a species have lead to the idea of "genetic importance" -- sets of individuals whose loss would deplete the gene pool, and who therefore should be accorded opportunity to breed. In this brief paper, two aspects of the idea of genetic importance are addressed. First it is important to recognise that it is not individuals singly, but groups of animals that are likely to be of genetic importance. It is the "genetic importance" of the total set of animals at risk of not contributing further to future generations that is relevant to the maintenance of the gene pool. Second, analyses of the conservation of gene pools have previously been restricted to single Mendelian loci. The genome is effectively a continuum of linked genes; failure to take this into account may lead to misleading results. A preliminary approach to the analysis of the conservation of genomes is presented.

Introduction

Preservation of the genetic variability of a species entails preservation of its gene pool to the maximal extent possible, by avoiding loss of surviving founder genes. Analyses of founder contributions and simulation analyses of gene extinction (MacCluer et al., 1986) have provided useful information on expected numbers of surviving genes in different segments of a small population whose pedigree is known. However, two types of dependence limit the usefulness of analyses that focus upon expectations. The first is dependence between individuals and the second is dependence between gene loci. Aspects of each of these are discussed in this paper.

Dependence between individuals in the survival of genes itself has two aspects. The first, discussed by Geyer and Thompson (1988) is dependence between founders, within and between founder lineages, in the survival of their genes. If genes from one founder or ancestral group survive in the population, or in any specified subset of the population, others are less likely to survive also. In the case of the Przewalski's horse pedigree, this leads to a variance of the number of distinct surviving founder genes that is only about 2/3 of the variance obtained on an assumption of independence. Over independent Mendelian loci, the variability in the number of distinct surviving founder genes is significantly less than is suggested by an analysis of survival probabilities of each founder gene separately.

In this paper, however, the focus is on a different aspect of dependence between individuals. The objective of retaining currently surviving genes has lead to the development of the idea of *genetic importance.* The genetic importance of an individual is a measure of the probability that it carries founder genes that are currently at risk of loss (MacCluer et al., 1986). More precisely, it can be defined, for a given individual and a single Mendelian locus, as the expected number of distinct founder genes carried that are not present elsewhere in the population. Such genes will be lost with the loss of the individual unless they are first transmitted to offspring. Due to genealogical structure, a founder gene may be limited to a small segment of the population, although it may be present in several copies within that subgroup. Because of this dependence between related individuals, it is necessary to extend the idea of genetic importance to groups of individuals whose collective loss may deplete the pool of surviving founder genes. For example, a founder gene that does not appear to be at risk may be limited to a single zoo collection. Even major subdivisions of a population may be at risk of collective loss from the management of the species, due to policy or politics. The Przewalski's horse pedigree provides two examples of subdivision, where the genetic importance of large groups of individuals may be key in developing a global management plan. This is the question addressed in section 2 of the paper.

Previous approaches and analyses to the study of gene survival in small populations have considered independently segregating Mendelian loci (Lacy et al., Chapter 7). Moreover, there has been a tendency to interpret a gene survival probability, p, as survival of a proportion p of the genome. An average over independent realisations is converted, explicitly or implicitly, to an average over the genome of an individual. As an expectation over genomes among individuals of identical ancestry, this is interpretation cannot be faulted. However, it provides a misleading picture of the distribution of the amount of original founder genome surviving in an individual or group of individuals. The genome is not a collection of independently segregating genes. All organisms have chromosomes; many have at least some large chromosomes. Failure to consider the genomic continuum may lead to misleading perceptions of gene identity and survival. Those endangered species for which detailed pedigree information is available typically have a small population size, and significant gene loss may occur over even a few generations. For such population sizes and time scales, the dependencies and disequilibria between closely linked genes will be substantial. Fisher (1949) introduced his *theory of junctions* as the first attempt to "move from a point theory to a strand theory" of population genetics. This theory provides a framework for the study of segregation of chromosome segments in a pedigree. The second objective of this paper is to present preliminary ideas in this direction.

2. Genetic importance of population subdivisions.

In order to evaluate the genetic importance of individuals or subgroups of a population of known pedigree, methods for computation of the probabilities of patterns of gene survival in large and complex pedigrees are required. These methods have been described elsewhere (Thompson, 1983) and have been applied principally to the Przewalski's horse pedigree. Geyer et al. (1989) analysed gene survival in population subgroups, and through history, in this large and complex pedigree. Additionally, they addressed

questions of "genetic importance", and extended the idea of genetic importance to pairs and small groups of individuals who collectively carry founder genes not present outside the group. Here we use the computations of Geyer et al. (1989) to consider a still wider view of genetic importance -- the genetic importance of the major subdivisions of the world population of Przewalski horses.

Historically, in the management of the Przewalski horse, there have been two ways of dividing the world population. The first is geographic. While there has been a North American management plan for some years (ref. SSP), it was only in 1990 that there was agreement on a global management plan, and the logistics of global management remain complex. In advocating global management, one issue was the genetic importance of those animals outside the population managed by the North American SSP. Numerically, this is the majority of the world population, but it was not clear how many distinct original founder genes could be lost by failure to manage the non-North American population. The second subdivision, which cuts across the geographic one, relates to the known domestic horse ancestry in many members of the species. The management policy considers separately animals having some descent from the known Mongolian domestic mare (DOM), and those that do not (non-DOM). The original policy would have lead to the eventual extinction of all of the founder genes currently present only in DOM-descended individuals. The genetic importance of the DOM-descended horses measures the cost in terms of the Przewalski's horse genes that would thereby be lost. These considerations have provided input in the development of a global management plan.

Table 1A shows the expected numbers of founder genes surviving in these four major world subdivisions. In these computations the DOM-genes are accounted separately; genetic importance is assessed with respect to the "desirable" non-DOM genes, whether in non-DOM or in DOM-descended animals. From Table 1B, we see that the world population carries some genes not present in the North American population; 1.24 of the total 9.91 surviving non-DOM founder genes would be lost if only the North American population were to survive. Among the non-DOM descended horses, the non-North American population contributes 0.76 genes of the total 7.30 expected distinct surviving founder genes. Thus more than 10% of surviving distinct original founder genes exist only outside the North American population; this is the extent of the genetic importance of the population outside of North America.

However, the genetic importance of the DOM-descended horses is substantially greater. From Table 1C, we see that both within North America, and within the total world population, about 25% of the distinct founder non-DOM genes are present only in DOM-descended horses. The DOM-descended horses are genetically important for these non-DOM genes. The non-DOM genes for which the DOM-descended horses are genetically important are principally genes from the Old Prague founder lineage (two Przewalski's horse founders and DOM). This is not surprising; through the original mating of DOM and subsequent population bottleneck, the DOM and Old Prague genes are completely confounded. One the other hand, note that there are a substantial number of non-Old-Prague genes for which the DOM-descended horses are genetically important. Due to matings early in the pedigree, this population subgroup received genes of other founders which did not segregate also into the non-DOM descended segment of the population.

3. Dependence between loci; a model for chromosome segments.

The above analyses have considered only single Mendelian loci. We turn now to the effects of recombination and linkage on questions of gene descent and gene survival. The simplest model of the recombination process is that crossovers occur as a Poisson process at some rate $lambda$. This model gives rise to independently exponentially distributed intervals (mean $1/ lambda $) between crossovers. The model does not allow for interference, and does not explicitly recognise heterogeneity of recombination along the chromosome. However, if chromosome lengths are measured in units of recombination distance, any such heterogeneity is implicitly subsumed. This model (proposed by Fisher, 1949) is a useful approximation to the process; it is a good model on the larger scale, although not on the small scale of intragenic processes. For simplicity, we do not here consider the potentially different recombination rates in males and females, although this type of heterogeneity could be modelled. If chromosomal distances are measured in Morgans (Haldane, 1919), the rate of the Poisson process is 1. (That is, $lambda =1 $.)

In this first analysis of genome inheritance under this model, we consider just two results, each of which typifies broader questions. The first is a question of shared genomes between relatives. The single-

locus gene identity by descent probabilities for relatives provide, of course, the expected proportions of genomes shared, but the proportion is a random variable whose distribution is a function of the pedigree and the recombination process. Secondly, there are questions of proportions of genome surviving. Again, the single-locus computations provide the expected proportion, but the distribution is more complex.

One of the simplest relationships between a pair of individuals involves the segregation from a single parent to two offspring. Due to the structure of small subpopulations of many species, this halfsib relationship is often of interest in conservation genetics. Maternal halfsibs arise as seedlings from a single maternal plant. Paternal halfsibs are cohort members in species where each group has a single breeding male. It is well known that at each locus, the offspring of a common mother share their maternal gene with probability 1/2; the expected proportion of maternal genome shared is 1/2. However, we wish to consider the distribution of this proportion, and, in particular, its variance.

Consider first a single chromosome length L Morgans. In the segregation to each offspring, the recombination process provides a maternal chromosome which is a sequence of segments deriving alternately from the maternal and paternal chromosomes within the mother. The *junctions* between these segments arise as a Poisson process of rate 1. Comparing the two realisations of this process (figure 1), we find alternating chromosome segments which are identical-by-descent and non-identical-by-descent between the two half-sib offspring. The points of interchange are the combined set of junctions in the two segregations, and thus form a Poisson process of rate 2 (figure 1). The total number of interchanges, N, is Poisson, with mean $2L$. The number of identical-by-descent segments, K, is $(N+1)/2$ if N is odd, and is $N/2$ or $(N+2)/2$, each with probability $half$, if N is even.

Conditional on N, the locations of interchanges are the order statistic of a sample of size N from a uniform distribution on an interval length L. Thus, standard theory provides a beta distribution for the proportion, Q, of shared genome; $Be(K , N+1-K) $. Thus, in this simple case, the exact distribution of length, QL, of chromosome shared is easily found. For an organism with k chromosomes, of lengths $L sub 1 , ... , L sub k $, the result is a sum of independent random variables, $Q sub i L sub i $, where $Q sub i$ has, conditional on $N sub i$ the appropriate beta distribution. For simplicity we consider here only the mean and variance of this distribution.

For a single chromosome length L, conditional on N, the mean proportion of chromosome shared is $1 over 2$ and the variance $1 over 4(N+2)$ if N is odd. If N is even, we have mean either $N over 2(N+1)$ or $N+2 over 2(N+1)$ and variance (in either case) $N over {4 (N+1) sup 2 }$. Combining these gives mean $1 over 2$ and variance $1 over 4(N+1)$. The unconditional variance is thus difficult to obtain exactly, but it is approximately $ 1 over 4 bE (1 over N+1) $, where N is Poisson with mean 2L; i.e. $(1- e sup -2L) /8L $. This expression reduces to the single-locus Binomial variance $1 over 4$ when L is small, and is approximately $1/8L$ for long chromosomes. Summing over independently segregating chromosomes the length of shared genome has mean $sum from i L sub i /2$ and variance $sum from i ˜L sub i (1- e sup {-2 L sub i })/8$.

It is of interest to compare this result with that obtained for a set of independently segregating unlinked loci. For m independent loci, the number of maternal genes shared by a pair of half sibs has a Binomial distribution, $Bi (m, half) $. The mean proportion of the loci shared is $1/2$, and the variance of this proportion is $1/4m$. In terms of the mean, 1 Morgan translates to 1 locus, but in terms of variance, in a long chromosome, the "rate of exchange" is 2 Morgans/locus. The exact "rate of exchange" depends on the individual chromosome lengths, as well as on the total genome length. Thus there is no simple way to convert a continuous genome into an "equivalent" number of independently segregating loci.

As an example of analysis of survival of genomes, rather than single genes, we consider the probability that the entire genome of an individual is present among its offspring. It will seldom be possible to preserve the entire genomes of individuals over multiple generations; some proportion of a founder's genome will be lost. However, it remains an interesting question as to how many offspring are required in order to have a reasonable chance of preserving the entire genome of an individual at least to the offspring generation. The answer, of course, depends on the length of the genome in question, and on its subdivision into chromosomes: as above, very small chromosomes will behave much like independently segregating genes.

The number of offspring required to have high probability of retaining entire genomes does have some practical relevance. A new founder may enter a highly inbred established population, as did the

founder "231" into the Przewalski's horse pedigree in the 1950s. For such a founder, attempting to preserve as much genome as possible may be a reasonable goal; if 10 offspring are required (for example) this may be realistic for one animal, even though not all members of the population can be permitted to be so prolific. A second application lies in frozen embryos; there the limitation of habitat for living animals does not apply. To ask how many embryos should be stored is a reasonable question, to which a practical answer can be given.

The required formula was given by Donnelly (1983); the probability that the whole of a chromosome-pair length L is preserved among the k offspring is approximately $exp(-kL/2 sup k-1)$. Again, it is of interest to compare this result with that for a set of unlinked loci. The probability that both parental homologous alleles survive at any given locus is $ q~=~1 ~-~ 2 sup -k+1 $, and this is thus also expected proportion of the total haploid genome at which both parental homologues survive. The probability that both alleles survive at all of m independently segregating loci is thus $q sup m $ or approximately $exp (-m/2 sup k-1)$. Note that the dependence on the number of offspring k is different in these two expressions; there is no number of independently segregating loci that can be considered "equivalent" for all k to a chromosome of length L. Donnelly (1983) gave also a more exact formula: $ exp (- (k(k-3)L + J (k-1))/(k-2)2 sup k-1)$ for a set of J pairs of autosomes of total length L. When k is not small this is effectively equal to $ exp (- (kL + J)/2 sup k-1)$, which, in the limit for small L, gives the independent loci result, whereas for a small number of large chromosomes gives the previous approximation $exp(-kL/2 sup k-1)$. Note that, although the number of chromosome pairs (J), and the total length (L), have some effect on these probabilities, the relative lengths of chromosome enters only as a higher order correction. Donnelly (1983) verified that any correction due to different relative lengths of chromosome is indeed slight.

Table 2 gives some numerical results, for varying sizes of genome and numbers of offspring. (The human genome is approximated by the middle combination; about 20 pairs of autosomes and total length about 30 Morgans.) Except for small numbers of offspring the dependence of the number of autosome pairs is slight; the major factor in the probability of passing on one's entire genome is the total length, measured in Morgans. With only 4 offspring there is little chance of passing on the whole of even the smallest example genome, consisting of 10 pairs of autosomes of total length 10 Morgans. With 16 offspring, there is a high probability of passing on the whole of even a very large genome, consisting of 30 pairs of autosomes of total length 60 Morgans. Of greater significance is that for 8 offspring the probabilities remain low, but are very substantially bigger for 12 offspring. For this question of preserving an entire parental genome, this is the critical range of offspring numbers. If one requires a high probability of passing on an entire genome, at least 12 offspring are necessary. For many species this is not a practical proposition, even for a few key individuals.

The question of survival of entire genomes to offspring can, of course, be extended. What is the chance of preserving at least x% of the genome if there are k offspring? What is the chance of preserving genomes among grandoffspring? H. Blossey (pers. comm.) has found that whereas impractically large numbers of offspring may be required to retain whole genomes, a substantially smaller number, such as 4 or 5, is required for high probabilities of preserving at least 90% of parental genome.

4. Discussion

In analyses of founder gene survival, the idea of "genetic importance" was first developed as a measure of individual animals, relating to single Mendelian loci (MacCluer et al., 1986). In this paper, two extensions of this idea are reviewed briefly. The first develops genetic importance as a characteristic of groups of individuals, or of major subsets of the population, as well as of single individuals. It is the survival of genes in the collective population than is of primary interest in preserving potential genetic variability. It is the collective loss of all genes that are currently represented only in individuals who may not breed that causes loss of variability.

The second direction of extension of analyses of gene survival is from measures relating to single loci to measures relating to the continuous genome of DNA strands. In conserving the genetic variation in a population, we need to think in terms of conservation of genomes, not just of individual genes. In this paper we consider only two preliminary questions; the distribution of the amount of genome shared by maternal half-sibs, and the probability that all the genome of an individual survives among its offspring.

One important general point emerges: There is no "equivalent Mendelian gene number" for genomes. The correspondence between genome length and loci numbers depends on the number of pairs of chromosomes, on the total chromosome length, and on the relative lengths. Moreover, it depends on the question; mean survival, variance of survival, and probabilities of total survival each provides a different "rate of exchange".

Acknowledgement

This research was supported in part by NSF grant BSR-8921839. I am grateful to Charles Geyer for the computations for Table 1, and for many helpful discussions.

References

Donnelly K.P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theor. Pop. Biol.* **23,** 34-63.

Fisher R.A. (1949). *The Theory of Inbreeding.* Academic Press, New York.

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 Biology* **7,** 313-327.

Geyer C. J., Thompson E. A. and O. A. Ryder. (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. *J. Genet.,* **8,** 299-309.

MacCluer, J. W., VandeBerg, J. L., Read, B. and Ryder O. A. (1986). Pedigree Analysis by Computer Simulation. *Zoo Biology,* **5,** 147-160.

SSP for Przewalski's horse. -- details ??

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

Table 1: Expected number of surviving founder genes within, and consequent $genetic˜importance sup 1$ of, some sub-groups of the Przewalski horse world population.

	World population		North-American population	
	Total world Population	Non-$DOM sup 2$ world Population	Total N.A. Population	Non-$DOM sup 2$ N.A. Population
A. Expected surviving genes:				
DOM genes	0.57	0.00	0.45	0.00
Non-$DOM sup 3$ genes	9.91	7.30	8.67	6.64
Total genes	10.48	7.30	9.13	6.64

B. Genetic importance of non-North-American horses:		
	All horses	non-$DOM sup 2$ horses
DOM genes	0.11	0.00
Non-$DOM sup 3$ genes	1.24	0.76
Total genes	1.36	0.76

C. Genetic importance of DOM-descended horses:		
	World	North America
DOM-genes	0.57	0.45
Other Old-Prague genes	1.55	1.27
Other founders' genes	1.06	0.77
Total genes	3.18	2.49

1. Genetic importance is the expected number of surviving non-DOM genes present only within the specified subgroup of the population.

2. Non-DOM horses are animals who do not have descent from DOM, the Mongolian domestic mare.

3. Non-DOM genes are genes that are not identical by descent to the genes in DOM.

Table 2: Probability that all of an individual's genome survives among its offspring.

		number of offspring			
J=number of autosome pairs	L=total length of genome (Morgans)	4	8	12	16
10	10	0.0126	0.5423	0.9436	0.9952
	30	0.0001	0.1913	0.8491	0.9862
	60	0.0000	0.0401	0.7249	0.9728
20	10	0.0019	0.4950	0.9385	0.9948
	30	0.0000	0.1747	0.8446	0.9858
	60	0.0000	0.0366	0.7210	0.9725
30	10	0.0003	0.4519	0.9334	0.9945
	30	0.0000	0.1595	0.8400	0.9855
	60	0.0000	0.0334	0.7171	0.9722