Whole genome sequence data: Characteristics resulting from demography and sequencing errors

I.M. MacLeod[*], B.J. Hayes[†] and M.E. Goddard[*†]

Introduction

The cost of genome sequencing is dropping rapidly so that it will soon be possible to fully sequence many individuals within a species. This whole genome sequence data (WGS) should be of great value to population genetics. It will give unprecedented detail concerning the history of a population and evidence of selection. However, in order to interpret WGS we need to know what to expect given a particular population history. For instance, to test for evidence of selection one would compare the observed data with data generated under a null hypothesis with no selection. In practice the observed data must be summarised by a manageable number of parameters so that it can be compared to that expected under a given set of assumptions. Inevitably, the high throughput sequencing methods and terabases of data produced come with a penalty of errors. For effective use of WGS, for example to estimate the population history, these errors must be screened out of the data without removing too high a proportion of the real SNPs.

Two important characteristics of sequence data are the frequency of heterozygous sites and the multi-locus linkage disequilibria (LD) between segregating sites. Both these characteristics can be described jointly in the following manner: Take two homologous chromosomes at random from the population and record the sites at which they differ (that is, sites that would be heterozygous if these two chromosomes were from the same individual) and the distances between adjacent sites. The distribution of the distances between heterozygous sites contains information about the rate of heterozygosity and LD. The distribution measures LD because long distances between heterozygous sites occur when the two chromosomes are identical by descent (IBD) over this part of the chromosome. It is the inheritance of IBD chromosome segments by multiple individuals that generates LD.

It is possible to generate the distribution of distances between heterozygous sites under a given set of assumptions by simulating sequence data, but this can be prohibitively costly in computer time for many animals of a species with a large genome. Recently we published an analytical method to predict the distribution of distances between heterozygous sites (in pairs of haplotypes) given the past history of effective population size (Macleod et al 2009). In this paper we use this method to predict the expected distribution of homozygous segment lengths, generate a large random sample of "observed" segments across the genome, and then test the effect of sequencing errors on the observed pattern of homozygous segments in sequence data.

[*] Melbourne School of Land and Environment, University of Melbourne, Victoria 3010, Australia

[†] Biosciences Research Division, Department of Primary Industries, Victoria, 3038, Australia

Material and methods

We use the analytical predictor developed by MacLeod *et al* (2009) to predict HH_n which is the probability of observing a run of homozygosity $\geq n$ bases when a pair chromosomes (either within the same animal or taken from different animals) are compared. The parameters in the model are: effective population size (*Ne)* at various times in the past (*G* generations), as well as per base pair (bp) mutation (μ) and recombination rates (r).

We contrast HH_n in two different populations and for both we used $\mu = r = 1x10^{-8}$ and chose Ne so that they share the same single loci heterozygosity (i.e. both have equal average length of homozygous runs in pairs of haplotypes). The contrasting populations are:

1. *"Bovine" population*: with previously estimated bovine demography (de Roos et al 2008) with changing effective population size. The parameters below describe from left to right, the most ancestral *Ne* (for given *G*) to the present day Ne:

Ne	90,000	20,000	9,000	6,000	3,000	2,000	1,000	300	120	90
G	900,000	70,000	20,000	7,000	2,000	700	270	20	8	2

2. *"Constant" population*; constant effective population size in equilibrium - Ne=11,125 and G=1,000,000.

Secondly, having predicted HH_n based on these 2 sets of parameters, we illustrate the sequence data obtained from these two populations by simulating the sequence of a pair of chromosomes where the heterozygous sites are marked. To do this we use HH_n to calculate F(n) the proportion of homozygopus runs that are < n bases long by $F(n) = 1 - \bar{h}(P(n) - P(n+1))$ where $\bar{h}$ is the average homozygous length (1/single locus heterozygosity), F(n) is the cumulative distribution of the length of homozygous runs so we can sample from this distribution by sampling uniform, random probabilities from 0 to 1 and back transforming these to homozygous segment lengths using the cumulative distribution ie $n = F^{-1}(x)$ where $x \sim U(0,1)$. We sample 500,000 of these lengths of homozygous runs and join them sequentially to simulate pairs of chromosomes.

Finally, we illustrate the effect of sequencing errors (where a homozygous site is incorrectly called heterozygous) on these simulated chromosome pairs by randomly adding false SNPs at a rate of 1 error into each window of 223,880 bp along a 75Mb long chromosome.

Results and discussion

Figure 1 shows the HH_n distributions for the two contrasting populations. The small recent population size in the Bovine model results in more chance of finding long homozygous segments which have not yet been broken up by mutation and recombination. For example there is a 10% chance of observing lengths of ≥ 0.3Mb in Bovine while in the Constant size population there is only a 0.4% chance of observing lengths ≥ 0.3Mb. Figure 1 demonstrates how the analytical prediction based on a specified demography with no selection could be generated and then compared with real observed WGS runs of homozygosity. The analytical predictor of HH_n can also be used to work backwards from an observed pattern of HH_n to define the best fit demography in the absence of selection (MacLeod, Hayes, Meuwissen *et al* 2009).

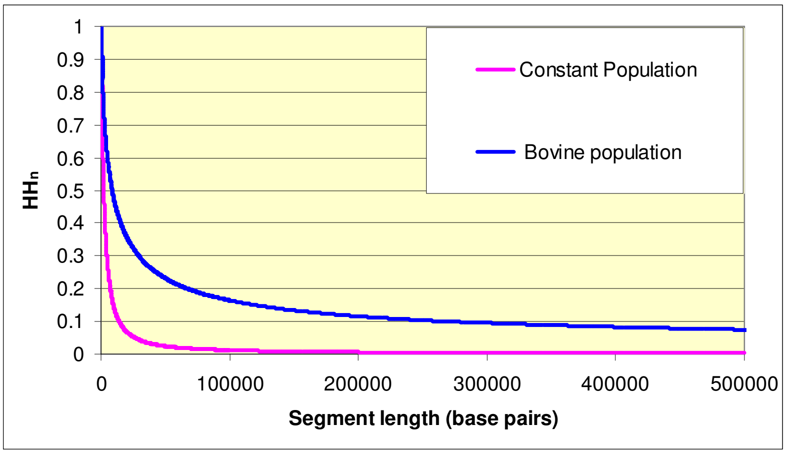

Figure 1. HHn or the probability of finding runs of homozygosity of ≥ n bases in the Bovine and Constant populations models.

Figure 2A shows two pairs of chromosomes from each population. Again this demonstrates very clearly how demography of the population will affect the expected distribution of heterozygous sites along the chromosome. In the Constant population most of the distances between heterozygous sites are short (< 50,000bp) with relatively little variation, while the Bovine chromosome pairs 1 and 2 show much greater variation in homozygous segment lengths. In Bovine data the very large ancestral population size results in many extremely short homozygous segments but also rarer long segments are present due to the recent very small population size. The observation of some very long homozygous runs could be mistakenly attributed to selection if the population history is not well specified for the null hypothesis.

The distribution in Fig 2A is that of error free data, while in reality high throughput sequencing technology results in a range of data errors including base pairs wrongly identified as heterozygous when in fact the pair is homozygous. Figure 2B shows the same chromosome pairs but with these “false SNPs” superimposed on them.. The result of this relatively low level of errors is quite striking. The y axis scales used in Fig 2A & B are different to draw attention to the fact that these errors have the strongest effect on the very long homozygous segments. Short homozygous segments are rarely disrupted by one error per 223,880 bp but long homozygous segments are cut into many short segments. This largely destroys the evidence for recent low *Ne*. Fig 2B could be wrongly interpreted as showing regions of low recombination and low mutation if the false SNPs go undetected.

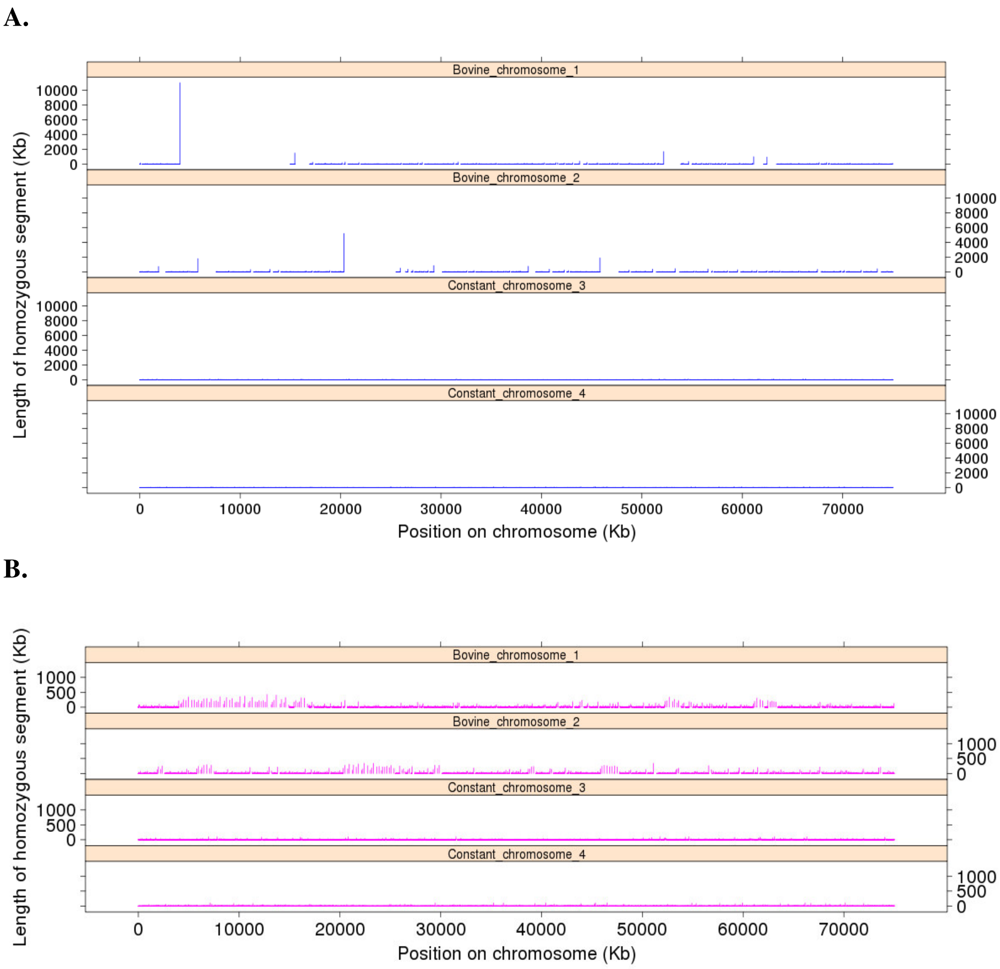

Figure 2 A & B: Distribution of homozygous segments sampled along 2 simulated chromosomes for each of the Bovine and Constant population models. A is error free data while B displays data from A with random errors superimposed.

Conclusion

This study demonstrates the importance of understanding the demographic history when trying to interpret patterns of homozygosity in WGS. It also highlights the likely effect of undetected false SNPs in the data, and importance of developing methods to account for and minimise errors when analysing genome sequence data.

References

de Roos, A.P.W., Hayes, B.J., Spelman, R.J., *et al.* (2008) *Genetics* 179, 1503-1512.
MacLeod, I.M., Meuwissen, T.H.E, Hayes, B.J., *et al*. (2009) *Genet. Res.* 91, 413-426.