Whole-genome sequencing data offer insights into human demography

Jonathan K Pritchard

Two new studies take distinct population genetic approaches to analyzing whole-genome sequencing data sets in order to estimate human demographic parameters. These papers refine our understanding of the relationships among human populations while illustrating both the possibilities and the statistical challenges of fitting demographic models to whole-genome data sets.

These are exciting times for human population genetics, as the ever-increasing number of human genome sequences promise to add greatly to our understanding of the evolution of modern humans[1]. The new genome sequence data will also shed light on other long-standing evolutionary questions such as the extent of recent human adaptation[2]. However, the availability of these large-scale data sets raises statistical and computational challenges. On page 1031 of this issue, Adam Siepel and colleagues[3] report new statistical methods to estimate the relationships among populations from whole-genome sequence data of multiple individuals. In a related paper, Li *et al.* report a different new method for historical inference, but using one genome sequence at a time[4]. Both papers represent important methodological advances in their ability to estimate detailed demographic information from whole-genome sequences. These studies refine several key aspects in human demographic models, including the timing of the population split between Africans and non-Africans.

Demographic models

Analysis of genetic variation within and between populations is an important tool for studying population demographic histories over the past tens to hundreds of thousands of years. Many aspects of population history can influence patterns of genetic variation observed in current data sets. For example, the amount of genetic differentiation between populations can be used to estimate population divergence times. Similarly, population sizes, as well as changes in size, affect the rate of genetic drift within populations and consequently help to determine the abundance and allele frequency distributions of variable sites[5].

Modern-day patterns of genetic variation result from a combination of historical and evolutionary processes, including demographic history as well as mutation, recombination and genetic drift. These processes are usually modeled using a stochastic process known as the 'coalescent' that describes the ancestral relationships among the observed sequences in any given region of the genome[6]. The goal of demographic inference is to use the patterns of genetic variation observed in genetic data to infer the historical parameters that produced those data. In recent years, a great deal of effort has gone into developing statistical approaches for doing this.

These approaches include so-called full-likelihood methods that aim to compute or approximate the probability of observing the precise configuration of genetic variation in a given data set, assuming the coalescent model and a given set of demographic parameters[7,8]. Such methods are appealing because according to statistical theory they make maximal use of all the information present in the data. However, despite progress, full-likelihood methods remain computationally demanding, especially for large data sets. There has also been interest in methods that compress the data into simpler summaries such as the allele frequency spectrum[9], albeit with some loss of information. Finally, there are several methods that rely on clever simplifications of the likelihood function to make it computationally tractable[10,11]. Such simplifications have proven valuable for problems including haplotype phasing, genotype imputation and estimation of fine-scale recombination rates.

Using whole genomes

The two new papers[3,4] both tackle the important question of how to scale up these statistical methods so as to take advantage of whole-genome sequence data (**Fig. 1**). Gronau *et al.*[3] analyze recently published whole-genome sequences of six individuals using a Bayesian coalescent-based approach to infer human genealogies. The individuals represent diverse ancestries and populations: three from sub-Saharan Africa, two from east Asia and one from Europe. Their primary interest is to improve estimates of the historical relationships among major human groups, including population sizes, divergence times and migration rates. Their computational approach is most closely related to the full likelihood methods described above, as they use a computational technique called Markov-chain Monte Carlo (MCMC) to sample from among the possible coalescent genealogies that could have generated the observed data in each genomic region. From this they estimate posterior distributions of the population demographic parameters. To make this ambitious scheme computationally feasible, they make two important simplifications. First, they analyze only a small fraction of the genome (approximately 37 Mb, divided into 1-kb regions located far from genes in order to reduce the impact of selection[12]). Second, because they consider segments of just 1 kb, they show that intralocus recombination

Jonathan K. Pritchard is at the Department of Human Genetics, University of Chicago, Chicago, USA.
e-mail: pritch@uchicago.edu

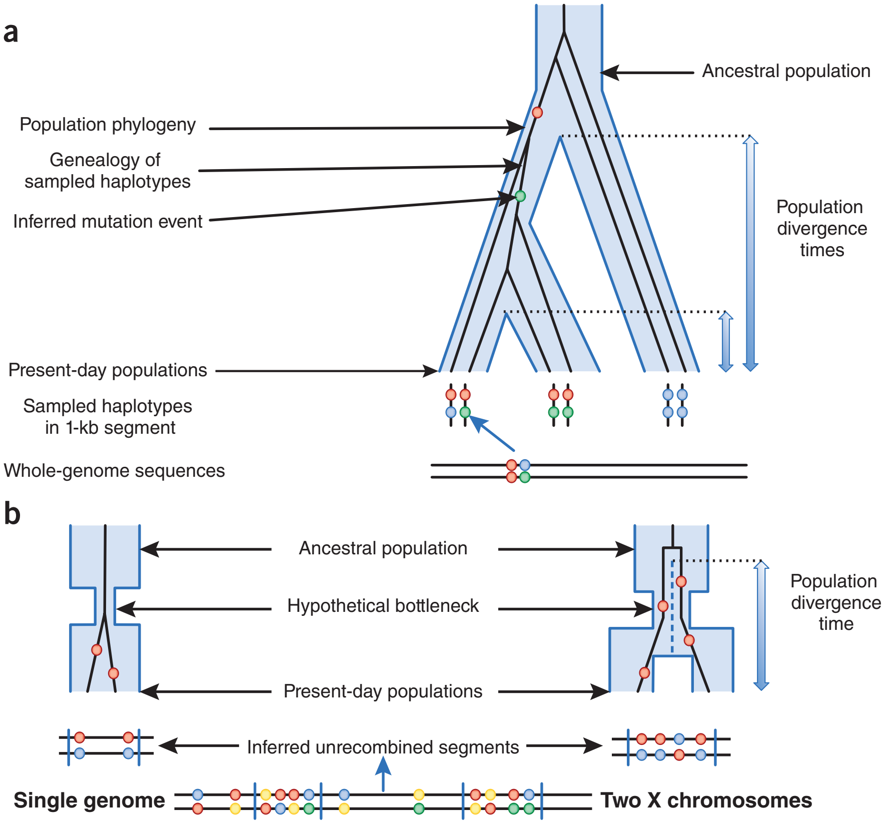

Figure 1 Illustration of methods for demographic inference from whole-genome sequence data sets. The two methods illustrated, reported by Gronau *et al.*[3] and Li *et al.*[4], use coalescent genealogies (black lines) to infer the demographic histories of sampled populations (here represented by the blue bands). Observed sequence differences are indicated by colored dots. (**a**) Gronau *et al.*[3] start by extracting 1-kb regions from across the genomes of representatives of the sampled populations. They then use a computational method to sample from the set of possible genealogies connecting the 1-kb haplotypes. This helps them to estimate historical parameters for the sampled populations, including divergence times, ancestral population sizes and migration rates. (**b**) Li *et al.*[4] use a hidden Markov model to estimate the distribution of coalescence times of genomic segments from a single individual, and thereby the history of the population sizes. According to theory, coalescent events occur at higher rates whenever the population size is small; thus, there would be particularly high rates of coalescence during population bottlenecks. To study population divergence, Li *et al.* pair X chromosomes from males from two different populations (African and non-African). Their analysis suggests that splitting occurred gradually (dashed line) with a corresponding low (but nonzero) rate of coalescence prior to the final separation of the populations.

is sufficiently rare that it can be ignored in this analysis. Their approach depends most strongly on the pattern of mutations within each segment: these provide partial information about the coalescent genealogies at each locus, which are informative about the underlying model parameters.

In contrast, the new method by Li *et al.*[4] uses the patterns of variation within a single genome to infer the history of population size changes over the past million years. They make few assumptions about ancestral population size, instead allowing size to vary over time in a flexible way (similar to Bayesian Skyline methods; ref. 13). Their method relies on the pattern of heterozygous sites observed in an individual whole-genome sequence. According to theory, recent coalescent events between an individual's two copies of each chromosome are likely to produce long stretches of low heterozygosity, whereas more ancient coalescent events are likely to produce short stretches of high heterozygosity. The rate of coalescent events in any given epoch is then informative about effective population size at that time, with small population sizes resulting in high rates of coalescence. Li *et al.*[4] use a model, based on a slightly modified version of the standard coalescent and fit using a computationally efficient hidden Markov model, to estimate the history of population size changes.

Dating the out-of-Africa split

One question addressed by both papers is the timing of the population split between Africans and non-Africans, a date of key importance in human history. Although it is now widely accepted that the ancestors of non-Africans emerged from Africa in the past 100,000 years, there remains considerable uncertainty as to the precise timing. Gronau *et al.*[3] estimate that this split occurred relatively recently, ~50,000 years ago (95% Bayesian credible interval 38,000–64,000 years), with minimal subsequent migration. Li *et al.*[4] consider this split by pairing X chromosomes from African and non-African males, and they suggest a slightly earlier split (~60,000–80,000 years ago). Importantly, Li *et al.* infer substantial migration until as recently as 20,000 years ago, suggesting that the population separation occurred gradually.

These estimates substantially reduce the uncertainty for this key date while highlighting the difficulty of estimating split times in the presence of migration. Li *et al.*[4] also estimate that the well-known bottleneck in non-African populations was strongest at around 20,000–40,000 years ago, and therefore the original migration event out of Africa probably did not directly cause this bottleneck, as has often been assumed (but see also ref. 14). Finally, another key result from Gronau *et al.* is their confirmation of the early split of the San from the Yoruba and Bantu populations (~130,000 years ago), long before the out-of-Africa event, thus highlighting the antiquity of the San population.

Next steps

Together these two papers highlight both the power and the challenges of demographic inference based on population genetic analyses of whole-genome sequence data sets. Gronau *et al.* and Li *et al.* each illustrate powerful model-based approaches, and together have substantially increased our understanding of the historical relationships among some of the major human population groups. Yet a more nuanced view of human history will require analysis of larger numbers of individuals, representing more populations. It remains very difficult to disentangle population divergence, gene flow and admixture events; however, a more detailed understanding of these parameters would be very helpful for interpreting other problems in population genetics such as determining the extent of recent natural selection[2]. Given the difficulties of estimating historical parameters in very complex models, there is also an important place for alternative approaches that place less emphasis on formal parameter estimation. Such approaches may be helpful for detecting interesting signals in complex data sets[15]. In any event, we will soon have large numbers of whole-genome sequences from diverse populations, and with this the opportunity to decipher human history at an unprecedented resolution.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. 1000 Genomes Project Consortium. *Nature* **467**, 1061–1073 (2010).
2. Hernandez, R.D. *et al. Science* **331**, 920–924 (2011).
3. Gronau, I. *et al. Nat. Genet.* **43**, 1031–1034 (2011).
4. Li, H. & Durbin, R. *Nature* **475**, 493–496 (2011).
5. Pool, J.E. *et al. Genome Res.* **20**, 291–300 (2010).
6. Nordborg, M. in *Handbook of Statistical Genetics* (eds. Balding, D.J., Bishop, M. & Cannings, C.) 179–212 (Wiley, 2001).
7. Beerli, P. & Felsenstein, J. *Proc. Natl. Acad. Sci. USA* **98**, 4563–4568 (2001).
8. Fearnhead, P. & Donnelly, P. *Genetics* **159**, 1299–1318 (2001).
9. Gutenkunst, R.N. *et al. PLoS Genet.* **5**, e1000695 (2009).
10. Li, N. & Stephens, M. *Genetics* **165**, 2213–2233 (2003).
11. McVean, G. *et al. Genetics* **160**, 1231–1241 (2002).
12. McVicker, G. *et al. PLoS Genet.* **5**, e1000471 (2009).
13. Drummond, A.J. *et al. Mol. Biol. Evol.* **22**, 1185–1192 (2005).
14. Keinan, A. *et al. Nat. Genet.* **39**, 1251–1255 (2007).
15. Hellenthal, G. *et al. PLoS Genet.* **4**, e100078 (2008).

Germline *BAP1* mutations and tumor susceptibility

Alisa M Goldstein

Two new studies describe germline mutations in *BAP1* in putatively dissimilar cancer-related syndromes. The spectrum of neoplasms associated with these germline mutations suggest that BAP1 has an important tumor suppressor function in multiple tissues.

BAP1 is a tumor suppressor gene located on chromosome 3p21 in a region that shows loss or deletions in numerous cancers, including lung and breast cancers as well as uveal melanoma and mesothelioma. Two recent studies reported high frequencies of somatic mutations in *BAP1* in uveal melanoma[1] and mesothelioma[2]. Harbour *et al.*[1] found inactivating somatic mutations of *BAP1* in 47% (28/60) of uveal melanomas (**Table 1**), with a much higher frequency (27/34, 79%) in metastasizing uveal melanomas. In addition, one tumor had a germline frameshift mutation, suggesting that this variant was a susceptibility allele. Bott *et al.*[2] identified somatic inactivating *BAP1* mutations in 23% (12/53) of a discovery set of malignant pleural mesotheliomas, half of them from individuals who reported asbestos exposure, and in 18% (12/68) of an independent set of malignant pleural mesotheliomas. In this issue, Wiesner, Bastian, Speicher and colleagues[3] and Testa, Carbone and colleagues[4] provide further links between *BAP1*, uveal melanoma and mesothelioma by identifying germline *BAP1* mutations in two putatively distinct cancer-related syndromes characterized predominantly by melanocytic tumors or mesothelioma, respectively, along with uveal melanoma in both cases.

Two new cancer-related syndromes?

Wiesner *et al.*[3] identified co-segregating germline mutations in *BAP1* in two families (1 and 2) with multiple members with melanocytic tumors that ranged histopathologically from epithelioid nevi to atypical melanocytic proliferations with features that overlapped with cutaneous melanoma. Both families had one member with uveal melanoma, and family 2 had multiple members with cutaneous melanoma. Affected family members had many (5 to >50) of the clinically and histopathologically distinct melanocytic tumors, whereas there were few melanomas, suggesting that the risk of malignant potential in individual tumors was low. Examination of a subset of the familial melanocytic tumors showed that the majority of tumors showed biallelic inactivation of *BAP1* by various somatic alterations.

Testa *et al.*[4] discovered co-segregating germline mutations in *BAP1* in two families (L and W) with five or more members with mesothelioma. The families had modest levels of asbestos exposure from having lived in asbestos-containing houses but did not have occupational asbestos exposure. The families also had multiple members with various malignancies, including two members with uveal melanoma in family L. Somatic alterations in the familial mesothelioma tumors indicated biallelic inactivation of *BAP1*. Testa *et al.*[4] also sequenced *BAP1* in germline DNA from 26 sporadic mesothelioma cases and uncovered two inactivating frameshift mutations in individuals who were subsequently found to have had earlier diagnoses of uveal melanoma. Given the rarity of both uveal melanoma and mesothelioma in the United States, the authors concluded that it was extremely unlikely for these two malignancies to have occurred in the same individuals by chance.

BAP1 mutations and other cancers

Beside uveal melanoma, mesothelioma and the distinct melanocytic tumors, numerous additional tumors were observed in the families[3,4]. Specifically, several cancers, including cutaneous

Table 1 Tumors with somatic BAP1 mutations

		Tumors with somatic *BAP1* mutations		
Tumor type	Total tumors analyzed	*n*	%	Ref(s).
Melanocytic				
Common nevi	29	0	0.0	3
Spitz nevi	17	0	0.0	3
Atypical Spitz nevi	18	2[a]	11.1	3
Cutaneous melanoma	60	3	5.0	3
Uveal melanoma, total	93	41	44.1	3,1
Wiesner *et al.* study	33	13	39.4	3
Harbour *et al.* study	60	28	46.7	1
Low metastatic risk	26	1	3.8	1
High metastatic risk	31	25	80.6	1
Metastatic	3	2	66.7	1
Mesothelioma, total	139	28	20.1	4,2
Testa *et al.* study	18	4	22.2	4
Bott *et al.* study	121	24	19.8	2
Breast	251	1	0.6	5,6
Lung	322	2	0.4	5,6
Ovary	59	2	3.4	5,6
Pancreas	30	0	0.0	5,6

[a]These tumors had morphologic features similar to the melanocytic tumors observed in families 1 and 2 of ref. 3.

Alisa M. Goldstein is at the Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland, USA.
email: goldstea@mail.nih.gov