2003S

Reconstruction of ancestral human haplotypes using genetic and genealogical data. *J.M. Granka[1], R.E. Curtis[2], K. Noto[1], Y. Wang[1], J.K. Byrnes[1], M.J. Barber[1], N.M. Myres[2], C.A. Ball[1], K.G. Chahine[2].* 1) AncestryDNA, San Francisco, CA; 2) AncestryDNA, Provo, UT.

The genomes of individuals who lived long ago may persist in modern populations in the form of genomic segments broken down by recombination and inherited by their descendants. We develop a novel computational method to reconstruct the chromosomal haplotypes of human ancestors given genetic data from a sufficient number of their present-day descendants. After genealogical information is used to identify descendants of an ancestor (and therefore also of their partner), we use phased genome-wide single nucleotide polymorphism (SNP) data to find regions of the genome that are identical-by-descent (IBD) among them. We develop a novel stitching algorithm to reconstruct up to four chromosomal haplotypes of an ancestor and their partner given the descendant IBD segments and haplotypes. The method aims to remove spurious IBD segments caused by false inference, inaccurate genealogical information, or multiple common ancestors. Short regions of the genome with missing data can also be imputed given flanking reconstructed haplotypes. Lastly, given descendants of other individuals related to the ancestral couple, we show that it is sometimes possible to tease apart the personal identity of each of the reconstructed chromosomal haplotypes (i.e., which are the ancestor's, and which are their partner's). Through simulations, we calculate the amount of the genome that can theoretically be reconstructed given the number of generations back to the ancestor and the number of actual and sampled descendants. Given sufficient data, we can reliably reconstruct the haplotypes of in silico ancestors with high precision and recall; performance is sensitive to genealogical tree quality and accuracy of inferred IBD. We apply our method to phased genome-wide SNP data, obtained from the AncestryDNA customer database, from several hundred individuals descended from an 18th century couple. In genomic regions with many inferred IBD segments, we can reconstruct the haplotypes of the couple, in some cases assigning each haplotype to a specific member of the pair. In regions of the genome with fewer segments, we are less able to discover all haplotypes with certainty. Finally, we demonstrate that given these reconstructed haplotypes, we can infer a given ancestor's ancestry and select physical features. Our study highlights the feasibility of reconstructing the genomes of human ancestors and has immediate applications in population genetics, medical genetics, and genealogy research.

2004M

Identity by descent between humans, Denisovans, and Neandertals. *S. Hochreiter, G. Povysil.* Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria.

We analyze the sharing of very short identity by descent (IBD) segments between humans, Neandertals, and Denisovans to gain new insights into their demographic history. Short IBD segments convey information about events far back in time because the shorter IBD segments are, the older they are assumed to be. The identification of short IBD segments becomes possible through next generation sequencing (NGS), which offers high variant density and reports variants of all frequencies. Only recently HapFABIA has been proposed as the first method for detecting very short IBD segments in NGS data. HapFABIA utilizes rare variants to identify IBD segments with a low false discovery rate. We applied HapFABIA to the 1000 Genomes Project whole genome sequencing data to identify IBD segments which are shared within and between populations. Some IBD segments are shared with the reconstructed ancestral genome of humans and other primates. These segments are tagged by rare variants, consequently some rare variants have to be very old. Other IBD segments are also old since they are shared with Neandertals or Denisovans, which explains their shorter lengths. The Denisova genome most prominently matched IBD segments that are shared by Asians. Many of these segments were found exclusively in Asians and they are longer than segments shared between other continental populations and the Denisova genome. Therefore, we could confirm an introgression from Deniosvans into ancestors of Asians after their migration out of Africa. While Neandertal-matching IBD segments are most often shared by Asians, Europeans share more than other populations, too. Again, many of the Neandertal-matching IBD segments are found exclusively in Asians, whereas Neandertal-matching IBD segments that are shared by Europeans are often found in other populations, too. Neandertal-matching IBD segments that are shared by Asians or Europeans are longer than those observed in Africans. This hints at a gene flow from Neandertals into ancestors of Asians and Europeans after they left Africa. Interestingly, many Neandertal- or Denisova-matching IBD segments are predominantly observed in Africans - some of them even exclusively. IBD segments shared between Africans and Neandertals or Denisovans are strikingly short, therefore we assume that they are very old. This may indicate that these segments stem from ancestors of humans, Neandertals, and Denisovans and have survived in Africans.

2005S

Inferring demographic history from whole genome using Approximate Bayesian Computation. *F. Jay, F. Austerlitz.* Laboratoire Eco-Anthropologie et Ethnobiologie, Muséum National d'Histoire Naturelle, Paris, Paris, France.

Reconstructing past demography from neutral genetic data is essential as it both improves our knowledge about human history and provides an accurate neutral model against which selective hypotheses can be tested. Approximate Bayesian Computation (ABC) has proven to be useful for inferring demography from microsatellite or SNP data. This approach consists in simulating genetic data under a large range of complex demographic scenarios and realistic biological processes. Simulations are then compared to observed data using a set of informative summary statistics. Whole-genome data are expected to be extremely rich in information about past demography but, because simulations were, until recently, computationally too costly, ABC methods have not been thoroughly tested on such very long sequences. Dense polymorphism data contain extra information that is not available from unlinked site polymorphisms, and will, therefore, hopefully improve the reconstruction of demographic history. They allow computing specific statistics, such as the decay of linkage disequilibrium with distance, the distribution of length of haplotypes shared between two or more individuals, or the "allele frequency identity-by-state" as described by Theunert et al. (2012). The power given by some of these statistics to infer demographic parameters has been investigated. However, studies were done independently on single classes of statistics, and not always under the approximate Bayesian framework.

Here, we examine how combining these "dense data statistics", with "classical statistics" (e.g. pairwise differences, heterozygosity) in an ABC framework improves the inference of demographic history. Furthermore, we describe how sequencing errors that are usually more frequent in full sequences than SNP data impact the summary statistics and the ABC inference. To diminish these effects we propose to either (i) filter data drastically and prune summary statistics that are highly sensitive to errors, or (ii) model errors within our ABC and incorporate them into simulations. We benchmark these different approaches for simple demographic scenarios, and focus more specifically on population expansion events that happened in recent human history.

Theunert C et al (2012) Inferring the history of population size change from genome-wide SNP data. MBE, 29, 3653-3667.

2006M

Genetic Structure of North-Indian Punjabi Population Based on Autosomal Microsatellite Loci. *M. Kaur, B. Badaruddoza.* Deptt. of Human Genetics, Guru Nanak Dev University, Amritsar, India.

The population of Punjab, India, possesses an exclusive genetic profile, primarily due to the many migratory events in this region which caused an extensive range of genetic diversity. Hence, the present study is an attempt to find out the genetic similarity and phylogenetic position of north-west Punjabi population with respect to past history of admixture of foreign populations, especially, Caucasoid Populations. In this study, six microsatellite markers: THO1, TPOX, CSF1PO, vWA, D7S820 and FGA have been analyzed among 516 samples from five endogamous population groups, Jat Sikhs, Mazhbi Sikhs, Brahmins, Ramdasias and Muslims of north-west border districts of Punjab. The number of alleles ranged from 8 to 12 at six STR loci. The exact test probabilities for HWE suggested some significant departures in certain loci and population groups. In general, the average observed heterozygosity was lower than expected heterozygosity in six STR markers among the five population groups. The average sub-ethnic genomic differentiation (Fst) among five population groups of northwest Punjab was 0.0335. The CSF1PO showed highest sub-ethnic differentiation (0.0649), whereas, the lowest Fst has been observed for FGA locus (0.013). The percentage of genomic diversity attributable to different populations relative to the total genomic diversity (Gst) varied between 6.1% for CSF1PO locus and 0.9% for D7S820. When all the loci were jointly considered, 3.0% of the total genomic diversity was attributable to the five population groups. The maximum gene flow was observed in FGA (Nm=18.77%), which was followed by THO1 (15.92%). The lowest amount of gene flow was observed in CSF1PO (3.6%). However, in general with all loci the gene flow was observed to be 7.2% among these studied population groups. To understand the extent of sub-structuring among five northwest Punjabi population. Structure analysis was also performed with different values of K. The log probability values and the membership proportion of each group showed clear sub-structure among the population groups. Overall, five Punjabi speaking population groups of northwest Punjab are regionally well differentiated and exhibit strong genetic affinity based on their origin, settlement and their shared ethno-historic background.