Genome Biology

RESEARCH HIGHLIGHT

A new era of human population genetics

Alexander Platt* and John Novembre

Abstract

The 1000 Genomes Project Consortium has recently published an important early contribution to a new generation of systematic surveys of rare human genetic variation.

The history of population genetics has involved a continual interplay between observations of genetic data and prevailing theories of the dominant processes shaping variation. As new methods to survey genetic variation have been developed, the consensus understanding of the forces generating variation has changed multiple times and in dramatic ways. A new study from the 1000 Genomes Project Consortium [1], which is officially the Phase 1 publication, marks a watershed moment in population genetics. The field is entering an era in which rare variants are systematically characterized, and in which expectations from these data include a more detailed understanding of how purifying selection, migration and recent demography shape genetic variation.

A classic example of the power of new data in population genetics is how the observation of the clock-like rate of amino acid substitutions between species, together with early surveys of allozyme polymorphism dating back to the 1960s, helped fuel the development of the neutral theory by Motoo Kimura and colleagues [2]. Prior to these findings, most observable mutations were associated with morphological, physiological or disease traits. Incidences of rare mutant phenotypes of these kinds, in the presence of a dominant wild type, created an emphasis among one major school of evolutionary geneticists on the balance between mutation and natural selection as the determinants of the abundance of genetic polymorphisms [3]. Once large numbers of molecular polymorphisms were revealed, it became hard to sustain the idea that all molecular variation was due to mutation-selection balance, as this would imply high rates of mutation and a substantial genetic load.

*Correspondence: alex@alexanderplatt.org
Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA 90095, USA

The neutral theory was developed to help explain these new observations, and posits that sites are either selectively important, and therefore invariant, or invisible to selection and therefore free of constraint, with levels of variation set by mutation-drift balance. Since its introduction, the strict neutral theory has given way to nearly neutral models that include weakly selected sites, and to a greater emphasis on recent positive selection. Indeed, the development of new theories and perspectives as a result of breakthroughs in data generation is a repeating feature in the history of population genetics.

The newest era of data generation in population genetics, in part ushered in by the 1000 Genomes Project, is one of comprehensive characterization of rare variants within a species. The 1000 Genomes Project has produced genomes for 1,092 individuals from a diverse set of populations, with an average coverage of 80× in the exomes and 4 to 5× coverage for the rest of the genome. This has empowered the project to detect variants with frequencies below 0.5% and to characterize their geographic distribution at a regional scale. Compared to their more common brethren, rare alleles have two distinctive properties, each of which opens up new types of analyses. Both of these avenues have been explored by the 1000 Genomes Project Consortium.

Rare alleles behave like neutral alleles

The 1000 Genomes Project has taken advantage of how most rare variants are only weakly affected by selection. In simple theoretical models of evolution, the expected change in frequency due to selection for an allele with frequency p and additive selection coefficient of s is $ps(1 - p)/(1 + ps)$. When allele frequencies are small (p is low) and selection is weak (s is small), this quantity falls well below the expected magnitude of allele change due to genetic drift. This phenomenon is further enhanced when the fitness effect of the allele is recessive. The result is that both neutral rare alleles and nearly neutral rare alleles behave similarly and the impact of natural selection is only manifest once alleles become more common. Thus, differences in the distributions of rare and common variants can be used to identify the degree to which selection is acting on mutations at different kinds of sites.

In its Phase 1 publication [1], the Consortium exploits this idea by comparing the ratio of non-synonymous to

synonymous polymorphisms in different frequency ranges to estimate the size of the slightly deleterious class of variants compared to neutral variants. They observe that there are approximately as many rare non-synonymous polymorphisms as rare synonymous polymorphisms, but that the ratio falls to nearly 1:2 among more common alleles. This suggests that up to one-half of variable non-synonymous variants are slightly deleterious, and that this fraction varies among genes in different types of pathways. The project used the difference between rare and common relative abundances in different functional categories of mutations to rank the categories by degree of selective constraint. At the extremes, less than 10% of mutations creating novel stop codons or disrupting splice sites are able to rise in frequency above 0.5%, compared to the 65% of synonymous substitutions that have done so. However, the Consortium also finds that the cross-species conservation of a site is a better predictor of whether or not a mutation can become common than its functional category is. This indicates that there is tremendous within-category variability in the degree of selective constraint a mutation is under. The only category for which this is not true is mutations that create stop codons. These always appear to be under greater constraint than other categories of mutations regardless of evolutionary conservation. Overall, these results support Ohta's nearly neutral theory [2], in which the contribution of weakly selected variants to overall genetic diversity cannot be neglected.

Rare alleles are young alleles

Since all mutations start as single-copy variants and then often increase in frequency or are lost, many rare variants have only recently arisen. Common variants, unless they are under moderately strong selection, can only have attained their frequency through the passing of time. The spread across populations of these older, more common, alleles reflects the combined demographic properties of an extended time period, whereas the distribution of rare alleles reflects only more recent historical events. The 1000 Genome Project's Phase 1 publication confirms that previous studies on common variant discovery were successful in identifying the vast majority (94%) of variants that occur at frequencies of 5% or higher; however, nearly half the variants that occur at a frequency between 0.5% and 5% were novel, and almost 90% of the variants with a frequency of less than 0.5% were previously unknown.

In European, Asian and American populations, the impact of population bottlenecks coming out of Africa can be seen in a loss of derived alleles of frequencies up to 40%. Clear evidence of subsequent explosive demographic growth, however, is only apparent in the strong excess of very rare variants. Rare and common variants also differ greatly in their geographic spread. Virtually all variants with frequencies greater than 1% are present on all continents and, for the most part, in all sampled populations; in contrast, more than half of variants occurring at a 0.5% frequency are restricted to a single population.

The rarest, and therefore youngest, demography-informative alleles are doubletons, alleles found exactly twice in the sample. For the majority of such variants, both copies were sampled from the same population, suggesting that they arose more recently than the formation of the distinct populations, even for populations occurring on the same continent. Where the two copies of a doubleton are found in different populations, this can be considered evidence of recent gene flow between the populations. One specific example of recent gene flow detected by the study was the finding that allele sharing between the Spanish and admixed American samples was greater than the amount of sharing between the Spanish sample and the remaining European samples; a likely explanation for this observation lies in the migrations that accompanied the historical colonization of the New World. The ability to study such recent gene flow patterns will be a powerful tool for studies of many other areas of human history.

The future of the study of rare alleles

The observations made in the 1000 Genomes Project Phase 1 publication [1] are among the most detailed of a bevy of recent studies describing the abundance of rare variants in sequencing data from very large samples [4-7]. As these studies grow in number, sample size and sampling locations, there will be an increasing push to develop new analytical tools and models. So far, studies, including that of the 1000 Genomes Project, have been predominantly descriptive. We are still largely in the phase of finding out what interesting patterns these new data show, a necessary precursor to developing improved analysis methods that will be able to exploit and understand the data in full detail. Where current methods can effectively assume that all variants are independently assorting, these new data invite the creation of methods to explain patterns of linkage disequilibrium around rare variants. Studying large samples of rare variants will also require methods with explicit spatial structure**,** as the genetic background on which a rare variant is found is highly dependent on the locale where it arose, and the migrational patterns of its carriers. Furthermore, studies of this type will likely be conducted, as the 1000 Genomes Project has been with at least a significant portion of low-coverage sequencing followed by imputation. This process is clearly an important technique for discovering and characterizing rare variants in a cost-effective manner, although exactly what artifacts are introduced is not fully understood. In addition, as sample sizes increase, some of the basic assumptions of coalescent theory begin to

break down because of the potential for simultaneous coalescent events; theory in this area will need to be further developed, and adapted to empower data inference.

A key development in the field of human population genetics over the past decade has been the tremendous effort made to ensure that large-scale studies are not only publicly accessible but also easily integrated with previous and future projects. The 1000 Genomes Project is no exception to this trend; beyond the direct value of the results described in its Phase 1 publication is the accompanying data resource, which will be of immense use to future studies investigating yet more populations and rarer variants.

Competing interests

The authors declare that they have no competing interests.

Published: 26 December 2012

References

1. The 1000 Genomes Project Consortium: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491:**56-65.
2. Ohta T, Gillespie, JH: **Development of neutral and nearly neutral theories.** *Theor Popul Biol* 1996, **49:**128-142.
3. Crow JF: **Population genetics history: a personal view.** *Annu Rev Genet* 1987, **21:**1-22.
4. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR, Wheeler DA, Sabo A, Lusk C, Weiss KG, Akbar H, Cree A, Hawes AC, Newsham I, Varghese RT, Villasana D, Gross S, Joshi V, Santibanez J, Morgan M, Chang K, Iv WH, Templeton AR, Boerwinkle E, Gibbs R, Sing CF: **Deep resequencing reveals excess rare recent variants consistent with explosive population growth.** *Nat Commun* 2010, **1:**131.
5. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, Warren L, Aponte J, Zawistowski M, Liu X, Zhang H, Zhang Y, Li J, Li Y, Li L, Woollard P, Topp S, Hall MD, Nangle K, Wang J, Abecasis G, Cardon LR, Zöllner S, Whittaker JC, Chissoe SL, Novembre J, *et al.*: **An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people.** *Science* 2012, **337:**100-104.
6. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM; Broad GO; Seattle GO; NHLBI Exome Sequencing Project: **Evolution and functional impact of rare coding variation from deep sequencing of human exomes.** *Science* 2012, **337:**64-69.
7. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, NhlbI Exome Sequencing Project, Akey JM: **Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants.** *Nature* 2012. doi: 10.1038/nature11690.

doi:10.1186/gb-2012-13-12-182
Cite this article as: Platt A, Novembre J: **A new era of human population genetics.** *Genome Biology* 2012, **13:**182.