Genome Medicine

COMMENTARY

The 1000 Genomes Project: new opportunities for research and social challenges

Marc Via*, Christopher Gignoux and Esteban González Burchard

Abstract

The 1000 Genomes Project, an international collaboration, is sequencing the whole genome of approximately 2,000 individuals from different worldwide populations. The central goal of this project is to describe most of the genetic variation that occurs at a population frequency greater than 1%. The results of this project will allow scientists to identify genetic variation at an unprecedented degree of resolution and will also help improve the imputation methods for determining unobserved genetic variants that are not represented on current genotyping arrays. By identifying novel or rare functional genetic variants, researchers will be able to pinpoint disease-causing genes in genomic regions initially identified by association studies. This level of detailed sequence information will also improve our knowledge of the evolutionary processes and the genomic patterns that have shaped the human species as we know it today. The new data will also lay the foundation for future clinical applications, such as prediction of disease susceptibility and drug response. However, the forthcoming availability of whole genome sequences at affordable prices will raise ethical concerns and pose potential threats to individual privacy. Nevertheless, we believe that these potential risks are outweighed by the benefits in terms of diagnosis and research, so long as rigorous safeguards are kept in place through legislation that prevents discrimination on the basis of the results of genetic testing.

Introduction

Now that the Human Genome and the HapMap Projects have been completed, the international scientific community is turning its attention to the 1000 Genomes Project [1], an international collaboration between China, Germany, the UK, and the USA. The initial goal was to discover most of the genetic variation that occurs at a population frequency greater than 1% by deep sequencing at least 1,000 individuals from different worldwide populations using next-generation platforms and technologies. The genomes of approximately 2,000 individuals, from at least 20 different populations representing Africa, Europe, East Asia, and the Americas, are being collected and sequenced. The populations included will each have approximately 60 to 100 samples sequenced. For some populations, trios (both biological parents and an adult child) have been collected. Many of the samples, including some from the children, are going to be densely genotyped using genome-wide arrays. The goal of this study design is to reconstruct the parental chromosomal phase using the information provided by the child.

Discussion

Promising results have already arisen from the Pilot 1 analysis, which has been referred to as the '1000 Genomes Low Coverage Pilot'. This analysis consists of 180 individuals from the four original HapMap populations sequenced at 2X to 4X coverage, meaning that an average number of two to four sequences are generated for every genomic position. In this initial phase, more than 9 million new single-nucleotide polymorphisms (SNPs), many novel indels (insertions/deletions), and some large structural variants have been identified.

The results that arise from the completion of this project will lead to a great leap in our knowledge of human genetic variation. At first, the results will allow scientists to identify population-specific genetic variation at an unprecedented degree of resolution. Of the 9 million novel SNPs identified so far from Pilot 1 analysis alone, approximately 8 million are seen in only one HapMap population. Most common variants have already been identified; the novel variants are disproportionately rare and thus more likely to be observed in only one of the studied populations. The identification of these variants will help the development of new population-specific

*Correspondence: marc.viagarcia@ucsf.edu
ELSI/Samples Committee, 1000 Genomes Project, and Institute for Human Genetics, University of California at San Francisco, Box 2911, San Francisco, CA 94143, USA

genotyping arrays. This will maximize genome coverage while minimizing the ascertainment bias that affects currently available arrays, especially for non-European populations. This unbiased survey of polymorphism in diverse groups will also help improve the imputation methods for genetic variants that are not represented on current arrays.

In addition to improving the resolution of population genetics, the clinical implications are endless. For example, genome-wide association studies (GWASs) are now routinely used to identify genomic regions associated with common human diseases. However, these studies rarely identify the precise causative genes or sequence variants. The reason for this is that the human genome contains regions of strong linkage disequilibrium, and a disease-associated locus can encompass several genes and multiple tightly associated polymorphisms. In addition, current arrays emphasize so-called 'tag SNPs', or SNPs that are highly correlated with their local linkage disequilibrium structure to provide more even coverage across the genome. These SNPs can usually be found in introns or intergenic regions, not coding regions or known regulatory elements that are likely to be functional. This makes it difficult to pinpoint causal variants by association mapping using genotyping arrays alone. Deep sequencing studies will identify novel or rare functional variants, thereby allowing scientists to find all potential disease-causing variants and genes. However, the strong associations among genetic variants in a given genomic region will require experimental studies to be performed to determine which of the associated genetic variants are actually functionally causal. Recent work has shown on a small level that resequencing of candidate regions from GWASs can often uncover rarer variants with higher effects and more direct functional consequences in common diseases [2].

The unprecedented level of sequence information that will arise from the 1000 Genomes Project will also improve our knowledge of genomic configurations that were shaped by evolutionary processes. For example, analyses of the distribution of SNP density along chromosomes will inform us about chromosomal regions that are more susceptible to selective pressures or differential patterns as a result of the expansion of humans throughout different continents. Recent findings from exonic resequencing have shown that patterns of population-specific rare, deleterious mutations (such as those that cause Mendelian recessive diseases) in coding regions can be largely explained by historical processes affecting specific populations [3]. However, understanding how these processes occur, which genes are affected, and which other populations in the future could have a higher prevalence of diseases caused by rare variants requires a large-scale resource for validation of population genetics methods, a resource such as the 1000 Genomes Project. The samples that are included in the 1000 Genomes Project do not have identifying information, phenotypes, or clinical data available. The project is providing a resource about human genetic variation that will be used in many studies of particular phenotypes, such as complex diseases or drug response.

The availability of full genome sequences from worldwide samples will directly improve the accuracy of direct-to-consumer genetic ancestry tests. Medical applications, such as determining drug efficacy and/or toxicity and prediction of disease and prognosis, will need further deep sequencing in clinical projects before public benefits from these new technologies are developed. The attention from the media will probably increase the public's expectations of the 1000 Genomes Project and its potential applications. However, scientists should take advantage of the increased public awareness to highlight the importance of genetic research and to encourage the participation of all communities in future research. Larger follow-up studies will be needed to achieve the required statistical power to establish conclusions in datasets that will include billions of variables integrating genomic data with future transcriptomic, proteomic, and epigenetic information.

The recent improvements in sequencing technology, which allow the sequencing of samples for the 1000 Genomes Project, presage the forthcoming availability of whole genome sequences at affordable prices. The sequencing of individual genomes raises concerns about potential threats to privacy and other ethical issues. Although the USA recently passed the Genetic Information Nondiscrimination Act (GINA), prohibiting genetic discrimination in employment and health insurance [4], the protection of individual rights varies from country to country. For example, the European Commission strongly recommends the prevention of discrimination as a result of genetic testing in insurance and employment [5]. However, the EU does not have uniform legislation regarding whether it is legal for insurance companies or employers to access genetic testing results or other medical records. Rather, the decision on whether to legislate the use of genetic test results has been left to the discretion of each individual country.

Conclusions

The information that the 1000 Genomes Project and the next generation of deep sequencing platforms will provide is unprecedented and will have important implications for genetics and health as we move closer and closer to the era of the ubiquitous personal genome as part of our medical record.

Despite our best intentions, history has demonstrated that it will be very difficult to get the genie back into the

bottle once it is opened. This is has been true for all scientific and medical advances. Although there are potential social costs associated with linking informative genetic data to individuals or populations, we believe that these potential costs are outweighed by the benefits in terms of diagnosis and research. Scientists should continue to use the genetic diversity that exists within humans as starting points for further research. We cannot stall scientific progress for fear of potential misuses of knowledge. Rather, we should keep rigorous safeguards in place to ensure that scientific advancements proceed and serve to benefit all humanity. Finally, it is important to establish a fluent communication with the general public and engage participation in genetic research.

Abbreviations

GINA, Genetic Information Nondiscrimination Act; GWAS, genome-wide association study; SNP, single-nucleotide polymorphism.

Acknowledgements

We acknowledge the support of National Institutes of Health (HL078885, HL088133, U19 AI077439, and ES015794), Flight Attendant Medical Research Institute (FAMRI), and the RWJ Amos Medical Faculty Development Award to EGB, Beatriu de Pinos Postdoctoral Grant (2006 BP-A 10144) to MV, the UCSF Chancellor's Fellowship to CG, the Sandler Center for Basic Research in Asthma and the Sandler Family Supporting Foundation. We also thank Lisa Brooks and Jean McEwen (NIH/NHGRI and the 1000 Genomes Project) for their useful comments.

Authors' contributions

MV, CG and EGB contributed to the design and writing of this commentary.

Competing interests

The authors declare that they have no competing interests.

Published: 21 January 2010

References

1. The 1000 Genomes Project [http://www.1000genomes.org]
2. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA: **Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.** *Science* 2009, **324:**387-389.
3. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD: **Proportionally more deleterious genetic variation in European than in African populations.** *Nature* 2008, **451:**994-997.
4. Genetic Information Nondiscrimination Act, Public Law 110-233, 122 Stat. 881.
5. European Commission: *EUR 21120-25 Recommendations on the Ethical, Legal and Social Implications of Genetic Testing.* Luxembourg: Office for Official Publications of the European Communities; 2004.

doi:10.1186/gm124
Cite this article as: Via M, *et al*.: The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Medicine* 2010, **2**:3.