Journal of Next Generation Sequencing & Applications

Gao, Next Generat Sequenc & Applic 2014, 1:1
DOI: 10.4172/2469-9853.1000110

Commentry Open Access

The Feasibility of WGS Association Studies for Complex Traits

Xiaoyi Gao*

Department of Ophthalmology and Visual Sciences, University of Illinois at Chicago, Chicago

***Corresponding author:** Xiaoyi Gao, Department of Ophthalmology and Visual Sciences, University of Illinois at Chicago, USA, Tel: +312-996-5825; Fax: +312-996-0759; E-mail: rgao@uic.edu

Rec Date: September 24, 2014, **Acc date:** September 26, 2014, **Pub date:** September 29, 2014

Introduction

Whole genome sequencing (WGS) reveals every base pair of the human genome and offers the ultimate information resolution to researchers. In January 2014, Illumina announced that the $1,000 genome is available. It was a significant achievement in high-throughput sequencing technology since it took more than 10 years to break this long-sought threshold after the completion of the first sequenced human genome, which cost nearly $3 billion. However, the $1,000 price tag is a marketing figure that does not account for analysis or data storage, both of which tend to be more expensive than the consumables. Moreover, it requires that sequencing centers invest in a $10 million system in order to achieve the $1,000 per genome cost. Although the cost of sequencing a single genome is quite reasonable, there are many other costs to consider for WGS. Is WGS for complex trait association studies in large cohorts finally feasible?

Numerous studies have shown that a large sample size is the key element needed in order for statistical power to detect genotype-phenotype associations. This has been clearly demonstrated in the widespread use of meta-analyses, which can have several hundreds of thousands of samples nowadays. The basis of a meta-analysis is each individual genome-wide association study (GWAS). For illustration purposes, we can assume that a large GWAS uses 5000 subjects, although many GWAS can have a daunting sample size of a quarter of a million subject [**1**]. With 5000 subjects at $1,000 per genome, it will take $5 million to sequence every individual. Under current funding situations, the cost of sequencing alone looks prohibitive for any but the wealthiest centers, not to mention the expensive, labor-intensive, and time-consuming downstream data analysis. Even if there are sufficient funds for sequencing, the generation of about 120 Gb of data at a depth of 30x per human genome can easily fill up 150 4 TB servers. This large storage requirement is mainly due to redundancy in the data since WGS typically needs at least 30x depth, i.e. each base in the genome is covered by 30 short sequence reads on average for accurate variant calling according to current sequencing technology. Furthermore, even more space is needed in order to process the sequences, e.g. quality control, sequence alignment, and variant calling. The large amount of data can also create network traffic issues; even if some intermediate files are compressed. These requirements stress all but the largest and best-funded research centers. The ideal genome sequencing technology would be one that reads all nucleotides once and provides base calls with high accuracy, like DNA replication in cells, while remaining inexpensive. However, such technology does not yet exist.

In order to use large sample sizes for sequencing studies, two compromises have been made by researchers to reduce the cost and computational burden. Low-depth sequencing and whole-exome sequencing (WES) are used in order to reduce either sequencing depth or genomic coverage. Instead of using 30x depth, the CHARGE (Cohorts for Heart and Aging Research in Genetic Epidemiology) consortium (web.chargeconsortium.com) used 6x depth to sequence almost 1000 participants in a study of high-density lipoprotein cholesterol [**2**]. The UK10K project (www.uk10k.org) sequenced 4000 individuals to 6x depth using WGS to study cardiometabolic and health related traits. The 1000 Genomes Project WGS data also were analyzed at low-depth, i.e. 2-6 [**3**]. The weakness of this approach is that rare variants cannot be accurately measured. This is may be a crucial flaw, since it has been speculated that a substantial part of the missing heritability may be explained by these rare variants. Many other sequencing studies used WES to focus on the coding regions of the genome, i.e. the exons. Despite the reduced cost and analysis time, WES does not cover regulatory and non-protein coding regions of the human genome, regions that WGS has shown to be important. One strategy to overcome this shortfall is to supplement the WES with targeted sequencing in regulatory regions.

GWAS arrays can be complemented with targeted sequencing. This approach essentially interrogates target regions using high-depth sequencing and off-target regions with low-depth or low-cost approaches and can be very powerful for identifying rare variants associated with complex traits. For example, by following previously discovered GWAS hits with targeted sequencing, rare variants in the C3 gene associated with age-related macular degeneration were discovered [**4**]. Another way to complement current GWAS arrays is to design specialized custom arrays based on sequences. For example, the Illumina HumanExome Bead Chip delivers supreme coverage of 250,000 high-value common and rare exonic variants selected from 12,000 individual exome and whole genome sequences.

Another hybrid approach is to reduce the number of individuals that need to be sequenced. For example, individuals at the tails of the phenotype distribution may be sequenced for use in variant discovery with the assumption that these individuals are more likely to carry rare variants with large effects. This increases the frequency of rare variants in the study sample and the statistical power to detect them. Once the variants are identified for a cohort, they can be included in GWAS or exome chips as custom content or can be used for designing custom arrays, which are a fraction of the cost of sequencing.

Sequencing technology has become faster, cheaper, and more accurate over the past several years. However, even at a price tag of $1,000 per genome, WGS association studies for complex traits are still very expensive. In the future, inexpensive technology options may be available for researchers in which the genome will be read with minimum redundancy, or even only once, without loss of accuracy of sequence calls. Until then, researchers have to take approaches to reduce sequencing depth, or sequencing coverage, or take hybrid approaches that combine the strength of both GWAS arrays and sequencing, or a combination of high-depth and low-depth sequencing, in order to be cost effective and practical.

Acknowledgements

Research reported in this publication was supported in part by the National Eye Institute of the National Institutes of Health under Award Number R01EY022651 (to XG). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Berndt SI, Gustafsson S, Magi R, Ganna A, Wheeler E, et al. (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat Genet 45: 501-512.
2. Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, et al. (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet 45: 899-901.
3. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
4. Zhan X, Larson DE, Wang C, Koboldt DC, Sergeev YV, et al. (2013) Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. Nat Genet 45: 1375-1379.