European Journal of Human Genetics (2017) 25, 395–396

www.nature.com/ejhg

LETTER

From exomes to genomes: challenges and solutions in population-based genetic association studies

European Journal of Human Genetics (2017) **25,** 395–396; doi:10.1038/ejhg.2016.206; published online 25 January 2017

Since the completion of the Human Genome Project and the technological revolution that it launched, the challenges of producing high-quality whole-genome sequence (WGS) data have largely been met. Today, many large-scale international studies are beginning to sequence the genomes of many thousands of individuals in order to understand the genetic etiology of common and rare, complex and Mendelian human traits and diseases.[1]

Recently, several smaller-sized projects, based on both low-coverage whole-genome and high-coverage whole-exome sequence (WES) data, have developed strategies to overcome many of the technical challenges (eg data compression, sequence alignment and genotype calling) associated with such large-scale projects.[2,3] Nevertheless, as large-scale WGS data are generated with the goal of discovering genetic associations that inform disease treatment and prevention, significant scientific and computational challenges remain. As projects such as the Precision Medicine Initiative are launched,[1,4] experiences from the era of large-scale WES can inform the design and analysis of large-scale WGS data.

For all studies an important consideration is the statistical power to detect associations, which is a function of the genetic architecture (ie, the allele frequencies and effect sizes). Because the vast majority of variants within the genome are rare with allele frequencies $<1\%$,[2,3] a comprehensive search for genetic associations must account for rare variants. As the results from recent studies attest, many rare-variant associations display moderate to modest effect sizes (odds ratios <1.4 for binary traits and effect sizes <0.5 trait standard deviations for quantitative traits).[5–8]

If these results hold true for rare variants generally, the power to detect rare-variant associations will be severely limited compared with common-variant association studies and sufficiently powered studies may require even larger sample sizes than genome-wide association studies of common variants. Future studies focused on rare-variant associations should consider the successes that innovative statistical techniques for rare-variant association testing as well as efficient study designs have made on addressing the issue of limited statistical power.

One study design that can be used to increase power to detect associations for quantitative traits, compared with random ascertainment, is the extreme trait sampling design.[9] Briefly, an extreme trait design considers the entire distribution of a quantitative trait and selects samples from the tails or 'extremes' of the distribution. This technique can be generalized to case–control studies as well by sampling from the extremes for various risk factors, for example, early onset cases and older, high-risk controls as this may enhance the power to detect associations. The power of an extreme trait design is driven by the size of the underlying population from which the samples are selected.[10,11] There are multiple examples of new rare-variant associations being discovered from extreme samples drawn from large, population-based studies.[12–14] Though extreme trait sampling represents a powerful approach, they may not be suitable for every type of study. First, conclusions from extreme trait designs may be difficult to generalize due to differences in genetic architecture at the extremes of a quantitative trait. For instance, the stature of individuals in the extremes of height is typically driven by very few large effect variants, whereas for individuals in the 'middle' of the distribution, height is typically determined by hundreds of loci of small effects.[15] Second, unless specialized analytical methods are used that specifically acknowledge the sampling design, the analysis of secondary traits from an extreme trait design can lead to biased- or false-positive findings.[16]

In addition to sample selection, large-scale sequencing projects need to weigh the trade-offs associated with sequencing depth and sample size. High-quality genotypes may be obtained from low-coverage sequencing (defined here as $<10\times$) of whole genomes by using haplotype aware genotype callers.[3] However, this is not possible with WES data as it is difficult to reconstruct accurate haplotypes based on exonic variants alone. Though previous studies have shown that lower read depth may increase power by increasing the available sample size,[17] currently WGS data are typically generated with $30\times$ coverage as this is the standard protocol on the ubiquitously used Illumina HiSeq instrument (San Diego, CA, USA). For studies investigating structural variants and very rare or private variants, deep sequencing of at least $30\times$ is the way to guarantee high-quality variant calls.

An attractive alternative approach to direct sequencing is to impute sequence data into a set of samples with genotyping array data. The data from the 1000 Genomes Project have become a popular 'reference panel' for this type of genotype imputation. Newer sequence data sets, such as from the UK10K project, have augmented the 1000 Genomes data to provide reference panels capable of accurately imputing variants of $<1\%$ allele frequency.[18] The Haplotype Reference Consortium has gathered WGS data on 64 976 individuals and claims accurate imputation down to 0.1% allele frequencies. With such rich, publically available resources, genotype imputation provides a cost-effective strategy for investigating low-frequency and rare-variant associations.

In addition to study design, statistical techniques can be leveraged to test for aggregate rare-variant associations based on the premise that multiple rare variants within a gene or region contribute to an association. There are many points to consider when conducting rare-variant tests and many models have been developed accordingly. The choice of a minor allele frequency cutoff for including variants in these tests has been debated since the first methods were proposed.[19] Typically a 1% (or 0.5%) cutoff is enforced, though methods such as Variable Threshold[20] have been developed so that this cutoff is not defined arbitrarily, and all possible cutoffs are considered. A related point is whether variants should be weighted according to minor allele frequency,[21] or a prediction score for whether the variant is likely to be deleterious[22] or effect protein structure.[23] Methods exist to test for many different genetic models; there are fixed effect models such as the combined multivariate collapsing (CMC) approach that tests a

dominance model,[19] and the GRANVIL approach (gene- or region-based analysis of variants of intermediate and low frequency) that tests an additive model;[24] and random effect tests such as the Sequence Kernel Association Test (SKAT) that tests for heterogeneity of effect across variants (this is also referred to as a 'variance component test').[25] There are adaptive tests that seek to estimate some of these parameters while simultaneously testing for association (eg, KBAC, VW-TOW),[26,27] as well as omnibus methods that test both fixed effect and variance component models simultaneously (eg, SKAT-O and MiST).[28,29] Lee *et al*[30] present a comprehensive review of the statistical issues related to rare-variant association testing.

Despite significant progress in the statistical literature on this topic, only a few empirical examples have emerged that demonstrate multiple rare variants within the same gene contributing to an association. Although there are examples of fixed effect tests (eg, CMC, GRANVIL) finding rare-variant associations that were undetected with random effect tests (eg, SKAT),[5,7] we are not aware of single example where a random effect test has found a rare-variant association that was not also found by a fixed effect test.

As studies transition from WES to WGS, considerations on the proper unit of analysis is important. Although the exome offers a natural unit of analysis (ie, a gene) for aggregate rare-variant association methods, it is unclear how best to aggregate association signals outside of coding regions and whether the genetic effects in enhancers, promoters and other elements related to gene regulation will be detectable by the same aggregate methods that are used for exomes. For WGS studies to uncover aggregate signals outside of coding regions, a better understanding of the genome outside of the coding region will be crucial. However, aggregate rare-variant testing is really a means for gaining statistical power to detect associations. With sample sizes in the hundreds of thousands, future studies may be well-powered to detect individual rare-variant associations of modest effect, rendering moot the need to aggregate signal across genomic regions.

As the data deluge from WGS technologies continues, new software tools capable of handling massive data volumes, high dimensions, and sophisticated, statistical and computational analyses will be needed. In the past, researchers developed robust, accessible and user-friendly software tools for previous generations of genetic studies (eg, the GWAS- and WES-eras).[31,32] A crucial component of success moving forward will be the development of new computational tools and paradigms (eg, cloud-based computing) that scale with the size and complexity of WGS data from large-scale studies.

Finally, and perhaps most importantly, data sharing among researchers will be critical for future WGS studies. In order to obtain the sample sizes needed for well-powered analyses, the sharing of data and/or summary statistics for meta-analysis are both critical. The scientific networks that permit data sharing also lead to increased sharing of expertise, a priceless commodity in the field of human genetics where the goals of biologists, statisticians, computer scientists and clinical practitioners align.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We wish to thank the reviewers for their thoughtful comments.

Paul L Auer[1] and Suzanne M Leal*[,2]

[1]*Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA;*

[2]*Department of Molecular and Human Genetics, Center for Statistical Genetics, Baylor College of Medicine, Houston, TX, USA*

E-mail: sleal@bcm.edu

1 Collins FS, Varmus H: A new initiative on precision medicine. *N Engl J Med* 2015; **372**: 793–795.
2 Consortium UK, Walter K, Min JL *et al*: The UK10K project identifies rare variants in health and disease. *Nature* 2015; **526**: 82–90.
3 Genomes Project C, Auton A, Brooks LD *et al*: A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.
4 Terry SF: Obama's precision medicine initiative. *Genet Test Mol Biomarkers* 2015; **19**: 113–114.
5 Auer PL, Teumer A, Schick U *et al*: Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. *Nat Genet* 2014; **46**: 629–634.
6 Peloso GM, Auer PL, Bis JC *et al*: Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56 000 whites and blacks. *Am J Hum Genet* 2014; **94**: 223–232.
7 TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, Crosby J *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med* 2014; 371: 22–31.
8 Iotchkova V, Huang J, Morris JA *et al*: Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nat Genet* 2016; **48**: 1303–1312.
9 Carey G, Williamson J: Linkage analysis of quantitative traits: increased power by using selected samples. *Am J Hum Genet* 1991; **49**: 786–796.
10 Auer PL, Reiner AP, Wang G *et al*: Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: lessons learned from the NHLBI Exome Sequencing Project. *Am J Hum Genet* 2016; **99**: 791–801.
11 Kryukov GV, Shpunt A, Stamatoyannopoulos JA *et al*: Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci USA* 2009; **106**: 3871–3876.
12 Do R, Stitziel NO, Won HH *et al*: Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* 2014; **518**: 102–106.
13 Emond MJ, Louie T, Emerson J *et al*: Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *Nat Genet* 2012; **44**: 886–889.
14 Lange LA, Hu Y, Zhang H *et al*: Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am J Hum Genet* 2014; **94**: 233–245.
15 Wood AR, Esko T, Yang J *et al*: Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014; **46**: 1173–1186.
16 Lin DY, Zeng D: Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol* 2009; **33**: 256–265.
17 Li Y, Sidore C, Kang HM *et al*: Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res* 2011; **21**: 940–951.
18 Huang J, Howie B, McCarthy S *et al*: Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* 2015; **6**: 8111.
19 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008; **83**: 311–321.
20 Price AL, Kryukov GV, de Bakker PI *et al*: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010; **86**: 832–838.
21 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; **5**: e1000384.
22 Kircher M, Witten DM, Jain P *et al*: A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; **46**: 310–315.
23 Adzhubei IA, Schmidt S, Peshkin L *et al*: A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
24 Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010; **34**: 188–193.
25 Wu MC, Lee S, Cai T *et al*: Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011; **89**: 82–93.
26 Liu DJ, Leal SM: A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 2010; **6**: e1001156.
27 Sha Q, Wang X, Wang X *et al*: Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genet Epidemiol* 2012; **36**: 561–571.
28 Lee S, Emond MJ, Bamshad MJ *et al*: Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012; **91**: 224–237.
29 Sun J, Zheng Y, Hsu L: A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol* 2013; **37**: 334–344.
30 Lee S, Abecasis GR, Boehnke M *et al*: Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 2014; **95**: 5–23.
31 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
32 Wang GT, Peng B, Leal SM: Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. *Am J Hum Genet* 2014; **94**: 770–783.