Genome Medicine

RESEARCH HIGHLIGHT

Testing for rare variant associations in complex diseases

Jennifer Asimit* and Eleftheria Zeggini

Abstract

The study of rare variants holds the promise of accounting for some of the missing heritability in complex traits. Next-generation sequencing technologies enable probing of variation across the full spectrum of allele frequencies. Multiple methods for the analysis of rare variants have been proposed and, recently, Ionita-Laza *et al*. have presented an approach with the theoretical capacity to detect risk and protective variants. The identification of rare risk variants could have major implications in understanding complex disease etiopathogenesis.

The importance of rare variants in complex disease

Genome-wide association scans have discovered common-frequency variants that play a role in complex disease susceptibility. However, these variants account for only a small fraction of the genetic component of disease. Rare variants are hypothesized to have larger effect sizes and may help fill in some of this heritability gap. In recent studies, rare variants have been shown to play an important role in complex disease etiology [1,2]. Assessing the role of rare variants in complex diseases is becoming increasingly feasible as next-generation sequencing technologies allow the efficient genome-wide sequencing of many individuals.

Single-variant association tests have been successful in detecting disease-associated common variants, but such tests have been shown to be sensitive to both allele frequency and effect size, so that they are underpowered in analyzing rare variants [3,4]. More powerful approaches to rare variant analyses are locus-based tests that collectively test all of the rare variants within a region, which may be a gene or other functional unit, such as a regulatory region. Several statistical methods have been proposed for rare variant analyses (as reviewed by Asimit and Zeggini [5]), but many of them experience a large drop in power when both protective and risk variants are present in a genetic region of interest. Indeed, over 20 different methods have been proposed in the last 2 to 3 years, each with distinct properties under different allelic architecture scenarios [5].

In a recent *PLoS Genetics* publication, Ionita-Laza *et al*. [6] introduced a testing approach, referred to as a replication-based strategy that is less sensitive to the presence of both risk and protective variants in the region. Here, we discuss the method and the authors' findings when comparing it with two existing rare variant analysis methods proposed by Li and Leal [3] and by Browning and Browning [7]. We also discuss the state of the field of rare variant analysis in complex traits, and the likely clinical implications when the proposed methods start to be applied to real data to identify novel disease associations at a wider scale.

The replication-based strategy

The method of Ionita-Laza *et al*. [6] is based on a weighted-sum statistic that evaluates whether an accumulation of rare variants occurs in a particular region at significantly higher frequencies in either cases or controls. Two one-sided test statistics may be formed: one for testing higher frequencies in cases than controls (risk variants) and one for the converse test of higher frequencies in controls than cases (protective variants). In forming each one-sided test statistic, the observed variants in cases and controls are partitioned into distinct groups according to the counts of the minor allele in cases, and the counts in controls.

The maximum of the two one-sided test statistics for testing higher frequencies of risk variants and higher frequencies of protective variants can then be used to test whether the region has an accumulation of rare variants that confer either protection or risk. In this way the strategy avoids combining the possibly different directional effects of the variants, which may dilute signals in cases where both risk and protective variants are present. As an alternative, Ionita-Laza *et al*. also investigated a combined test statistic that takes the sum of the two one-sided test statistics, rather than their maximum [6].

*Corresponding author: Jennifer Asimit ja11@sanger.ac.uk
Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH, UK

In simulation studies comparing the maximum weighted-sum statistic from the replication-based strategy with the collapsing method of Li and Leal [3] and the weighted-sum statistic of Browning and Browning [7], the maximum test statistic consistently attained the highest power. When both protective and risk variants existed in the region, the collapsing method and weighted-sum statistic experienced large decreases in power from the ideal situation of only one direction of effect for causal variants. In contrast, the approach proposed by Ionita-Laza *et al.* [6] was not as sensitive to the mixture of effect directions, although it still incurred moderate power losses.

Comparisons between the two versions of the replication-based strategy yielded interesting results. The maximum statistic showed a substantial power advantage over the combined statistic in a situation where only risk variants were present. However, as the number of protective variants in the region increased, the power of the combined statistic was found to surpass that of the maximum statistic. Therefore, both replication-based strategy approaches can provide useful information, but the choice of which of the two to use depends on the underlying assumptions of the directions of effect for the causal variants in the region under investigation. This is a welcome step forward in the analysis of rare variants, but the development of further tests for simultaneously handling the effects of both protective and risk variants are required.

Disease-related implications of rare variant associations

Low frequency and rare variants conferring susceptibility to common complex disease may have larger effect sizes compared to the common-variant associations identified to date. Higher penetrance may also indicate higher predictive power of disease, although it is clear that complex traits are underpinned by combinations of common and low frequency sequence variants, environmental factors and their interactions. Rare variant associations are also expected to more readily point to the causal locus or functional unit, potentially enhancing the translational potential of these findings.

Early examples of low frequency and rare variant associations with complex traits have arisen from targeted resequencing experiments. For example, Cohen *et al*. [1] detected a significant over-representation of non-synonymous variants in individuals with low plasma levels of high-density lipoprotein cholesterol compared with those with high plasma levels of high-density lipoprotein cholesterol. More recently, Nejentsev *et al*. [2] followed a pooled resequencing approach to identify four rare variants within the *IFIH1* (interferon induced with helicase C domain 1) gene that lower risk of type 1 diabetes independently of each other. Rare variant minor alleles are expected to potentially confer increased risk or protection against disease.

Challenges for the detection of rare disease-associated variants

Many tests for associations with rare variants have been proposed, but the majority experience a massive loss in power when both protective and risk variants are present in the region under analysis. The novel testing strategy proposed by Ionita-Laza *et al.* [6] is more robust to the presence of both directions of effect for causal variants at a locus. Their replication-based strategy using the maximum statistic was demonstrated to outperform two existing methods for rare variant analyses, and did not lose as much power as these methods. An alternative form of their statistic, taking the sum rather than the maximum of two one-sided test statistics, showed an even larger power increase when both protective and risk variants were present. Both the maximum and combined statistics represent an improvement over the existing methods they were compared against, but each performs better than the other under different assumptions. Indeed, it is widely accepted that different rare variant analysis methods perform optimally under different allelic heterogeneity scenarios. Due to the relative paucity of empirical data, there is no emerging consensus genetic architecture on which to focus further development. The advent of next-generation sequencing technologies has empowered a new genre of genetic association studies that focus on low frequency and rare variants. An appropriate analytical toolbox is being developed, with probably more methods than empirical datasets available at this point in time. It is expected that rare variation will play a role in the etiology of common complex diseases, and that new discoveries will shed further light into the biological underpinnings of pathological processes.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
JA and EZ are supported by the Wellcome Trust (WT088885/Z/09/Z).

Published: 27 April 2011

References

1. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 2004, **305**:869-872.
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. 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.
4. Bodmer W, Bonillna C: Common and rare variants in multifactorial

susceptibility to common diseases. *Nat Genet* 2008, **40:**695-701.
5. Asimit J, Zeggini E: **Rare variant association analysis methods for complex traits.** *Annu Rev Genet* 2010, **44:**293-308.
6. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C: **A new testing strategy to identify rare variants with either risk or protective effect on disease.** *PLoS Genet* 2011, **7:**e1001289.
7. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet* 2007, **81:**1084-1097.

doi:10.1186/gm238
Cite this article as: Asimit J, Zeggini E: **Testing for rare variant associations in complex diseases.** *Genome Medicine* 2011, **3:**24.