1728S

Utilizing Private Variants in Large Genome-Wide Association Studies: Issues, Techniques, Experiences. *U. Bodenhofer, S. Hochreiter.* Institute of Bioinformatics, Johannes Kepler University, Linz, Austria.

High-throughput sequencing technologies have facilitated the identification of large numbers of single-nucleotide variations (SNVs), many of which have already been proven to be associated with diseases or other complex traits. Several large sequencing studies, such as, the 1000 Genomes Project, the UK10K project, or the NHLBI-Exome Sequencing Project, have consistently reported a large proportion of private SNVs, that is, variants that are unique to a family or even a single individual. The role that private SNVs play in diseases and other traits is currently poorly understood — which is largely due to the fact that it is statistically very challenging to consider private SNVs in association testing. While it is generally impossible to use single-marker tests for private SNVs, burden tests are potentially able to deal with private SNVs, but only if the number of private SNVs occurring in a region is correlated with the trait under consideration. Moreover, burden tests have a disadvantage if deleterious and protective SNVs occur together in the same region. Non-burden tests like the popular SNP-set (Sequence) Kernel Association Test (SKAT) are typically utilizing correlations between SNVs — a strategy that is not applicable to private SNVs either, since singular events are generally uncorrelated. We propose the Position-Dependent Kernel Association Test (PODKAT), which is designed for detecting associations of very rare and private SNVs with the trait under consideration even if the burden scores are not correlated with the trait. PODKAT assumes that, the closer two SNVs are on the genome, the more likely they have similar effects on the trait under consideration. This assumption is fulfilled as long as deleterious, neutral, and protective variants are grouped sufficiently well along the genome. This contribution focuses on the use of PODKAT for large whole-genome studies. On the one hand, we will discuss issues related to data handling, computational complexity, and statistical significance. On the other hand, we will present results obtained for UK10K whole-genome cohorts that unveil the potential of considering private and very rare SNVs in genome-wide association studies.

1729M

A non-threshold region-specific method for detecting rare variants. *D.P. Chen[1], A.R. Hsieh[2], C.S.J. Fann[1].* 1) Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan; 2) Graduate Institute of Biostatistics, China Medical University, Taichung, Taiwan.

Rare variants have a proven role in some complex diseases. Many statistical methods proposed for the detection of rare variants associated with diseases have some limitations, such as the threshold of rare variants, and the direction of effects. Accordingly, we developed a region-specific method that do not use the threshold for defining rare variants and take the directions of effects into account. Our method also considers the linkage disequilibrium (LD) within the region, and can handle common and rare variants simultaneously. Our region-specific method used the concept of weighting variants according to their minor allele frequencies and odds ratios (OR) to combine effects of common and rare variants on disease occurrence into a single score, and provided a test statistic in assessing the significance of the score. To evaluate the performance of our method, we simulated extensively under different effect sizes according to Basu and Pan (2011). We found that the power of our method increased as the effect sizes increased. The type I error of our method was controlled well in spite of the simultaneous variations. Moreover, we compared our proposed method to several currently available methods, including kernel-based adaptive cluster (KBAC) and Sequence Kernel Association Test (SKAT). We found our method can generate comparable or better power in simulations. Results from our method showed a 15% increase in power comparing with SKAT (61% vs 47%) under small OR and lower LD, and 36% increase in power comparing with KBAC (98% vs 62%) while variants have different directions. However, our method performs well in 2-direction setting, but moderate in independent-variant scenario. We conclude that our proposed method can be used as a complementary tool with others to assist the dissection of the etiology of complex diseases.

1730T

Evaluating the calibration and power of three gene-based association tests for the X chromosome. *C. Ma, M. Boehnke, S. Lee.* Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI.

While genome-wide association studies (GWAS) have identified thousands of trait-associated genetic variants, the proportion of findings on the X chromosome lags behind those on the autosomal chromosomes. Existing X chromosome analysis methods focus on single marker association analysis. For analysis of rare variants (minor allele frequency < 0.5%), gene-based tests where multiple markers are analyzed jointly as a unit can be more powerful than single marker tests. To date, there are no gene-based tests designed to analyze the X chromosome. Using simulated case-control and quantitative trait (QT) studies, we evaluate the calibration and power of three gene-based tests for the X chromosome: burden, Sequence Kernel Association Test (SKAT), and optimal unified SKAT (SKAT-O). Specifically, we evaluate the impact of different ratios of males and females in cases and controls, and different coding of males alleles with X-inactivation (coding minor alleles as X=2) and without (X=1).

For case-control studies, all three tests are well-calibrated or slightly anti-conservative for all scenarios evaluated. As previously shown, power of the three tests depends on the underlying genetic architecture of the genomic region analyzed; burden is most powerful for multiple causal variants with the same direction of effect, SKAT is most powerful for causal variants with opposite directions of effect, and SKAT-O is generally powerful. For variants simulated assuming X-inactivation, coding male minor alleles as X=2 is slightly more powerful; for variants simulated assuming no X-inactivation, coding male minor alleles as X=1 is slightly more powerful. However, the power loss for misspecifying the generally unknown model is small. Different ratios of males and females in cases and controls have little effect on power. For QT studies, burden and SKAT are well-calibrated, while SKAT-O can be slightly anti-conservative across all scenarios. Power comparisons between tests for QTs are very similar to those for binary traits. We demonstrate that these three gene-based tests are well-calibrated and powerful for both binary and quantitative trait data, and can be directly applied to analyze rare variants on the X chromosome.