BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 9 2010, pages 1269–1270
doi:10.1093/bioinformatics/btq125

Genetics and population analysis

Advance Access publication March 22, 2010

varLD: a program for quantifying variation in linkage disequilibrium patterns between populations

Rick Twee-Hee Ong[1,2] and Yik-Ying Teo[3,4,*]

[1]NUS Graduate School for Integrative Science and Engineering, [2]Centre for Molecular Epidemiology, [3]Department of Statistics and Applied Probability and [4]Department of Epidemiology and Public Health, National University of Singapore, Singapore

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Linkage disequilibrium (LD) differences between populations can potentially result in failure to replicate primary signals of trait association in independent genome-wide association studies (GWAS). However, such inter-population LD differences can be leveraged to narrow the search for common causal variants responsible for the association signals observed. The ability to assess and quantify LD variations among populations is thus expected to contribute to both replication and fine-mapping stages of GWAS.

Availability: The program varLD is available for download from http://www.nus-cme.org.sg/software/varld.html

Contact: statyy@nus.edu.sg

Received on January 26, 2010; revised on March 8, 2010; accepted on March 17, 2010

1 INTRODUCTION

The aim of genome-wide association studies (GWAS) of common diseases and complex traits is to find statistically significant markers, usually single nucleotide polymorphisms (SNPs), which are associated with the phenotype of interest. These SNPs are seldom the genetic variants responsible for the phenotype, but are markers in linkage disequilibrium (LD) with the underlying causal variants. Thus, the findings from GWAS serve as indicators for genomic regions that are likely to possess the genetic variants directly responsible for the functional changes and differences in phenotypic expression.

To reduce statistical noise, it is necessary to validate the initial findings in replication studies performed in independent cohorts, which are often from different populations. Increasingly, meta-analysis of multiple GWAS studies of the same trait, involving tens of thousands of individuals are currently being performed (Easton *et al*., 2007; Kolz *et al*., 2009; Levy *et al*., 2009; Lindgren *et al*., 2009). In addition to confirming the initial GWAS findings, larger effective sample sizes from these meta-analyses also increase statistical power for identifying novel associations with smaller genetic effects. The large number of individuals in these studies meant it is highly unlikely they will come from the same population but rather from multiple populations potentially possessing diverse genetic architectures. The portability of association signals in these trans-population studies is thus highly dependent on the similarity of LD patterns between the associated markers and the underlying causal variants in the different populations. Inter-population heterogeneity in LD patterns can confound replication, leading to false negatives in some populations. The ability to quantify the extent of inter-population LD variation is thus the first step towards understanding whether a failure to reproduce an association signal in another population may be a result of differential LD architecture with the causal variant.

The presence of LD among SNP markers has allowed commercial genotyping platforms with incomplete SNP coverage to successfully detect genotype–phenotype associations. However, this becomes a liability in fine-mapping studies that aim to identify the actual causal variants, since it becomes difficult to distinguish between the causal variant and the neighbouring markers in high or perfect LD with the causal variant. Assuming there exists a common causal variant among different populations, diverse patterns of LD surrounding the SNPs that emerge from different GWAS studies can be used to refine the boundaries where the causal variant may be found, narrowing the search space and reducing the cost of genotyping and sequencing during fine-mapping studies (Teo *et al*., 2010).

Here, we introduce a java program (*varLD*) that quantifies the extent of LD variation between the two populations. The program allows genome-wide assessment of LD variation, as well as targeted analysis of a specific genomic region. The outcome from these analyses can be used to identify regions that display evidence of the inter-population LD variation. As these regions are likely to possess diverse haplotypic patterns across the populations (Teo *et al*., 2009a), association signals emerging from GWAS that fall within these regions will benefit from leveraging this genomic diversity in trans-population fine-mapping studies.

2 METHOD AND IMPLEMENTATION

The program adopts the approach described by Teo and colleagues (Teo *et al*., 2009a). Briefly, sliding windows of a pre-determined number of SNPs common in both populations are considered, and the LD between every pair of SNPs in each window is calculated. The LD here is quantified by the signed r^2 metric, defined as the r^2 metric with the sign of the D' metric. This yields a symmetric LD matrix for each population, and the extent of inter-population LD differences between the SNPs in each window is quantified by the varLD score, which sums the absolute differences between the ranked eigenvalues of the matrices. For genome-wide assessment, each score is compared against the distribution of the scores calculated from all possible windows across the genome, and typically scores existing in the extreme right tail of the distribution are defined as candidate regions displaying significant LD differences. For targeted analysis of a specific genomic region,

*To whom correspondence should be addressed.

a Monte Carlo (MC) statistical significance is calculated by comparing the score for the region to the scores generated after resampling the appropriate sample sizes for the two populations from the combined data produced by merging both populations (Teo *et al*., 2009a).

The program *varLD*, accepts tab-delimited genotype input files of the following format across the columns: rs-id/snp-id, position, genotype-1, genotype-2, genotype-3, etc. One file is expected for each chromosome or a region, and the SNPs in each file are assumed to be arranged such that the positions are in ascending order. Genotypes are coded as numerals 1,2,3,4 with 1 being the homozygote AA, 2 as heterozygote Aa, 3 as homozygote aa and 4 as missing.

varLD is implemented in Java, and runs on any platforms that supports v1.5 of the Java runtime environment. It requires the open source libraries of: (i) Apache Commons CLI library; (ii) Apache Commons Primitives library; and also (iii) JAMA, a Java Matrix package. Running on a 2.0 GHz machine with 16 GB of RAM, *varLD* with default parameters can perform the LD comparison for about one million SNPs common to two populations in approximately an hour where we run the analysis for the 22 autosomal chromosomes sequentially in a single batch run.

The program has the following features implemented: (i) user-specified threshold on minor allele frequency for the SNP exclusion; (ii) user-specified threshold on extent of missing genotypes for the SNP exclusion; (iii) user-specified number of SNPs to consider in each window during genome-wide assessment; (iv) allows the use of the same input file for genome-wide assessment and for assessment in a targeted genomic region, where the latter allows the user to specify the start and end position to consider; (v) generates a MC *P*-value in the targeted genomic region assessment, allowing the user to specify the number of iterations to perform.

3 EXAMPLE APPLICATION

Warfarin is a widely prescribed anticoagulant drug to reduce blood clotting in order to protect patients from stroke, thrombosis and heart attack. However, the dose requirement and drug response have been shown to vary significantly between patients from different population groups (Lee *et al*., 2006; Rieder *et al*., 2005). For example, Asian Indians have been observed to display warfarin resistance, thus requiring a higher dose as compared to Chinese and Malays (Lee *et al*., 2006). A genetic association has been observed for this inter-population variation in warfarin activity, with the identification of population-specific haplotypes in the VKORC1 gene that correlate with differences in warfarin dosage and response (Lee *et al*., 2006; Rieder *et al*., 2005; Schwarz *et al*., 2008). Thus, VKORC1 provide a convenient example to illustrate the use of varLD, since a priori we expect the LD patterns in this gene region to differ between these populations. The program *varLD* was used to calculate and compare the LD differences around the VKORC1 gene between populations in the International Hapmap Project (International HapMap Consortium, 2007) and the Singapore Genome Variation Project (Teo *et al*., 2009b). Figure 1 shows the normalized scores from the genome-wide assessment of LD variation in numerous pairs of populations from these two projects,

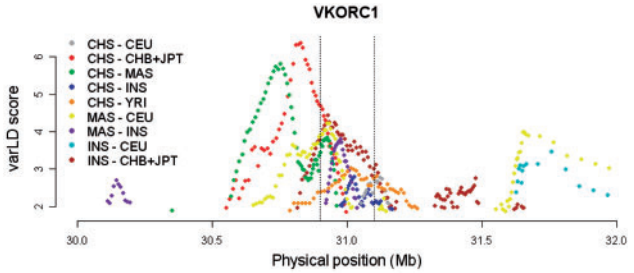

Fig. 1. Normalized varLD scores in the top 5% of the genome-wide distributions for comparisons between pairs of populations from the Singapore Genome Variation Project (SGVP) and International Hapmap Project (Hapmap). The vertical dotted lines represent the start and end of the VKORC1 gene. CEU, Hapmap Utah residents with ancestry from northern and western Europe; CHB, Hapmap Han Chinese from Beijing; CHS, SGVP Chinese from Singapore; INS, SGVP Indian from Singapore; JPT, Hapmap Japanese from Tokyo; MAS, SGVP Malay from Singapore; YRI, Hapmap Yoruba from Nigeria.

indicating strong evidence of inter-population LD variations in and around the VKORC1 gene. Similarly, in a targeted comparison within the VKORC1 gene with 10 000 MC iterations, most of the population pairs indicate moderate to strong evidence (MC $P < 0.01$) of LD variations, except that between Hapmap CEU and SGVP INS (MC $P = 0.5044$).

Conflict of Interest: none declared.

REFERENCES

Easton,D.F. *et al*. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.

International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Kolz,M. *et al*. (2009) Meta-analysis of 28 141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet.*, **5**, e1000504.

Lee,S.C. *et al*. (2006) Inter-ethnic variability in warfarin requirement is explained by VKORC1 genotype in an Asian population. *Clin. Pharmacol. Ther.*, **79**, 197–205.

Levy,D. *et al*. (2009) Genome-wide association study of blood pressure and hypertension. *Nat. Genet.*, **41**, 677–687.

Lindgren,C.M. *et al*. (2009) Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet.*, **5**, e1000508.

Rieder,M.J. *et al*. (2005) Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N. Engl. J. Med.*, **352**, 2285–2293.

Schwarz,U.I. *et al*. (2008) Genetic determinants of response to warfarin during initial anticoagulation. *N. Engl. J. Med.*, **358**, 999–1008.

Teo,Y.Y. *et al*. (2009a) Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res.*, **19**, 1849–1860.

Teo,Y.Y. *et al*. (2009b) Singapore Genome Variation Project: A haplotype map of three South-East Asian populations. *Genome Res*., **19**, 2154–2162.

Teo,Y.Y. *et al*. (2010) Methodological challenges of genome-wide association analysis in Africa. *Nat. Rev. Genet.*, **11**, 149–160