BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 14 2014, pages 2086–2088
doi:10.1093/bioinformatics/btu171

Genetics and population analysis

Advance Access publication April 1, 2014

SECA: SNP effect concordance analysis using genome-wide association summary results

Dale R. Nyholt

Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane 4000, Queensland, Australia

Associate Editor: Gunnar Ratsch

ABSTRACT

Summary: The genomics era provides opportunities to assess the genetic overlap across phenotypes at the measured genotype level; however, current approaches require individual-level genome-wide association (GWA) single nucleotide polymorphism (SNP) genotype data in one or both of a pair of GWA samples. To facilitate the discovery of pleiotropic effects and examine genetic overlap across two phenotypes, I have developed a user-friendly web-based application called SECA to perform SNP effect concordance analysis using GWA summary results. The method is validated using publicly available summary data from the Psychiatric Genomics Consortium.

Availability and implementation: http://neurogenetics.qimrberghofer.edu.au/SECA.

Contact: dale.nyholt@qimrberghofer.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 22, 2013; revised on February 20, 2014; accepted on March 24, 2014

1 INTRODUCTION

Epidemiological and clinical studies indicate many human complex disorders co-occur within an individual, whereas family and twin studies suggest correlations in familial and genetic liabilities. Since 2007, >1700 genome-wide association studies (GWAS) have been performed to identify common single nucleotide polymorphisms (SNPs), associated with >890 phenotypes (disease endpoints or quantitative traits) (Hindorff *et al.*, 2011; www.genome.gov/gwastudies, accessed September 18, 2013). A surprising finding of GWAS is that many loci show pleiotropic effects by being associated with more than one distinct phenotype. A study of 1380 genes and 1687 SNPs listed in the NHGRI Catalog of Published GWAS (www.genome.gov/gwastudies, accessed February 4, 2011) found 16.9% of genes and 4.6% of SNPs show pleiotropic effects (Sivakumaran *et al.*, 2011).

Identifying and taking advantage of genetic overlap across phenotypes can improve detection of genetic risk factors because when risk is correlated across phenotypes, pooled analyses will be better powered than individual-disorder analyses. Genetic overlap can also help determine whether our current classification and categorization of diseases are valid or whether genetic similarities traverse current divisions. Understanding pleiotropic effects is also important for drug development. For example, if a gene has opposing effects on different common diseases, drug development and marketing will be greatly complicated. Knowledge of pleiotropic associations could help to predict and avoid adverse side effects.

Polygenic prediction, or genetic risk score (GRS), analysis (Purcell *et al.*, 2009) and bivariate linear mixed model (bivariate-GCTA) analysis (Lee *et al.*, 2012) have recently been developed to assess the genetic overlap across phenotypes.

GRS analysis requires GWAS summary results for a *discovery* sample (*dataset1*) to construct a score in an independent *target* sample (*dataset2*) by forming the weighted sum of associated alleles (i.e. a GRS) within each subject (thus requiring individual-level GWA SNP genotype data for the target sample). Association between the GRS and a phenotype in the target sample implies genetic overlap between the discovery and target phenotypes.

The bivariate-GCTA approach requires individual-level population-based GWA SNP genotype data for both the discovery and target sample to estimate the genetic correlation explained by SNPs between pairs of quantitative traits or pairs of binary traits. For example, the SNP-based genetic correlation ($r_{g\ SNP}$) will be positive when the cases of one (discovery) trait show higher genetic similarity (i.e. pairwise genetic relationship estimated from genome-wide SNP genotype data) to the cases of the other (target) trait, than they do to their own (discovery) controls.

The sharing of individual-level genotype and phenotype data requires application, material transfer agreement, informed consent and ethical consideration (McGuire *et al.*, 2011). Given most large GWA studies meta-analyze summary results from multiple independent samples, the sharing of individual-level GWA SNP genotype data is often difficult, if not impossible. Therefore, to overcome these challenges, I have developed a user-friendly web-based application called SECA to perform SNP effect concordance analysis using GWAS summary results.

2 METHODS

SECA analyzes two sets of GWAS summary results (Supplementary Fig. S1), each containing five essential columns: (i) reference SNP cluster 'rs' ID (SNP), (ii) effect allele (EA), (iii) non-effect allele (NEA), (iv) *P*-value from association test (*P*) and (v) regression coefficient (BETA), odds ratio (OR) or signed z-score (ZSCORE) for the EA relative to the NEA.

SECA first aligns the SNP effects across the two GWAS summary results (*dataset1* and *dataset2*) to the same EA, and extracts a subset of independent SNPs via linkage disequilibrium (LD) clumping using the PLINK program (Purcell *et al.*, 2007). The default clumping procedure is '*P*-value informed'. The approach iterates from the first to last SNP on each chromosome sorted from smallest to largest *P*-value in *dataset1* (P_1)

that has not already been clumped (denoting this as the index SNP) and forms clumps of all other SNPs that are within 1 Mb and in LD ($r^2 > 0.1$) with the index SNP based on HapMap or 1000 Genomes Project (1 kGP) genotype data. A second round of LD clumping is performed to ensure none of the round 1 index SNPs within 10 Mb of each other are in long-range LD ($r^2 > 0.1$). The default approach identifies the subset of independent SNPs with the most significant association *P*-values in *dataset1*. An alternate LD clumping approach identifies an effectively random subset of independent SNPs by setting all *dataset1* *P*-values (P_1) to 0.5 during the clumping rounds.

Restricting to SNPs associated with $P_1 \leq \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ in *dataset1*, SECA first performs exact binomial statistical tests using the R statistical package (R Development Core Team, 2009) to determine whether there is an excess of SNPs associated in both datasets for the subset of SNPs associated with $P_2 \leq \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ in *dataset2*. For each of the 144 SNP subsets (generated using these 12×12 *P*-value threshold combinations), binomial tests of associated SNPs in *dataset1* (P_1) and *dataset2* (P_2) are performed. A binomial test 'heatmap' plot is generated to graphically summarize the proportion of SNP subsets with an excess [observed (obs) ≥ expected (exp)] or deficit (obs < exp) number of associated SNPs, and empirical *P*-values (adjusted for testing all 144 subsets) are calculated via permutation. The permutation procedure first creates uncorrelated datasets by randomly shuffling the observed SNP effect (BETA) and corresponding *P*-value (*P*) between SNPs in *dataset1*, and then repeats the analysis of the 144 SNP subsets. A permuted *P*-value ($P_{\text{BTsig-permuted}}$) is estimated for the observed number of binomial tests with obs ≥ exp and $P_{\text{BT}} \leq 0.05$ (n_{BTsig}), and for the single most significant excess ($P_{\text{BTmin-permuted}}$). We note that for polygenic traits, a well-powered GWAS should produce an excess of smaller *P*-values (Yang *et al.*, 2011). Therefore, instead of specifying P_2 as the expected proportion of SNPs with P_1 and P_2, the proportion of SNPs associated in *dataset2* (irrespective of its *P*-value in *dataset1*) is used as the 'overlap' (null) probability (i.e. 'expected proportion') in the binomial tests. For example, the observed proportion of total SNPs with $P_1 \leq 1$ and $P_2 \leq 0.05$ is used as the expected proportion of SNPs with $P_1 \leq \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and $P_2 \leq 0.05$.

SECA next performs Fisher's exact statistical tests to determine whether there is an excess of SNPs where the effect directions (BETA, OR or ZSCORE) are concordant across *dataset1* and *dataset2* for each of the 144 SNP subsets. Fisher's exact tests of SNP effects in *dataset1* (e.g. OR_1) and *dataset2* (OR_2) are performed on 2×2 tables containing the number of SNPs with $OR_1 < 1$ ($BETA_1/ZSCORE_1 < 0$) and $OR_2 < 1$, $OR_1 \geq 1$ ($BETA_1/ZSCORE_1 \geq 0$) and $OR_2 < 1$, $OR_1 < 1$ and $OR_2 \geq 1$ and $OR_1 \geq 1$ and $OR_2 \geq 1$. A Fisher's test 'heatmap' plot is generated to graphically summarize the proportion of SNP subsets with concordant (Fisher's test odds ratio, $OR_{\text{FT}} \geq 1$) and discordant ($OR_{\text{FT}} < 1$) SNP effects, and an empirical *P*-value ($P_{\text{FTsig-permuted}}$) is calculated via permutation for the observed number of subsets (n_{FTsig}) with nominally significant concordance ($OR_{\text{FT}} \geq 1$ and $P_{\text{FT}} \leq 0.05$). A permuted *P*-value is also estimated for the single most significant concordant test ($P_{\text{FTmin-permuted}}$).

In addition to a summary of the above analyses, SECA reports detailed results from binomial and Fisher's tests for three practical subsets of SNPs, which are nominally associated ($P_2 \leq 0.05$) in *dataset2* and (i) genome-wide *significant* ($P_1 \leq 5 \times 10^{-8}$), (ii) genome-wide *suggestive* ($P_1 \leq 1 \times 10^{-5}$) and (iii) nominally associated ($P_1 \leq 0.05$) in *dataset1*.

A 'grid' search is also performed across 150 *P*-value thresholds to identify the subset of SNPs overlapping *dataset1* and *dataset2* that produce the minimum binomial test 'pleiotropic' *P*-value and minimum Fisher's test (effect concordance) *P*-value. The search is performed at 5 equidistant *P*-value $\{P_1, P_2\}$ increments for the interval $[5 \times 10^{-8}, 9 \times 10^{-8}]$, i.e. $\{5 \times 10^{-8}, 6 \times 10^{-8}, 7 \times 10^{-8}, 8 \times 10^{-8}, 9 \times 10^{-8}\}$; 9 equidistant increments for each of the intervals $[1 \times 10^{-7}, 9 \times 10^{-7}]$, $[1 \times 10^{-6}, 9 \times 10^{-6}]$, $[1 \times 10^{-5}, 9 \times 10^{-5}]$, $[1 \times 10^{-4}, 9 \times 10^{-4}]$, $[1 \times 10^{-3}, 9 \times 10^{-3}]$; and 100 increments of 0.01 between 0.01 and 1.0. The results from all $150 \times 150 = 22\,500$ SNP subsets are output to enable post hoc examination of SNP overlap and concordance across the full range of statistical significance in *dataset1* and *dataset2*.

Additionally, SECA generates quantile–quantile (Q-Q) and true discovery rate (TDR) plots for *dataset2* *P*-values (P_2) stratified (conditioned) on *dataset1* *P*-values ($P_1 \leq \{0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1.0\}$) to visualize whether there is an excess of overlapping 'pleiotropic' SNPs. A leftward shift in the Q-Q or TDR curve corresponds to an excess of SNPs with smaller *P*-values. In the presence of pleiotropy, we expect the Q-Q and TDR curves for *dataset2* SNP *P*-values (P_2) to deviate further left of the identity line because we condition on them having smaller *P*-values (P_1) in *dataset1*. Finally, 'pleiotropy-informed' conditional false discover rate (FDR) results are output for *dataset2* *P*-values (P_2) stratified on *dataset1* *P*-values ($P_1 \leq \{1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.01, 0.001, 0.0001, 0.00001\}$). These results may identify SNPs associated in *dataset2* after conditioning on their association significance (P_1) in *dataset1* and help prioritize SNPs for follow-up studies. Coincidently, Andreassen *et al.* (2013) developed in parallel (but without accompanying software), similar Q-Q and TDR plot and FDR-based approaches to identify pleiotropic effects. An overview of the SECA approach is provided in Supplementary Figure S2.

Resources to customize LD clumping, an alternate version of SECA (iSECA), which assumes the SNPs in *dataset1* to be independent, and an R script allowing users to alter the number of replicates (default = 1000) for the permutation of binomial and Fisher's test *P*-values are also provided.

3 RESULTS

Using publicly available GWAS summary data, SECA corroborates recent results from the Cross-Disorder Group of the Psychiatric Genomics Consortium (Smoller *et al.*, 2013), finding significant polygenic overlap between bipolar disorder, major depressive disorder, schizophrenia and autism spectrum disorder, but generally not with attention deficit hyperactivity disorder (further details and results are provided in the Supplementary Material).

Funding: The Australian Research Council (FT0991022) and National Heath and Medical Research Council (APP0613674).

Conflict of Interest: none declared.

REFERENCES

Andreassen,O.A. *et al.* (2013) Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors, *Am. J. Hum. Genet.*, **92**, 197–209.

Hindorff,L.A. *et al.* (2011) A Catalog of Published Genome-Wide Association Studies. Available at: http://www.genome.gov/gwastudies (2 April 2014, date last accessed).

Lee,S.H. *et al.* (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, **28**, 2540–2542.

McGuire,A.L. *et al.* (2011) Ethical and practical challenges of sharing data from genome-wide association studies: the eMERGE Consortium experience. *Genome Res.*, **21**, 1001–1007.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*, **81**, 559–575.

Purcell,S.M. *et al.* (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.

R Development Core Team. (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sivakumaran,S. *et al.* (2011) Abundant pleiotropy in human complex diseases and traits, *Am. J. Hum. Genet.*, **89**, 607–618.

Smoller,J.W. *et al.* (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, **381**, 1371–1379.

Yang,J. *et al.* (2011) Genomic inflation factors under polygenic inheritance, *Eur. J. Hum. Genet.*, **19**, 807–812.