BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 11 2009, pages 1445–1446
doi:10.1093/bioinformatics/btp159

Genetics and population analysis

Synergy Disequilibrium Plots: graphical visualization of pairwise synergies and redundancies of SNPs with respect to a phenotype

John Watkinson and Dimitris Anastassiou*

Center for Computational Biology and Bioinformatics, Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, NY 10027, USA

Received on February 9, 2009; revised on March 7, 2009; accepted on March 15, 2009

Advance Access publication March 17, 2009

Associate Editor: Martin Bishop

ABSTRACT

Summary: We present a visualization tool applied on genome-wide association data, revealing disease-associated haplotypes, epistatically interacting loci, as well as providing visual signatures of multivariate correlations of genetic markers with respect to a phenotype.

Availability: Freely available on the web at: http://www.ee.columbia.edu/~anastas/sdplots

Contact: anastas@ee.columbia.edu

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

1 INTRODUCTION

Linkage disequilibrium (LD) plots provide direct visualization of pairwise associations of single nucleotide polymorphisms (SNPs). They can be computed from large genotype sample sets, such as those from the HapMap project, or from a genome-wide association (GWA) study by plotting the matrix of association values for each pair of SNPs.

LD plots have been used to identify blocks of SNPs associated with a trait or disease for a given population (Duerr *et al*., 2006). However, a more direct approach incorporates the presence of the trait directly into the disequilibrium metric. The metric should measure the amount of cooperative or redundant association of the SNPs with the trait, allowing detection of 'epistasis', occurring when the effects of a genetic factor on a trait is modified by another factor. Given the limited success of identifying significant individual risk-conferring variants for some disorders, it is hoped that the discovery of responsible epistatic interactions among genetic variants reflecting molecular elements in complex pathways will elucidate novel disease mechanisms. The increasingly available large biological datasets coming from GWA studies provide a unique opportunity to discover such multivariate correlations. Here we provide a software package introducing synergy disequilibrium (SD) plots as visual tools identifying disease-associated haplotypes, as well as epistatically interacting loci with respect to disease.

2 DESCRIPTION

Synergy is an information theoretic quantity well-suited for discovering epistatic interactions. The synergy between two SNPs S_i and S_j with respect to a disease C (or any phenotype or trait) is defined (Anastassiou, 2007) as the amount of information conveyed by the pair of SNPs about the presence of the disease, minus the sum of the corresponding amounts of information conveyed by each SNP:

$$I(S_i, S_j; C) - \left[I(S_i; C) + I(S_j; C)\right]$$

This is consistent with the definition (American Heritage Dictionary) of synergy as 'the interaction of two or more agents or forces so that their combined effect is greater than the sum of their individual effects'. Thus, synergy quantifies the amount of association between two SNPs and a phenotype that is due to purely cooperative effects among the factors. Equivalently, the synergy is equal to the increase or decrease of the information that an SNP provides about the presence of the disease as a result of knowledge of the other SNP.

Large positive synergy suggests an epistatic interaction mechanism, as it can be seen as a part of the information conveyed by the pair of SNPs about the presence of the disease that is attributable to a purely cooperative interaction between the two SNPs. On the other hand, negative synergy of large magnitude indicates that the two SNPs are redundantly associated with the presence of the disease, i.e. each SNP by itself is associated with disease, while including both SNPs in combination does not significantly enhance this association. For example, multiple pairwise redundant SNPs may define a 'disease-associated haplotype' when there is a strong LD connecting these SNPs and all of them appear individually associated with disease, suggesting that a 'causal' biological hotspot may be located within that region.

We can use the GWA data to define random variables (three-valued in the case of genotyped SNPs) by creating probabilistic models from relative frequencies after counting the number of healthy samples as well as the number of diseased samples encountered in each joint state. Once we have the model defined, then we can readily evaluate all the information theoretic quantities directly from these counts (Anastassiou, 2007; Varadan and Anastassiou, 2006).

When coupled with additional biological knowledge and validation, SD analysis has the potential to shed light on the nature of etiological mechanisms, e.g. by analyzing gene product isoforms from the identified disease-associated haplotypes or by exploring known pathways connecting high-synergy genes.

3 EXAMPLES OF SD PLOTS

We used GWA data generated from the Wellcome Trust Case Control Consortium (WTCCC) including about 2000 cases for

*To whom correspondence should be addressed.

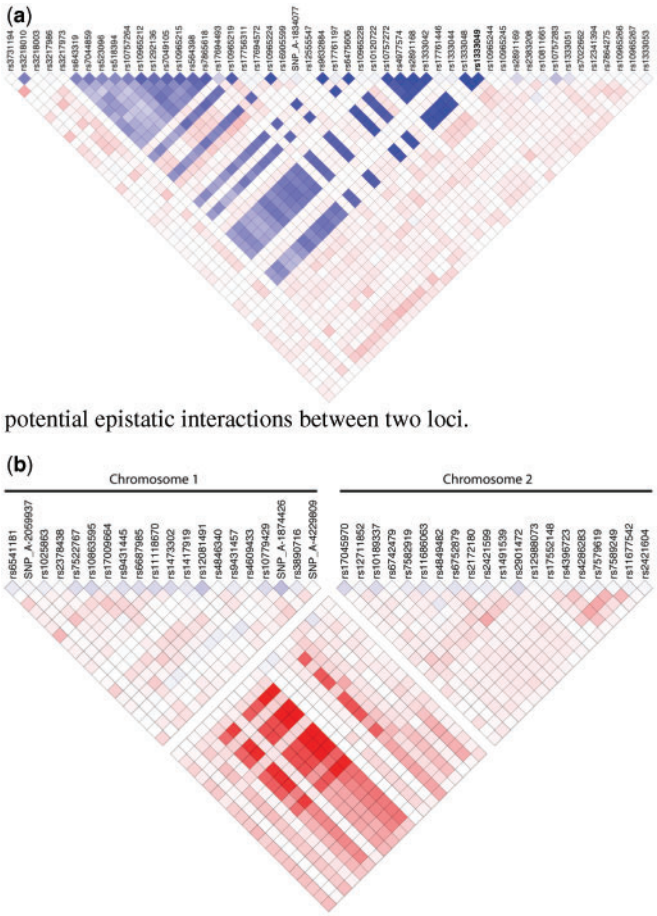

Fig. 1. (**a**) SD plot of a disease-related haplotype. (**b**) SD plot of potential epistatic interactions between two loci.

each of several diseases and 3004 controls (Wellcome Trust Case Control Consortium, 2007). These synergies are plotted in a manner analogous to that used for LD plots, resulting in an SD plot.

For example, Figure 1a shows an SD plot that includes the SNPs belonging to a coronary artery disease (CAD)-associated haplotype. The magnitude of synergy of an SNP pair is depicted by the intensity of the corresponding red diamond-shaped dot, while the magnitude of the redundancy of an SNP pair is depicted by the intensity of the corresponding blue diamond-shaped dot. This plot was generated by considering the region around SNP rs1333049, which was previously found to be the most strongly associated with CAD (Wellcome Trust Case Control Consortium, 2007). That SNP appears at the rightmost end of the haplotype (set of mutually redundant SNPs with respect to the disease) defined by the blue entries. Therefore, it is seen that the disease-associated haplotype extends upstream of rs1333049. In such cases, the precise haplotype can be easily found from the SD plot by phasing (inferring the SNP values in each chromosome from the genotypes) and analyzing the corresponding genomic area. The 'top row' of dots in SD plots depicts the 'self-synergy' of an SNP with respect to disease, which measures the SNP's correlation with disease.

On the other hand, Figure 1b shows an area of potential epistatic interaction between two genomic regions at different chromosomes with respect to Crohn's disease, which we also generated using the WTCCC data. The plot in Figure 1b is 'broken', because it connects two genomic regions in different chromosomes. The two high-synergy genomic regions are defined by the high-intensity red dots.

Another example concerns the known effects of the major histocompatibility complex class II region, on Type 1 diabetes. Many susceptible and protective combinations of allelic variations are known and epistatic interactions between alleles have been hypothesized (Erlich *et al*., 2008; Koeleman *et al*., 2004). In addition to the WTCCC dataset, we obtained an independent validation dataset from the Type 1 Diabetes Genetics Consortium with 4844 cases and 3552 controls (without any overlap with the 5004 samples of the WTCCC dataset). After imputing SNP values, we produced the two corresponding SD plots (Supplementary Figure), which were remarkably similar, demonstrating that SD plots provide robust and reliable visual 'signatures' depicting aspects of multivariate correlations with respect to particular phenotypes.

ACKNOWLEDGEMENTS

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who con-tributed to the generation of the data is available from www.wtccc.org.uk.

Funding: The project was funded by the Wellcome Trust under award 076113. It also utilizes resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and Juvenile Diabetes Research Foundation International (JDRF) and supported by U01 DK062418.

Conflict of Interest: none declared.

REFERENCES

Anastassiou,D. (2007) Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.*, **3**, 83.

Duerr,R.H. *et al.* (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, **314**, 1461–1463.

Erlich,H. *et al*. (2008) HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. *Diabetes*, **57**, 1084–1092.

Koeleman,B.P. *et al*. (2004) Genotype effects and epistasis in type 1 diabetes and HLA-DQ trans dimer associations with disease. *Genes Immun.*, **5**, 381–388.

Varadan,V. and Anastassiou,D. (2006) Inference of disease-related molecular logic from systems-based microarray analysis, *PLoS Comput. Biol.*, **2**, e68.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.