BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 23 2006, pages 2960–2961
doi:10.1093/bioinformatics/btl518

Genetics and population analysis

ssSNPer: identifying statistically similar SNPs to aid interpretation of genetic association studies

Dale R. Nyholt

Genetic Epidemiology Laboratory, QIMR, 300 Herston Road, Brisbane, Queensland, 4006, Australia

Received on September 8, 2006; revised and accepted on October 5, 2006
Advance Access publication October 12, 2006
Associate Editor: Martin Bishop

ABSTRACT

Summary: ssSNPer is a novel user-friendly web interface that provides easy determination of the number and location of untested HapMap SNPs, in the region surrounding a tested HapMap SNP, which are statistically similar and would thus produce comparable and perhaps more significant association results. Identification of ssSNPs can have crucial implications for the interpretation of the initial association results and the design of follow-up studies.

Availability: http://fraser.qimr.edu.au/general/daleN/ssSNPer/

Contact: daleN@qimr.edu.au

Most papers on genetic association conclude with the obligatory genuflection that the identified variant may not be causal, but in tight linkage disequilibrium (LD) with another that is. Here, I try to quantify this problem and highlight the utility of identifying 'statistically similar' single nucleotide polymorphisms (ssSNPs) in the interpretation of initial association results and in the design of follow-up studies.

Given practical limitations on genotyping, investigators are forced to select and prioritize test SNPs from the ~12 million SNPs currently present in the dbSNP (build 126) database (http://www.ncbi.nlm.nih.gov/projects/SNP/). It has therefore become routine to utilize correlations (LD) among nearby variants to guide the selection of informative 'tag' SNPs; the most extensive genome-wide resource being the International HapMap Project, currently containing ~6.3 million SNPs (http:/www.hapmap.org/).

The degree of LD between alleles at two SNPs can be described in terms of the common measure r. The LD measure r, is the correlation coefficient for a 2×2 table and is usually squared to remove the arbitrary sign introduced when the SNP alleles are labelled. The resulting metric (r^2) is useful in association analyses as it expresses the proportion of variance in SNP 1 'explained' or predicted by SNP 2 and vice versa. As a consequence, r^2 provides a measure of statistical similarity between SNPs. Put another way, r^2 measures the statistical power SNP 1 has to predict the genotypes of SNP 2 and vice versa (Hill and Robertson, 1968).

In candidate gene-wide association studies, SNPs may be selected based on plausible biological causality [e.g. nonsynonymous coding SNPs (nsSNPs), predicted transfactor binding sites, predicted miRNA target sites, etc.] and/or tagSNPs may be selected until an $r^2 \geq 0.8$ (for example) is exceeded for all sites (Carlson *et al*., 2004). The latter case also pertains to positional genome-wide association studies where 100 000s of tagSNPs are required to satisfactorily screen the entire human genome.

Although it is becoming common practice for investigators to utilize LD among nearby variants in the design of association studies, I wish to emphasize that once a significant association is obtained it is important to carefully examine inter-marker LD, as it can have crucial implications for the interpretation of the initial association results and the design of follow-up studies.

To explain, first consider a pair of SNPs with $r^2 = 1.0$; these SNPs are 'statistically indistinguishable' (siSNPs) [also previously termed 'genetically indistinguishable', giSNPs (Lawrence *et al*., 2005)] and barring genotype error will produce statistically identical association results. In addition, untyped SNPs with $r^2 \geq 0.8$ can be considered to be of high enough similarity (i.e. SNP 1 has 80% power to predict the genotypes of SNP 2 and vice versa) to produce similar and importantly, perhaps more significant association results compared to the original associated test SNP. Therefore, it is recommended that researchers investigate ssSNPs (especially where $r^2 \geq 0.8$) and their surrounding plausible causative region(s) in addition to the region(s) initially implicated by the test SNP.

To demonstrate the utility of identifying HapMap ssSNPs, I examined the 1 Mb region surrounding the test SNPs associated in the four studies (Amundadottir *et al*., 2006; Graham *et al*., 2006; Grant *et al*., 2006; Smyth *et al*., 2006) recently appearing and discussed (Todd, 2006) in Nature Genetics. Briefly, HapMap SNPs (release #21) genotyped in the CEPH trios within 500 kb either side of the most significantly associated (HapMap) test SNP in each study were examined via my novel ssSNPer web interface (described below).

Only one HapMap SNP (rs12243326, $r^2 = 0.969$) 20 087 bp upstream would likely produce an association result similar (i.e. with $r^2 \geq 0.8$) to the original association of rs12255372 with type 2 diabetes (Grant *et al*., 2006). Given that both rs12243326 and rs12255372 are within the *TCF7L2* gene, it is reasonable to conclude that further studies aimed at identifying the causal variant(s) should concentrate solely on *TCF7L2*.

Similarly, only one HapMap SNP (rs2111485, $r^2 = 0.882$) would have high power to produce an association result comparable to the type 1 diabetes (T1D) associated nsSNP rs1990760 (exon 15) (Smyth *et al*., 2006). The ssSNP rs2111485 is 13 515 bp downstream from rs1990760, lies 13 053 bp downstream of *IFIH1* and 23.5 kb upstream from the next nearest gene (*FAP*). Hence the ssSNP information supports the biological evidence indicating *IFIH1* to be the strongest candidate. However, two ssSNPs {rs984971 (intergenic), rs2068330 (intron 14)} with $r^2 = 0.72$ were identified within the other functional candidate gene in the region mentioned by the authors (*KCNH7*). Importantly, these two SNPs

were also reported by Smyth *et al*., (2006) to be highly associated with T1D. Hence, although *IFIH*1 remains the more likely candidate due to association of the nsSNP rs1990760 and further studies should (initially) concentrate on this gene, the possibility remains that there are other regulatory variants within the *IFIH*1 region, including *KCNH*7 and *FAP*.

In contrast, there are 17 ssSNPs (Table 1) ranging from 15 305 bp upstream to 89 846 bp downstream of the rs2280714 SNP most strongly associated with elevated *IRF*5 expression levels, which is associated with increased risk to systemic lupus erythematosus (Graham *et al*., 2006), making the association interpretation more complex. In particular, all but one of the 17 ssSNPs (rs752637) is 3′ to *IRF*5 and lie within the *TNPO*3 gene. The authors provide a strong case for a role of *IRF*5 based upon expression data and biological plausibility; however they note that the SNPs most strongly associated with *IRF*5 expression are well downstream of *IRF*5 and do not lie in a recognizable regulatory region and hypothesize that an additional (unknown) genetic variant in tight LD with rs2280714 may drive the expression phenotype. Consequently, to thoroughly investigate the association and identify the causal variant(s), further analyses should involve both the *IRF*5 and neighbouring *TNPO*3 gene.

Correspondingly, 19 ssSNPs (Table 1) ranging from 815 bp 5′ to 54 322 bp 3′ were identified for the rs1447295 SNP associated with prostate cancer (Amundadottir *et al*., 2006). Intriguingly, there are currently no genes identified within the 55 137 bp region spanned by the 19 ssSNPs. However, there are numerous conserved regions and two pseudogenes (*DQ*515896 and *DQ*515897) within the ssSNP-identified region which provide likely targets for further study. Of course, as in all HapMap-based analyses, other (ss)SNPs, not currently genotyped in the HapMap may exist which have the potential to assist investigators identify the causal variant(s), thus highlighting the need for continued SNP discovery.

The ssSNPer web interface is a simple yet powerful tool, which allows users to merely upload to our web server a file specifying a list of test HapMap refSNP (rs) IDs and a file containing HapMap SNP genotype data (step-by-step instructions for obtaining HapMap genotype data for regions up to 5 Mb are provided) to determine the number and location of ssSNPs in the surrounding region. Single test SNP ssSNPer output first includes a graph providing an overview of the statistical similarity (r^2), based on expectation-maximization (EM) estimated haplotype frequencies (Excoffier and Slatkin 1995), between the test SNP and all surrounding HapMap SNPs. Following the graph is a list of the surrounding SNPs which have at least 50% similarity (i.e. $r^2 \geq 0.5$) first sorted from highest to lowest r^2 and then according to map order (i.e. from pter to qter). Map distances are given as base pair from the test SNP. The final section reports all r^2 values (i.e. $0 \leq r^2 \leq 1$) between the test SNP and all HapMap SNPs across the submitted region in map order; these data are suitable for plotting along with other map landmarks (e.g. genes, conserved regions, etc). Multiple test SNP ssSNPer output is restricted to the section listing ssSNPs with $r^2 \geq 0.5$ and is also useful for choosing replacement SNPs in the late stages of association study design, where ssSNPs can be preferentially chosen with respect to their r^2 values to replace previously selected SNPs which do not fulfil assay design requirements.

Table 1. ssSNPs for two recent studies published in Nature Genetics

Graham *et al*., (2006) Test SNP: rs2280714			Amundadottir *et al*., (2006) Test SNP: rs1447295		
ssSNP	bp from test SNP	r^2	ssSNP	bp from test SNP	r^2
rs7789423	26 479	1.000	rs6470519	−815	1.000
rs6948928	28 372	1.000	rs7818556	−639	1.000
rs3857852	62 552	1.000	rs10109700	926	1.000
rs13221560	68 450	1.000	rs7826179	4161	1.000
rs1072767	78 480	1.000	rs9643226	9443	1.000
rs921403	81 999	0.955	rs1447296	10 321	1.000
rs10279821	88 822	0.955	rs10808558	16 112	1.000
rs10156169	89 846	0.931	rs7832031	31 914	1.000
rs752637	−15 305	0.888	rs4242382	32 535	1.000
rs6969930	35 588	0.834	rs4314621	32 977	1.000
rs12155080	64 014	0.834	rs4242384	33 516	1.000
rs10239340	73 785	0.834	rs7812429	35 135	1.000
rs10229001	4672	0.832	rs7812894	35 441	1.000
rs2272347	24 690	0.832	rs7814837	37 164	1.000
rs7807018	45 463	0.832	rs4582524	43 397	1.000
rs7385716	10 869	0.831	rs13255059	45 578	1.000
rs6965542	61 193	0.800	rs11986220	46 651	1.000
			rs10090154	47 099	1.000
			rs7837688	54 322	0.824

ACKNOWLEDGEMENTS

The author would like to thank Professors Nick Martin, Peter Visscher and Grant Montgomery and the anonymous reviewers for valuable comments on earlier drafts of this manuscript. The author is supported by an NHMRC Fellowship (#339462). This research was supported in part by NHMRC Program Grant (#389892).

Conflict of Interest: none declared.

REFERENCES

Amundadottir,L.T. *et al*. (2006) A common variant associated with prostate cancer in European and African populations. *Nat. Genet*., **38**, 652–658.

Carlson,C.S. *et al*. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet*., **74**, 106–120.

Excoffier,L. and Slatkin,M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol*., **12**, 921–927.

Graham,R.R. *et al*. (2006) A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. *Nat. Genet*., **38**, 550–555.

Grant,S.F. *et al*. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet*., **38**, 320–323.

Hill,W.G. and Robertson,A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet*., **38**, 226–231.

Lawrence,R. *et al*. (2005) Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. *Genome Res*., **15**, 1503–1510.

Smyth,D.J. *et al*. (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat. Genet*., **38**, 617–619.

Todd,J.A. (2006) Statistical false positive or true disease pathway? *Nat. Genet*., **38**, 731–733.