BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 7 2010, pages 969–970
doi:10.1093/bioinformatics/btq068

Genetics and population analysis
Advance Access publication February 19, 2010

CandiSNPer: a web tool for the identification of candidate SNPs for causal variants

Armin O. Schmitt*, Jens Aßmus, Ralf H. Bortfeldt and Gudrun A. Brockmann
Department for Crop and Animal Sciences, Humboldt-Universität zu Berlin, Invalidenstraße 42, 10115 Berlin, Germany

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Human single nucleotide polymorphism (SNP) chips which are used in genome-wide association studies (GWAS) permit the genotyping of up to 4 million SNPs simultaneously. To date, about 1000 human SNPs have been identified as statistically significantly associated with a disease or another trait of interest. The identified SNP is not necessarily the causal variant, but it is rather in linkage disequilibrium (LD) with it. CandiSNPer is a software tool that determines the LD region around a significant SNP from a GWAS. It provides a list with functional annotation and LD values for the SNPs found in the LD region. This list contains not only the SNPs for which genotyping data are available, but all SNPs with rs-IDs, thus increasing the likelihood to include the causal variant. Furthermore, plots showing the LD values are generated. CandiSNPer facilitates the preselection of candidate SNPs for causal variants.

Availability and Implementation: The CandiSNPer server is freely available at http://www2.hu-berlin.de/wikizbnutztier/software/CandiSNPer. The source code is available to academic users 'as is' upon request. The web site is implemented in Perl and R and runs on an Apache server. The Ensembl database is queried for SNP data via Perl APIs.

Contact: armin.schmitt@agrar.hu-berlin.de

Received on October 8, 2009; revised on February 4, 2010; accepted on February 16, 2010

1 INTRODUCTION

Whole-genome genotyping has become established in medical, biological and agricultural research to elucidate the genetic basis of phenotypic traits such as disease or economically important features. Typically, the result of a genome-wide association study (GWAS) consists of a list of single nucleotide polymorphisms (SNPs) with *P*-values that characterize the strength of association with the phenotype under investigation. Further analysis is necessary to associate a significant SNP with a gene or to decide if a significant SNP is a candidate for a causal variant for a disease or just in linkage disequilibrium (LD) with it. A causal variant is a genomic locus that has a direct qualitative or quantitative effect on the phenotype. SNPs can alter a protein directly (non-synonymous SNPs, stop gained or stop lost SNPs, frameshift SNPs or SNPs in splice sites) or they can influence gene expression if they are located in regulatory regions. SNAP (Johnson *et al.*, 2008) proved to be a valuable tool for the analysis of SNPs that are in LD with a significant SNP, but has the shortcoming that only SNPs that are present on SNP chips of the major genotyping platforms are considered.

The main motivation to develop CandiSNPer was that much more human SNPs are known than systematically genotyped within the HapMap project or included on the major genotyping platforms. These SNPs represent potential candidate SNPs for causal variants, although they will not be observed directly as significant SNP in a GWAS. The number of reference SNPs (rs-SNPs) is currently 17 804 034 (dbSNP build 130; genome build 36.3), of which 6 573 584 are validated. In the Caucasian population, 4 030 774 SNPs were genotyped by the HapMap project (Thorisson *et al.*, 2005) (release #27, February 2009) and 4 million SNPs can be typed, for example, with Illumina's HumanOmni1-Quad BeadChip. Thus, restricting the search for the causal variant of a phenotype to SNPs present on a genotyping chip or to HapMap SNPs would mean to neglect three-quarters of all known SNPs.

To simplify the maintenance of CandiSNPer and to guarantee up-to-dateness genotype and SNP data are retrieved from the Ensembl database during runtime. This is at the cost of a response time that is higher than if primary data were stored locally and LD was precomputed.

2 METHODS

For a given SNP (termed start-SNP) that was found significant in a GWAS, all HapMap-SNPs of a certain population in a flanking region of user-defined length (default: 300 kb up- and downstream) are fetched from the Ensembl database. r^2 and D' are calculated between these SNPs and the start-SNP. The first and last SNP on the flanking region with an r^2 (or D', respectively) value above a user-defined threshold (default: 0.2) are considered as start and end, respectively, of the LD region around the start-SNP. Next, all SNPs with rsIDs in the LD region are listed with basic annotation: functional class, position, r^2 and D' with the start-SNP, nearest gene and distance to it as well as the difference between observed and expected conservation score (Cooper *et al.*, 2005). Hyperlinks to the SNP and the gene are provided in a HTML list. A plot showing r^2 or D' for all HapMap SNPs in the flanking region versus the chromosomal position is generated for download in five formats. The positions of non-HapMap SNPs are indicated as ticks on the x-axis. The user can assign colors to the data points so that SNPs belonging to different functional classes can be distinguished. Runtime for one start-SNP is ~5 min for a set of 100 individuals and 3000 SNPs. The CandiSNPer web interface is programmed in HTML/Perl CGI (v 5.8) code, running on an Apache web server (v1.3.37). Graphics are created by R (v 2.6.2).

3 RESULTS AND DISCUSSION

3.1 Analysis of significant SNPs in intergenic regions

Intergenic SNPs are SNPs that are located at least 5 kb up- or downstream of a gene. In general, they are not associated with a

*To whom correspondence should be addressed.

Table 1. Occurrence of SNPs of selected functional classes in LD regions around intergenic SNPs

Functional class of SNP	No. of cases	Percentage
Non-synonymous coding	226	41.5
Splice site	166	30.5
Regulatory region	118	21.7
Stop lost	10	1.8
Stop gained	55	10.1
Frameshift	139	25.5
None of the above	303	55.6

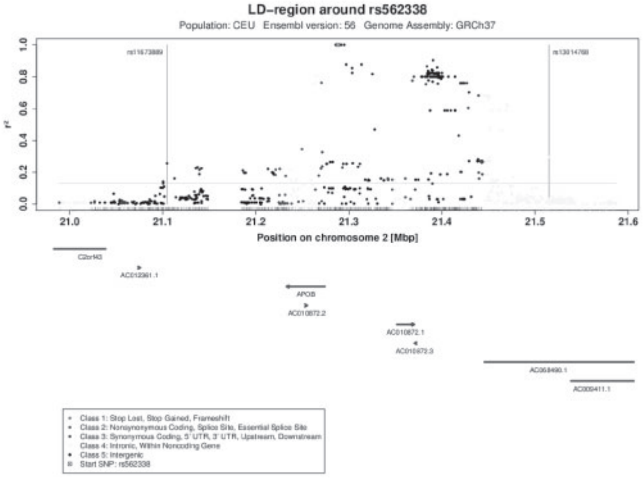

Fig. 1. r^2 values in the LD region around rs562338. Genes are shown as horizontal lines below the plot.

Table 2. Selection of candidate SNPs found in the vicinity of intergenic SNP rs562338 on human chromosome 2

SNP	Position	D'	Functional class	Gene	CS
rs72654415	21227545	NA	non-syn. cod.	*APOB*	0.76
rs72653079	21236405	NA	splice site	*APOB*	−3.30
rs62122485	21225819	NA	stop gained	*APOB*	0.10
rs1041962	21233202	NA	stop gained	*APOB*	2.50
rs35621726	21228575	NA	frameshift	*APOB*	1.70
rs35708286	21230838	NA	frameshift	*APOB*	−1.92
rs35068532	21235276	NA	frameshift	*APOB*	−3.56
rs12691202	21249716	0.91	non-syn. cod.	*APOB*	2.50
rs423836	21449098	0.91	non-cod. gene	*AC068490.1*	NA

If more than one functional class was attributed to an SNP only one is given. D' is given if genotypes were available, else they were denoted NA. CS, conservation score; non.-syn. cod., non-synonymous coding. Some variables are not shown because of limited space.

gene and not located in a known regulatory region. We applied CandiSNPer to all 545 intergenic SNPs which were at present found in the Caucasian population significantly associated with a phenotype (Hindorff *et al.*, 2009). Table 1 lists the number of cases in which LD regions around these SNPs host SNPs of functional classes that are deemed as strong candidates for a causal variant. In almost half of the cases at least one candidate SNP is located in the LD region around a significant intergenic SNP. In more than half of the cases, however, no such candidate SNP could be identified. This suggests that the identification of causal variants will remain a hard problem.

3.2 Example: analysis of an intergenic SNPs associated with LDL cholesterol

As an example, we studied all SNPs that are in the LD region of rs562338, an intergenic SNP ~21 kb downstream of *APOB* on chromosome 2 that was found to be associated with LDL cholesterol (Sandhu *et al.*, 2008). Figure 1 depicts the r^2 values between this SNP and SNPs in its vicinity. The LD region extends between ~21.1 and ~21.5 Mb (perpendicular lines). The horizontal line shows the average of r^2 values for scrambled genotypes plus 3 standard deviations ($\langle r^2 \rangle = 0.018$, SD = 0.025; $\langle D' \rangle = 0.12$, SD = 0.09). We extracted all reference SNPs in the LD region around this SNP and identified candidate SNPs a selection of which is listed in Table 2. They can have an effect on the functioning of the genes they are located in. The frameshift and stop gained SNPs are even likely to have a severe effect on the 3D conformation of the protein that *APOB* is coding for.

ACKNOWLEDGEMENTS

We thank A. Scherag (University of Duisburg-Essen, Germany) for fruitful discussion and the EBI (Hinxton, England) for the kind permission to query the Ensembl database during CandiSNPer runs.

Funding: German Federal Ministry of Education and Research (NGFNPlus 01GS0829).

Conflict of Interest: none declared.

REFERENCES

Cooper,G.M. *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Johnson,A.D. *et al.* (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, **24**, 2938–2939.

Sandhu,M.S. *et al.* (2008) LDL-cholesterol concentrations: a genome-wide association study. *Lancet*, **371**, 483–491.

Thorisson,G.A. *et al.* (2005) The International HapMap Project Web site. *Genome Res.*, **15**, 1592–1593.