SNPtoGO: characterizing SNPs by enriched GO terms

Daniel F. Schwarz¹,* , Oliver Hädicke¹,⁴, Jeanette Erdmann³, Andreas Ziegler¹, Daniel Bayer² and Steffen Möller²,*

¹Institut für Medizinische Biometrie und Statistik, ²Institut für Neuro- und Bioinformatik, ³Medizinische Klinik II, Universität zu Lübeck, Ratzeburger Allee 160, 23538 Lübeck and ⁴Current Address: Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstrasse 1, 39106 Magdeburg, Germany

Received on May 31, 2007; revised on October 12, 2007; accepted on October 31, 2007

Advance Access publication November 17, 2007

Associate Editor: John Quackenbush

ABSTRACT
For the analysis of complex polygenic diseases, one does not expect all patients to share the same disease-associated alleles. Not even will disease-causing variations be assigned to the identical sets of genes between patients. However, one does expect overlaps in the sets of genes that are involved and even more so in their assigned molecular processes. Furthermore, the assignment of single nucleotide polymorphisms (SNPs) to genes is highly ambiguous for intergenic SNPs. The tool presented here hence adds external information, i.e. GeneOntology (GO) terms (Gene Ontology Consortium), to the analysis of SNP data.
Availability: A web interface and source code are offered at https://webtools.imbs.uni-luebeck.de/snptogo
Contact: schwarz@imbs.uni-luebeck.de

1 INTRODUCTION
Full genome SNP chip experiments comprise hundreds of thousands of features. The challenge is to reduce the number of explanatory variables without losing relevant biological information. For the analyses of SNP data, an abstraction towards haplotype blocks or regions of little recombination is evident. The integration of external information in that process, e.g. allows for a filtering by the difference of allele frequencies in populations external to the study hand (Möller et al., 2004). By characterizing dominant features shared by multiple SNPs, the number of features is reduced and the results may become statistically more powerful.

Equivalent difficulties affect the analysis of gene expression data. This field has strongly advanced in embracing external data for the data analysis, a consequence from the direct availability of an avalanche of manually curated and automatically deduced annotations (The UniProt Consortium, 2007). Such approaches comprise the inspection of molecular pathways (Chung et al., 2004; Mlecnik et al., 2005), and, as it is the focus of this work, tools for the analysis of the enrichment (Gentleman, 2004; Wrobel et al., 2005) of GO terms (Gene Ontology Consortium, 2006).

The approach presented here extends the prior towards the analysis of single nucleotide polymorphisms (SNPs). The GO terms of a SNP are the same as those of the gene that has the most proximal chromosomal location. This assignment is ambiguous because of genes overlapping on chromosomes and because of SNPs being located between genes. The analysis of a set of SNPs will determine subsets of SNPs that are each dominated by a respective molecular process—as represented by an entry in GO.

2 APPROACH
This work presents a web interface to analyse the distribution of GO terms that are associated with a set of SNPs. The assignment is performed according to the Ensembl database (Hubbard et al., 2007) with a user-specified maximal distance between SNP and gene to include intergenic SNPs in the analysis. The maximal distance can be set by user in the submission mask. An option was added to restrain the acceptance of neighbouring genes to intergenic SNPs only. A SNP could be assigned to several genes due to overlapping genes. Thus, multiple GO terms could be associated with the same SNP. GO terms that are found overrepresented are reported both graphically and in a table. To constrain the search, a minimal distance from the root can be specified. Also, with prior knowledge that, e.g. adhesion is of concern in inflammatory processes (Gierer et al., 2005), respective negative lists can be given.

A GO term’s dominance is characterized by the ratio of the number of observed appearances in a particular set of selected SNPs versus the number of expected appearances for a random selection. Statistically, the problem is addressed with the Fisher’s exact test. For this implementation, the principles of the elim approach described for the topGO tool (Alexa et al., 2006) are directly applied. The elim algorithm (Alexa et al., 2006) uses the tree structure of GO for a top-down hierarchical selection of significant GO terms. A naïve approach that assumes the independence of the assigned GO terms would indicate too many statistically relevant GO terms that share a parent term that is already statistically significant.

The significance level for each Fisher’s exact test is set to α/(number of GO terms selected), the Bonferroni correction for multiple testing. That number is calculated in advance as the
directly assigned terms plus their minimal path in GO from
the root.

Genes with many historically well characterised SNPs may
appear with more probes on a SNP chip than others. This bias
is taken into account by the statistics. Nevertheless, the tool
offers a gene-based approach that for each GO term counts
only the number of different genes assigned to it, not the
number of SNPs.

The calculation is performed in multiple stages as shown in
Figure 1. The SNPs of interest are submitted in an entry mask
and parameters for maximal distances for intergenic SNPs set.
Results are presented as an HTML table with hyperlinks for
GO Terms via AmiGO (Gene Ontology Consortium, 2006), to
SNPs via Ensembl and dbSNP (Wheeler et al., 2007), and to
genes also via Ensembl. For each GO term, the calculated
P-values and test scores are presented aside absolute numbers
for each term indicating the dominance. The terms can be
sorted by any of these numbers. The user has additional
information thanks to the list of SNPs that was assigned to the
GO term and the gene that established this link.

In a second stage, the degree of overexpression is expressed
graphically. Presented in a heatmap, the x-axis represents the
SNPs, the y-axis the GO terms. The intensity represents the
P-value. Either axis has its entries reordered to cluster similar
SNPs and GO terms.

3 METHODS

The MySQL database of Ensembl version 46 was mirrored on a local
Debian Linux server. An assignment of SNPs to GO terms was
prepared as a separate table to lower response times. All calculations,
graphics and HTML generation are performed by an R script
(R Development Core Team, 2006) with support of the libraries
RMySQL (James and DebRoy, 2006), CGIwithR (Firth, 2003) and
gplots (Warnes et al., 2007).

4 DISCUSSION

This application is the first to apply principles of the analysis
of GO terms to SNP data. The key difference between tools
previously developed for gene expression data and this
approach is in the treatment of ambiguities in the assignment
of SNPs to genes.

A major concern for the analysis are the intergenic SNPs.
These are compensated for optionally including neighbouring
genes within a specified distance of, e.g. 100 kb. A SNP close to
enhancer regions may then (amongst others) be assigned to the
gene it controls (Blackwood and Kadonaga, 1998). One may
argue that it is not unlikely to find interacting genes
chromosomally neighboured (Wang et al., 2004), thus con-
tributing to a reduced error by the otherwise included unrelated
biological processes.

Intuitively, the statistically most associated SNPs of a
polygenic disease are likely to affect different biological
processes. If they were on the same, then these could not be
compensated. One may speculate that the number of SNPs
needed until GO terms can be identified may be a measure for
the polygenicity of the disease investigated.

The approach presented here is ignorant of the linkage
disequilibrium between SNPs or other information in the raw
data like the copy number variation. Also, the comparison of
SNP data with gene expression data of the same individuals
may yield additional insights for a selection of genes and
associated SNPs. Such investigations were not addressed
because of the huge amount of data that would be required
to be transferred for a complete service. Users are instead suggested to submit sets of linked SNPs separately. To support an automation of that process by in-house systems, all source code is made available.

Both the data from SNP chips and gene expression microarrays may be analysed in conjunction with GO. SNP data has the advantage to directly indicate the chromosomal location of putative causes for a genetic disease and its cofactors. Furthermore, a variant is detected independently from the tissue that is analysed. The challenge is to combine both types of data in the analysis.

5 CONCLUSION
The tool may be particularly beneficial for the analysis of defects in polygenic diseases, making use of the redundancy of defects in metabolic or regulatory pathways.

ACKNOWLEDGEMENTS
This work was funded by the DFG (K02250/3–1) and the EU projects KnowARC (032691) and Cardiogenics (037593). The authors thank Thomas Martinetz, Silke Szymczak and the anonymous reviewers for their support and critical reading of the manuscript.

Conflict of Interest: none declared.

REFERENCES


