

TAMAL: An Integrated Approach to Choosing SNPs for Genetic Studies of Human Complex Traits

Bradley M. Hemminger¹, Billy Saelim¹ and Patrick F. Sullivan^{2,3*}

¹School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

²Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill NC, USA.

³Department of Medical Epidemiology & Biostatistics, Karolinska Institutet, Stockholm, Sweden

Associate Editor: Frank Dudbridge

Summary: Investigators conducting studies of the molecular genetics of complex traits in humans often need rationally to select a set of single nucleotide polymorphisms (SNPs) from the hundreds or thousands available for a candidate gene. Accomplishing this requires integration of genomic data from distributed databases and is both time consuming and error-prone. We developed the TAMAL web site to help identify promising SNPs for further investigation. For a given list of genes, TAMAL identifies SNPs that meet user-specified criteria (e.g., haplotype tagging SNPs or SNP predicted to lead to amino acid changes) from current versions of online resources (i.e., HapMap, Perlegen, Affymetrix, dbSNP, and the UCSC genome browser).

Availability: TAMAL is a platform independent web-based application available free of charge at <http://neoref.ils.unc.edu/tamal/>.

Contact: Patrick Sullivan (pfsulliv@med.unc.edu)

Supplementary Information: <http://neoref.ils.unc.edu/tamal/>

1 INTRODUCTION

Investigators conducting studies of the molecular genetics of complex traits in humans often need rationally to select a set of single nucleotide polymorphisms (SNPs) from the hundreds or thousands available for a candidate gene. For example, for a study of the genetics of type 2 diabetes mellitus, alcoholism, or schizophrenia, an investigator may wish comprehensively to genotype SNP markers in dozens or even hundreds of candidate genes. With the completion of the initial sequencing of the human genome (Lander, Linton et al. 2001) and the considerable progress afforded by the International HapMap project (The International HapMap Consortium 2003; Altshuler, Brooks et al. 2005), many genes contain more SNPs than can be affordably genotyped. For example, the neuregulin-1 gene contains around 4,000 SNPs, more than is practically feasible to genotype (even as genotyping costs continue to plummet). Our application provides a rational methodology for reducing the number of SNPs to evaluate while still capturing directly or indirectly a considerable portion of the genetic variation found in the genomic region.

Accomplishing this task for a set of dozens or hundreds of genes is currently time consuming and error-prone as the integration of

genomic data from disparate databases is required. We developed the TAMAL (Technology And Money Are Limiting) web-based application to help streamline the task of choosing SNPs for further investigation (see Figure 1 for a screenshot of the TAMAL application).

2 METHODS

TAMAL is designed to be interactive, so that in addition to displaying suggested SNPs, the researcher can dynamically filter the results based on any of the application's controls. On the left panel in Figure 1, the user inputs the standard gene name for a single gene or uploads a list of genes. The standard gene name is generally that approved by HUGO (<http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl>), e.g. *COMT* for catechol-O-methyltransferase. All genomic locations are per the hg16 UCSC build.

The middle panel shows the result of querying the TAMAL database for the gene(s) input by the user. Optionally, the user can also limit the search to the most 5' and 3' extent of the gene or extend the search by a specified number of bases in either direction (20,000 bases by default). The SNP set is limited to those with evidence of variation in any of the major SNP databases (dbSNP, HapMap, Perlegen, and Affymetrix).

The right panel lists sets of criteria that can be used to filter the set of SNPs according to flexible criteria. At the top, the user can select the Gabriel (Gabriel, Schaffner et al. 2002) or TAGGER method of selecting haplotype tag SNPs from any or all of the four HapMap ancestry groups (The International HapMap Consortium 2003) as determined by HaploView (Barrett, Fry et al. 2004). It is important to note that some genomic regions may not be amenable to this approach (Wall and Pritchard 2003; Wall and Pritchard 2003). At the middle of the right panel, the user can select SNPs that lead to non-synonymous or synonymous amino acid changes augmented with *in silico* prediction of functionality (Karchin, Diekhans et al. 2005) or alter an intronic splice site. At the bottom, the user can select SNPs that occur in certain types of genomic features – SNPs that are: in a predicted promoter (*in silico* prediction but with biological validation) (Trinklein, Aldred et al. 2003), in a region of predicted regulatory potential (Blanchette, Kent et al. 2004), or a predicted transfactor binding site (TRANSFAC v6.0, <http://www.gene-regulation.com>), along with SNPs that are in regions with conservation scores $\geq 99^{\text{th}}$ percentile genomewide

The screenshot displays the NeoRef Open Archive Search interface. The main search area shows a search for the gene *COMT*, resulting in 123 total hits and 45 filtered hits. The search criteria include a 36% filter and an extension of 2000 bp in 5' and 3' directions. The results are presented in a table with columns for rname, chrom, Length, Notes, All Known SNPs, SNPs to Genotype, and Genome Browser. The filtered results for *COMT* on chromosome 22 are shown, with 31,222 bp length, 'Mapping okay' notes, 123 SNPs, and 45 SNPs to Genotype. Below the table are buttons for 'Download XML', 'Download Excel', and 'Reset'. To the right, there are filter panels for 'htSNP', 'Coding', and 'Other' categories, each with a table of attributes and counts. The 'htSNP' filter includes options for 'Gabriel Method' and 'Tagger Method', and lists panels like 'Caucasian Panel' (8), 'Chinese Panel' (12), 'Japanese Panel' (11), and 'Yoruban Panel' (16). The 'Coding' filter includes options for 'Lead to Synonymous Mutation' (7), 'Lead to Nonsynonymous Mutation' (6), 'Alter Splice Site' (0), and 'In Intron' (9). The 'Other' filter includes options for 'Predicted Promoter' (12), 'In Region of Predicted Regulatory Potential' (0), 'In Predicted Transfactor Binding Site' (0), and 'Conservation Score above 99th Percentile' (9). An inset window shows the UCSC Genome Browser visualization for the *COMT* gene on Human July 2003 Assembly, displaying various genomic tracks including SNPs, repeats, and conservation scores.

Fig. 1. TAMAL screenshot, showing the result of the user querying with input of a single gene, *COMT*. Inset into the bottom middle is the UCSC browser visualization for this result (normally this would appear as a pop up window on top of the TAMAL window).

for human-chimp-rat-mouse-chicken alignment via a hidden Markov model (Siepel and Haussler 2003).

The user can inspect the choice of SNPs by clicking on the down arrow next to a gene in the middle panel. This opens the UCSC genome browser in a separate window (see insert in Figure 1) so the user can inspect the SNP coverage and ensure that the SNPs selected are a reasonable subset of all those potentially available. Finally, at the lower edge of the middle panel users can download the results into an EXCEL file (commonly used by researchers) or in XML format (for exchange with other applications).

TAMAL is provided as a good faith effort to assist the human genetics community. No such tool should be considered as a foolproof “black box”. There are some genes that will be difficult to study with typical SNP methods, and there are additional databases for some genes that should be consulted (e.g., for genotyping members of the large CYP gene family). Nonetheless, provided that users remain cognizant of its limitations, TAMAL can greatly assist with relational SNP selection.

We will endeavor to update TAMAL on a quarterly basis to incorporate updates to the primary databases as well as new features.

ACKNOWLEDGEMENTS

We thank the Carolina Center for Exploratory Genetic Analysis for computational support (P20RR20751), and the Informatics and Visualization Laboratory (<http://www.ils.unc.edu/bmh/ivlab>) at the School of Information and Library Science for hosting this service.

REFERENCES

Altshuler, D., L. D. Brooks, A. Chakravarti, F. S. Collins, M. J. Daly and P. Donnelly (2005). "A haplotype map of the human genome." *Nature* 437(7063): 1299-320.

Barrett, J. C., B. Fry, J. Maller and M. J. Daly (2004). "Haploview: analysis and visualization of LD and haplotype maps." *Bioinformatics*.

Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler and W. Miller (2004). "Aligning multiple genomic sequences with the threaded blockset aligner." *Genome Res* 14(4): 708-15.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly and D. Altshuler (2002). "The structure of haplotype blocks in the human genome." *Science* 296(5576): 2225-9.

Karchin, R., M. Diekhans, L. Kelly, D. J. Thomas, U. Pieper, N. Eswar, D. Haussler and A. Sali (2005). "LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources." *Bioinformatics*.

Lander, E. S., et al. (2001). "Initial sequencing and analysis of the human genome." *Nature* 409: 860-921.

Siepel, A. and D. Haussler (2003). Combining phylogenetic and hidden Markov models in biosequence analysis. Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB 2003): 277-286.

The International HapMap Consortium (2003). "The International HapMap Project." *Nature* 426(6968): 789-96.

Trinklin, N. D., S. J. Aldred, A. J. Saldanha and R. M. Myers (2003). "Identification and functional analysis of human transcriptional promoters." *Genome Res* 13(2): 308-12.

Wall, J. D. and J. K. Pritchard (2003). "Assessing the performance of the haplotype block model of linkage disequilibrium." *Am J Hum Genet* 73(3): 502-15.

Wall, J. D. and J. K. Pritchard (2003). "Haplotype blocks and linkage disequilibrium in the human genome." *Nat Rev Genet* 4(8): 587-97.