

SNP special interest group



SNP-SIG Meeting

**Identification and annotation of SNPs
in the context of structure, function, and
disease.**

ISMB 2012
July 14th 2012, Long Beach (California), USA

<http://snps.uib.es/snp-sig>



Invited Speakers



Russ Altman

Stanford University, Palo Alto (CA), USA

Defining levels of evidence supporting clinically actionable SNP associations.



Steven Brenner

University of California, Berkeley (CA), USA

CAGI Experiments.



David Haussler

University of California, Santa Cruz (CA), USA

Somatic mutations in cancer as assessed by whole genome sequencing.



Olivier Lichtarge

Baylor College of Medicine, Houston (TX), USA

Missense Mutations in Action: an Analytic View.



Zemin Zhang

Genentech, San Francisco (CA), USA

Identification and computational analysis of somatic mutations in cancer samples.

SNP-SIG Organizers

Yana Bromberg, Rutgers University, New Brunswick (NJ), USA

Emidio Capriotti, University of Balearic Islands, Palma de Mallorca, Spain

Round table Discussion

Andre Franke, Christian-Albrechts-University, Kiel, Germany

Rachel Karchin, John Hopkins University, Baltimore (MD), USA

Sean Mooney, Buck Institute, Novato (CA), USA

Shamil Sunyaev, Harvard Medical School, Boston (MA), USA

SNP-SIG Meeting Programme - July 14th 2012, Long Beach (CA), USA

08:20 – 08:30 Welcome from the committee

Session 1: Annotation and prediction of structural/functional impacts of coding SNPs

- 08:30 – 09:20 **Highlight Speaker: Zemin Zhang**, Genentech, San Francisco (USA)
Identification and computational analysis of somatic mutations in cancer samples.
- 09:20 – 09:45 **Gustavo Parisi**, Universidad Nacional de Quilmes, Buenos Aires (Argentina)
Improving the prediction of disease-related variants using protein dynamism.
- 09:45 – 10:10 **Maricel Kann**, University of Maryland, Baltimore (USA)
Protein Domain-Centric Approach to Study of Cancer Somatic Mutations.
- 10:10 – 10:30 **Coffee Break**
- 10:30 – 10:55 **Lei Xie**, Hunter College, New York (USA)
Multiscale Modeling of Causal Phenotypic Associations of nsSNPs: Application to Hypoxia
- 10:55 – 11:20 **Nouf Alnumair**, University College London (UK)
The SAAP database and other tools to analyze the impact of mutations and predict the pathogenicity of a novel mutation.
- 11:20 – 11:45 **Frank Schacherer**, BIOBASE GmbH.
Predicting the disease potential of gene mutations with MutationTaster
- 11:45 – 12:25 **Keynote: Olivier Lichtarge**, Baylor College of Medicine, Houston (USA)
Missense Mutations in Action: an Analytic View.
- 12:25 – 12:40 **Company Presentation: Frank Schacherer**, BIOBASE GmbH.
Manually curated databases for SNP analysis.
- 12:40 – 13:20 **Lunch Break and Poster Session with the Authors**

SII: SNPs as effectors of change: disease and evolution

- 13:20 – 14:10 **Highlight Speaker: David Haussler**, University of California, Santa Cruz (USA)
Somatic mutations in cancer as assessed by whole genome sequencing.
- 14:10 – 14:35 **Boris Reva**, Memorial Sloan-Kettering Cancer Center, New York (USA)
Revealing selection in cancer using the predicted functional impact of cancer mutations. Application to nomination of cancer drivers
- 14:35 – 15:00 **Janita Thusberg**, Buck Institute for Research on Aging, Novato (USA).
Predicting Pharmacogenetic Protein Variants.
- 15:00 – 15:25 **Adam Kowalczyk**, University of Melbourne (Australia).
Detection and analysis of interactions in multiple GWAS studies using standard desktop PC.
- 15:25 – 15:45 **Coffee Break**
- 15:45 – 16:10 **Yana Bromberg**, Rutgers University, New Brunswick (USA)
Noise in Biology: defining the neutral range of protein function.
- 16:10 – 16:50 **Keynote: Russ Altman**, Stanford University, Palo Alto (USA).
Defining levels of evidence supporting clinically actionable SNP associations
- 16:50 – 17:20 **CAGI Report: Steven Brenner**, University of California, Berkeley (USA).
CAGI Experiments
- 17:20 – 18:10 **Round Table Discussion**
- 18:10 – 18:20 Closing remarks from the committee

Selected Presentations

SNP-SIG Meeting – ISMB 2012, July 14th Long Beach (CA), USA

THE SAAP DATABASE AND OTHER TOOLS TO ANALYZE THE IMPACT OF MUTATIONS AND PREDICT THE PATHOGENICITY OF A NOVEL MUTATION

Nouf S. A. Alnumair^{*} and Andrew C. R. Martin^{*}

^{*}University College of London, UK

email: nouf.alnumair.09@ucl.ac.uk,

andrew.martin@ucl.ac.uk

The Single Amino Acid Polymorphism data analysis pipeline (SAAPdap) is a pipeline for the analysis and visualization of the structural effects of mutations coupled with a predictor of whether mutations will be damaging. SAAPdap has been built along the same principles as SAAPdb, <http://www.bioinf.org.uk/saap/db/> our database of effects of mutations identified by mapping single nucleotide polymorphisms (SNPs, from dbSNP) and pathogenic deviations (PDs, from OMIM and several locus-specific mutation databases), to protein structural data from the Protein Data Bank. Both SAAPdap and SAAPdb perform an automated analysis of likely local structural effects that may disrupt protein folding, function or stability and therefore may be related to a harmful phenotype. SAAPdb used fixed thresholds for defining potentially damaging effects for all analyses. However assigning a mutation as (not) having a structural effect has been shown to be very sensitive to precise structural details. In SAAPdap, we provide continuous (rather than Boolean) values for each of the analyses. For example, in analysing clashes, rather than defining a damaging clash as any sidechain that has at least 3 van der Waals overlaps with other atoms, we now perform a more complete energy calculation incorporating Lennard-Jones and torsion energies. Another example is that analysis of mutations from-glycine and to-proline (the 'structural' amino acids which show an unusual Ramachandran distribution); we have moved from a simple set of allowed boundaries to a pseudo-energy potential based on the Ramachandran plot. Analysis of the data in SAAPdb shows clear differences in the sequence and structural characteristics of SNPs and PDs: as might be expected, PDs have additional, and more severe, structural effects. This indicates that there is a clear signal in the data that can be used to predict the pathogenicity of a novel mutation. This has been exploited using a Random Forest predictor and initial results out-perform any other available method. The work presented here includes an update of the data, initial results from the improved analysis of clashes, from-glycine and to-proline mutations and performance of the initial pathogenicity-predicting model.

NEUTRALITY DILEMMA: WHEN IS A MUTATION REALLY A FUNCTIONAL POLYMORPHISM?

Yana Bromberg^{*}

^{*}Rutgers University (NJ), USA

email: yanab@rci.rutgers.edu

Scientists have been trying to establish the genetic basis of disease ever since the first forays into genome sequencing. Non-synonymous single nucleotide polymorphisms (nsSNPs; mutations resulting in a single amino acid substitution in the encoded protein sequence) have rightfully been a focus of many studies due to their high level of involvement with disease. Numerous publications discuss experimental annotations of mutants associated with disease. Meta-resources, including the Protein Mutant Database (PMD) and Swiss-Prot, employ manual curation to isolate the necessary data from these publications. Additionally, natural language processing (NLP) methods aim to fill the speed-of-processing gap between rates of manual curation and speed of appearance of new publications. Regardless of the advantages and limitations of the various modes of literature data extraction, the experimental data itself suffers from the lack of specificity. Proving that a mutation has an effect on protein function (non-neutral mutation) only requires one set of the successful (disruptive) experiments. On the other hand, showing that a mutation has no effect (neutral mutation, identical to wild type) requires probing all possible ways that all functions of a given protein may have been disrupted. Computational predictions of mutation effects are cheaper, less time-consuming, and often more generic (measuring all functions of a protein at once) than experimental evaluations. However, predictions of neutrality often disagree with experimental observations. Do the computational methods overpredict damaging effects or is the tolerated range of protein function much wider than we think? Recent developments in high-throughput sequencing have finally generated enough data to consider these questions in more detail.

IMPROVING THE PREDICTION OF DISEASE-RELATED VARIANTS USING PROTEIN DYNAMISM

Emidio Capriotti, Maria Silvina Fornasari, Juritz Ezequiel, Pier Luigi Martelli, Piero Fariselli, Rita Casadio and Gustavo Parisi

**Universidad Nacional de Quilmes, Argentina
email: gustavo@unq.edu.ar*

Single Nucleotide Polymorphisms (SNPs) are the major source to study human genetic variability. In particular, non-synonymous coding SNPs (also called single amino acid polymorphism or SAPs) could be associated with alteration in protein function and the occurrence of diseases. Several methods have been developed to predict the effect of SAPs. Most of these methods are based on sequence and structural information of proteins. As protein function highly relies on protein dynamism, in this work we explored how the use of sequential, structural and protein dynamical information contributes to the prediction of disease-related SAPs. To this end we used a set of proteins with different extent of conformational diversity derived from PCDB database, a conformer database. These proteins are associated with 803 SAPs where 482 are disease related. Each protein is then represented by a set of different conformers and the effect of each mutation was studied in each of them. We found that a joint prediction taking into account all the conformers outperforms the prediction made using a single structure for each of the studied proteins. Our results suggest that the consideration of conformational diversity can improve the discrimination of neutral and disease related protein SAPs.

DETECTION AND ANALYSIS OF INTERACTIONS IN MULTIPLE GWAS STUDIES USING STANDARD DESKTOP PC

Benjamin Goudey^{*}, Qiao Wang, Fan Shi, Linda Stern, Geoff Macintyre, Cheng Soon Ong, Michael Inouye and Adam Kowalczyk^{*}

**The University of Melbourne, Australia
email: bwgoudey@gmail.com,
Adam.kowalczyk@nicta.com.au*

It has been hypothesised that genetic interactions (i.e. epistasis) may resolve the problem of “missing heritability” currently observed in genome-wide association studies (GWAS). Methods developed for their detection are typically focusing on SNPs, and are benchmarked on synthetic data for reasons of both scalability and interpretation. However, benefits from evaluations of this type are restricted by our limited understanding of the mechanisms of interactions between observable SNP markers, especially when the observed SNPs are surrogates for causal mutations in linkage disequilibrium. Systematic analysis of interactions in real life data should bridge this knowledge gap. Here, we apply a novel and fast methodology, named Genome-Wide Interaction Search (GWIS), for exhaustively evaluating SNP pairs. It can search an entire Wellcome Trust Case Control Consortium (WTCCC) dataset in less than 3 hours for a CPU and below 30 minutes for a GPU implementation on a standard desktop PC. This is significantly faster than comparable algorithms, facilitates more systematic analysis and has potential for exhaustive search for three-way SNP interactions. We present sample results from application of our algorithm to seven WTCCC datasets. All datasets clearly show SNP pairs that have considerably stronger association with disease than their individual component SNPs. In a separate experiment, running our algorithm over five independent Celiac disease datasets, we have detected the same pairs of SNP-markers replicating across independent cohorts and ethnicities. Many of these pairs show non-standard patterns of interaction, not compatible with typical models used for generation of synthetic benchmark data.

PROTEIN DOMAIN-CENTRIC APPROACH TO STUDY OF CANCER SOMATIC MUTATIONS

Thomas Peterson and Maricel Kann*

**University of Maryland, Baltimore County, USA
email: mkann@umbc.edu*

The fight against cancer has been hindered by its highly heterogeneous nature. Recent genome-wide sequencing studies have shown that individual malignancies contain many mutations that range from those commonly found in tumor genomes to rare cancer somatic mutations present only in a small fraction of lesions. For instance, the genome of a colorectal cancer in one patient can have somewhere between 50 to 100 somatic mutations, but might share only 2 or 3 mutated genes with colorectal tumor genomes from other patients. Somatic mutations that are frequently found in tumor genomes often play a significant role in tumor development and are thus classified as cancer driver mutations. However, efforts to correlate somatic mutations found in one or few individual tumor genomes with critical functional roles in tumor development have so far been unsuccessful. In this paper, we analyze cancer somatic mutations from lung cancer patients using a new approach based on aggregation of mutational data at the protein domain level. Our preliminary analysis confirms that our approach creates a framework for leveraging structural genomics and evolution into the analysis of somatic cancer mutations. We found that by incorporating information about classification of proteins and protein sites we are able to detect novel clusters of lung cancer somatic mutations.

REVEALING SELECTION IN CANCER USING THE PREDICTED FUNCTIONAL IMPACT OF CANCER MUTATIONS. APPLICATION TO NOMINATION OF CANCER DRIVERS.

Boris Reva* and Yevgeniy Antipin

**Memorial Sloan-Kettering Cancer Center, USA
email: reva@cbio.mskcc.org*

Every malignant tumor has a unique spectrum of genomic alterations including numerous protein mutations. There are also hundreds of personal germline variants to be taken into account. The combinatorial diversity of potential cancer-driving events limits the applicability of statistical methods to determine tumor-specific “driver” alterations among an overwhelming majority of “passengers”. An alternative approach to determining driver mutations is to assess the functional impact of mutations in a given tumor and predict drivers based on a numerical value of the mutation impact in a particular context of genomic alterations. Recently, we introduced a functional impact (FI) score, which assesses the mutation impact by the value of entropic disordering of the evolutionary conservation patterns in proteins. The FI score separates disease-associated variants from benign polymorphisms with an accuracy of ~80%. Can the FI score be used to identify functionally important non-recurrent cancer-driver mutations? Assuming that cancer-drivers are positively selected in tumor evolution, we investigated how the FI score correlates with key features of natural selection in cancer, such as the non-uniformity of distribution of mutations, the frequency of affected tumor suppressors and oncogenes, the frequency of other concurrent genomic alterations; as a control, we used presumably non-selected silent mutations. Using mutations of 6 cancers (TCGA), we found that predicted high-scoring functional mutations tend to be evolutionarily selected as compared to low-scoring and silent mutations. This result justifies prediction of mutations-drivers using ~ a 10 times shorter list of predicted high-scoring functional mutations, rather than the “long tail” of all mutations.

PREDICTING THE DISEASE POTENTIAL OF GENE MUTATIONS WITH MUTATIONTASTER

Jana Marie Schwarz, Dominik Seelow
and Frank Schacherer*

**BIOBASE International, Germany*
email: frank.schacherer@biobase-international.com

The evaluation of the disease potential of DNA alterations by wet lab means is time and cost intensive. Especially when it comes to Next Generation Sequencing projects and thousands of variations have to be tested for their possible effect, in silico approaches are inevitable.

We hence developed MutationTaster, a web-based software for disease potential prediction. It has already become a widely used prediction tool with up to 1,000,000 queries by external groups each day. It performs various tests both on protein and DNA level which are then scored by a Bayes classifier. In contrast to similar tools, MutationTaster is not limited to single amino acid substitutions. Besides, it has a higher performance and is much faster. Results from Next Generation Sequencing projects can be submitted to MutationTaster in batch query mode. We offer different converting tools to streamline the analysis, moreover, a dedicated query engine for automated analysis of NGS data from VCF file(s) will very soon be available.

Since its official release in 2009, the protein and genetic data used by MutationTaster was updated several times and a better splicing model was developed. MutationTaster now integrates the 1000 Genomes genotypes to filter out known polymorphisms. The latest improvement makes use of disease mutations from the Human Gene Mutation Database (HGMD) to train the Bayes classifier.

PREDICTING PHARMACOGENETIC PROTEIN VARIANTS

Chet Seligman, Janita Thusberg*, Emidio Capriotti, Biao Li, Jackson Miller, Jim Auer, Michelle Whirl-Carrillo, Teri Klein and Sean Mooney

**Buck Institute for Research on Aging*
email: jthusberg@buckinstitute.org

Little is known about the nature of pharmacogenetic (PGx) variants as compared to disease-causing genetic variants and neutral polymorphisms. Similarly to disease-causing variants, the pharmacogenetic SNPs may become invaluable for personalized medicine. We are employing bioinformatic methods, to annotate the protein level consequences of pharmacodynamic (PD) and pharmacokinetic (PK) variants in the PharmGKB® database with the goal of predicting candidate PGx variants. We are working towards a pharmacogenetic fingerprint of features that describe protein variants. Our research involves two separate strategies to separate PGx from neutral and PD variants from PK. The first involves gene and protein attributes determined from GO, KEGG, Reactome, HRPD and co-expression networks where known PGx entities are substantially enriched relative to the human genome. The second strategy operates at the variant level wherein we employ a machine-learning classifier (Random Forest) to known PGx variants, in order to evaluate features that will differentiate these from neutral SNPs or disease-causing mutations. Our dataset includes 143 known PGx variants and over 32,000 known and presumed neutral SNPs. The data imbalance problem was addressed by repetitive undersampling the most abundant class, and also by generating a separate negative dataset of SNPs occurring in the proteins harboring PGx SNPs. We will summarize our findings and describe how a PGx variant compares to the molecular attributes that describe other variants, both disease-associated and neutral. Also, a novel method for the prediction of candidate PGx variants is presented.

**MULTISCALE MODELING OF CAUSAL
PHENOTYPIC ASSOCIATIONS OF nsSNPs:
APPLICATION TO HYPOXIA**

Lei Xie^{*}, Philip Bourne, Li Xie, Dan Zhou, Gabriel Haddad, Raoul Valencia, Barbara Ferreira, Thahmina Ali, Vincent Xue and Maliha Tanweer.

^{*}*Hunter College, CUNY (NY), USA
email: lxie@sdsc.edu*

It is a major challenge to reveal the functional roles of nsSNPs on complex phenotypes. Statistical techniques establish correlations between genotypes and phenotypes but may fail to infer the biologically-relevant causal relationships. The emerging paradigm of Network-based Association Studies aims to address this problem of statistical analysis. However, a mechanistic understanding of how individual molecular components work together in a system requires knowledge of molecular structures, and their interactions. To address these challenges, we have developed a structural systems biology approach to modeling nsSNPs at multi-scales - from atomic details of molecular interactions to emergent properties of biological networks. We apply our approach to reveal the functional roles of nsSNPs associated with hypoxia tolerance in *Drosophila melanogaster*. Three major findings have emerged from our preliminary studies. First, a large fraction of the driver mutations are not located in functional sites, but rather regulate protein activity through allosteric transitions. Second, nsSNPs in two genes, *DYS* and *CalNac-T2*, are identified to be responsible for the up-regulation of the Notch signaling pathway that is critical to hypoxia tolerance. Last, the cross-talk between the heat shock regulation pathway and the Notch signaling may contribute to hypoxia. Our results demonstrate that the consolidation of statistical, structural, and network views of biomolecules and their interactions can provide new insight into the functional role of nsSNPs, which neither the knowledge of molecular structures nor of biological networks alone could achieve. Thus, multiscale modeling of nsSNPs may offer a powerful tool for Genome Wide Association Studies.

Selected Posters

SNP-SIG Meeting – ISMB 2012, July 14th, Long Beach (CA), USA

PREDICTING THE EFFECT OF SINGLE POINT MUTATIONS ON PROTEIN STABILITY USING EVOLUTIONARY INFORMATION

Emidio Capriotti^{*}, Piero Fariselli, Pier Luigi Martelli and Rita Casadio

^{*}*University of Balearic Islands, Spain*
email: emidio.capriotti@uib.es

Protein folding is the mechanism by which a protein reaches its native state that corresponds to a stable three-dimensional structure. Although several interactions concur to the stability of the proteins, they are only marginally stable and the folding free energy change (ΔG) is comparable with the energy of few hydrogen bonds. According to this scenario, a single-point mutation could be responsible for protein misfolding and loss of its function. In this work we analyze the impact of protein variants on the prediction of protein stability. We highlight the effect of protein variations on the predicted free energy change ($\Delta\Delta G$) indicative of protein stability considering protein evolution. In details, first we selected a set of 1,678 single point protein mutations including their reverse from the ProTherm database for which the $\Delta\Delta G$ value had been experimentally determined. Then we calculated the frequencies of the wild-type and mutant residues using a multiple sequence alignment from the BLAST output. This analysis reveals that wild-type residues that correspond to mutations decreasing the protein stability tend to be more conserved than those related to stabilizing mutations. Finally, we implemented a new version of the I-Mutant algorithm including information from protein sequence profile. When tested in cross-validation on the prediction of the $\Delta\Delta G$ sign (negative: destabilization; positive: increase in stability), the new version of I-Mutant scores with 86% accuracy, 0.73 correlation coefficient and 0.90 area under the ROC curve, all values higher than before. Our data add to the role of evolution information in improving predictor efficiency.

A NOVEL COMPUTATIONAL METHOD TO INFER THE DERIVED ALLELE FREQUENCIES AND THE NUMBER OF SINGLE-NUCLEOTIDE POLYMORPHISMS FROM A POOLED DNA SEQUENCING DATA

Natalia Rodchenko^{*}, Joseph P. Dunham, Andrew D. Smith and Sergey Nuzhdin

^{*}*University of Southern California (CA), USA*
email: rodchenk@usc.edu

Important conclusions about evolutionary processes such as various types of selection and demographic changes including population expansion and subdivision that define the genetic structure of a given population can be made based on the derived allele frequencies and the number of polymorphic sites in the genome. Given a pooled DNA sequencing experiment it is of great interest to infer the derived allele frequencies and proportion of polymorphic sites from the data. However, this is a challenging task often exacerbated by a low coverage and a high sequencing error rate in pooled datasets making it impossible to distinguish true mutations from sequencing errors. Additionally, information about haplotypes present in the sequenced sample can not be easily obtained from this type of experiment. We present a novel probabilistic model that incorporates the derived allele frequencies and the number of single-nucleotide polymorphisms as its parameters and allows us to infer these quantities from pooled DNA sequencing data using an EM algorithm. One of the key advantages of our approach is that we are able to explicitly model and correct the parameter estimates for the sequencing errors. Furthermore, our model allows us to use a whole dataset without discarding sites having low coverage and avoid SNP calling at each genomic location as this is unachievable for poorly covered sites. In order to assess the performance of the model we use simulated data as well as the pooled sequencing experiments from a population of somatic cells in the flatworm species *Schmidtea mediterranea*.

ACKNOWLEDGMENTS

The SNP-SIG meeting organizers would like to acknowledge the invited speakers:

- Russ Altman, Stanford University, Palo Alto (CA), USA
- Steven Brenner, University of California, Berkeley (CA), USA
- David Haussler, University of California, Santa Cruz (CA), USA
- Olivier Lichtarge, Baylor College of Medicine, Houston (TX), USA
- Zemin Zhang, Genentech, San Francisco (CA), USA

the following scientists involved in the organization of the roundtable discussion:

- Andre Franke, Christian-Albrechts-University, Kiel, Germany
- Rachel Karchin, John Hopkins University, Baltimore (MD), USA
- Sean Mooney, Buck Institute, Novato (CA), USA
- Shamil Sunyaev, Harvard Medical School, Boston (MA), USA

The organizers also acknowledge **BIOBASE International** (www.biobase-international.com) for its financial support.

AUTHOR INDEX

Ali, Thahmina	8	Nuzhdin, Sergey	9
Alnumair, Nouf S. A.	4	Ong, Cheng Soon	5
Antipin, Yevgeniy	6	Parisi, Gustavo	5
Auer, Jim	7	Peterson, Thomas	6
Bourne, Philip	8	Reva, Boris	6
Bromberg, Yana	4	Rodchenko, Natalia	9
Capriotti, Emidio	5,7,9	Schacherer, Frank	7
Casadio, Rita	5,9	Schwarz, Jana Marie	7
Dunham, Joseph P.	9	Seelow, Dominik	7
Ezequiel, Juritz	5	Seligman, Chet	7
Fariselli, Piero	5,9	Shi, Fan	5
Ferreira, Barbara	8	Smith, Andrew D.	9
Fornasari, Maria Silvina	5	Stern, Linda	5
Goudey, Benjamin	5	Tanweer, Maliha	8
Haddad, Gabriel	8	Thusberg, Janita	7
Inouye, Michael	5	Valencia, Raoul	8
Kann, Maricel	6	Wang, Qiao	5
Klein, Teri	8	Whirl-Carrillo, Michelle	7
Kowalczyk, Adam	5	Xie, Lei	8
Li, Biao	8	Xie, Li	8
Macintyre, Geoff	5	Xue, Vincent	8
Martelli, Pier Luigi	5,9	Zhou, Dan	8
Martin, Andrew C. R.	4		
Miller, Jackson	7		
Mooney, Sean	7		