SNP special interest group

SNP-SIG Meeting

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

ISMB/ECCB 2013
July 19th 2013, Berlin (Germany)

http://snpsig.biofold.org/

Invited Speakers

Steven Brenner
University of California, Berkeley (CA), USA
CAGI Experiments.

Paul Flicek
EMBL European Bioinformatics Institute, Hinxton, UK
Using comparative genomics to restrict the search space of regulatory variants causing rare disease.

Alon Keinan
Cornell University, Ithaca (NY), USA
Recent human population growth: rare variants, mutation load, and complex disease.

Manolis Kellis
Massachusetts Institute of Technology, Cambridge (MA), USA
Regulatory Genomics and Epigenomics of complex disease.

Ruth Nussinov
National Cancer Institute, Frederick (MD), USA
Mapping SNPs and oncogenic mutations onto structural pathways of cancer and inflammation.

SNP-SIG Organizers
Yana Bromberg, Rutgers University, New Brunswick (NJ), USA
Emidio Capriotti, University of Alabama at Birmingham, Birmingham (AL), USA

Round Table Discussion
Paul Flicek, EMBL European Bioinformatics Institute, Hinxton, UK
Alon Keinan**,** Cornell University, Ithaca (NY), USA
Manolis Kellis**,** Massachusetts Institute of Technology, Cambridge (MA), USA
Sean Mooney, Buck Institute, Novato (CA), USA
Ruth Nussinov, National Cancer Institute, Frederick (MD), USA

SNP-SIG Meeting Programme - July 19th 2013, Berlin, Germany

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: Ruth Nussinov**, National Cancer Institute, Frederick (USA)
Mapping SNPs and oncogenic mutations onto structural pathways of cancer and inflammation.

09:20 – 09:45 **Christopher Yates.** Imperial College, London (UK)
Disease-Propensity and nsSNP Phenotype Prediction.

09:45 – 10:10 **Abdul Sattar**. Griffith University, Brisbane (Australia)
Improving mutation-induced stability changes prediction in unseen non-homologous proteins with feature-based multiple models.

10:10 – 10:30 **Coffee Break**

10:30 – 10:55 **Martin Kircher**. University of Washington, Seattle (USA)
A general framework for estimating the relative pathogenicity of human genetic variants.

10:55 – 11:20 **Bjoern Stade.** Christian-Albrechts-University, Kiel (Germany)
snpActs: A versatile analysis tool for annotating and prioritizing SNV data sets.

11:20 – 12:10 **Keynote: Manolis Kellis**, MIT, Cambridge (USA)
Regulatory Genomics and Epigenomics of complex disease.

12:10 – 12:25 **Company Presentation: Frank Schacherer,** BIOBASE GmbH.

12:25 – 13:15 **Lunch Break and Poster Session with the Authors**

Session 2: SNPs as effectors of change: disease and evolution

13:20 – 14:10 **Highlight Speaker: Paul Flicek,** EMBL-EBI, Cambridge (UK)
Using comparative genomics to restrict the search space of regulatory variants causing rare disease.

14:10 – 14:35 **Graham Ritchie.** EMBL-EBI, Cambridge (UK).
Functional annotation of non-coding variants.

14:35 – 15:00 **Andrey Grigoriev.** Rutgers University, New Brunswick (USA).
Visualization of nucleotide substitutions in the transcriptome.

15:00 – 15:25 **John Moult.** University of Maryland, Rockville (USA)
GWAS and Drug Targets.

15:25 – 15:45 **Coffee Break**

15:45 – 16:35 **Keynote: Alon Keinan**. Cornell University, Ithaca (USA).
Recent human population growth: rare variants, mutation load, and complex disease.

16:35 – 17:15 **CAGI Report: Steven Brenner,** University of California, Berkeley (USA).
CAGI Experiments

17:15 – 18:05 **Round Table Discussion**

18:05 – 18:15 Closing remarks from the committee

Invited Presentations

USING COMPARATIVE GENOMICS TO RESTRICT THE SEARCH SPACE OF REGULATORY VARIANTS CAUSING RARE DISEASE

Paul Flicek

EBI-EMBL, Hinxton, UK
email: flicek@ebi.ac.uk

Exome sequencing enables the discovery of coding sequence variants and is a potentially powerful technique to discover variants causing rare disease by targeting disease families with signatures of segregation or de novo mutation. Recent reports suggest that exome sequencing fails to identify causative variants in 50-90% of cases, and at least some of these must be caused by regulatory variants. Unfortunately regulatory regions are tissue specific and cover a fraction of genome so large that whole genome sequencing is likely to be more cost effective and determining the tissue-specific regulatory variants likely to cause disease remains a significant challenge. We have developed a method that leverages functional rather than sequence conservation that identifies a few thousand regulatory regions in liver that are enriched for disease causing variants and demonstrate the effectiveness by identifying the molecular cause of a type of haemophilia that had been unknown for more than 20 years.

USING COMPARATIVE GENOMICS TO RESTRICT THE SEARCH SPACE OF REGULATORY VARIANTS CAUSING RARE DISEASE

Alon Keinan

Cornell University, Ithaca (NY), USA
email: ak735@cornell.edu

Human populations have experienced recent explosive growth since the Neolithic revolution. We demonstrate how such growth predicts an abundance of rare variants, and show that it has not been captured by earlier demographic modeling studies mostly due to small sample size. Recent studies that sequenced a very large number of individuals observed an extreme excess of rare variants, and provided clear evidence of recent population growth, though demographic estimates have varied greatly among studies. These studies were based on protein-coding genes, in which variants are also impacted by natural selection. Hence, we introduce new targeted sequencing data for studying recent human history with minimal confounding by natural selection. Modeling recent demographic history based on the allele frequency spectrum of these data, our models fit very well and shed light on the discrepancies among recent studies. Another important question is how negative selection operates during a recent epoch of rapid population growth, when the population is not at equilibrium. We examined the trajectories of mutations with different fitness effects in forward-in-time simulations and conclude that each individual carries a larger number of deleterious alleles than expected in the absence of growth, but the average fitness effect of these alleles is less deleterious. Combined, our results point to increased load of rare variants with small effect size playing a role in the individual genetic burden of complex disease risk.

REGULATORY GENOMICS AND EPIGENOMICS OF COMPLEX DISEASE

Manolis Kellis

MIT, Cambridge, (MA), USA
email: manoli@mit.edu

To understand the molecular basis of complex disease, we integrate genomic variation, epigenomic variation, and functional genomics datasets with genome-wide association studies. We use reference epigenomic maps of 98 human tissues and cell types to dramatically expand the annotation of non-coding regions and to link active enhancers to their upstream regulators and their downstream target genes using coordinated patterns of activity across cell types. We use the resulting regulatory predictions to revisit disease-associated loci, identifying SNPs that disrupt or create predicted enhancer and causal regulatory motifs, and providing mechanistic hypotheses for the observed associations for individual loci. Beyond the few genome-wide significant loci retained by traditional GWAS, we find functional enrichments across thousands of type-1-diabetes-associated SNPs in cell type-specific enhancers using a rank-based statistical test for enrichment. Beyond reference epigenomes, we study both genomic and epigenomic variation in Alzheimer's disease in a study of 750 individuals, revealing a global hyper-methylation signature in brain-specific enhancers containing specific motifs and methyl-QTLs for 60,000 probes. We systematically validate thousands of our regulatory predictions using Massively Parallel Reporter Assays by disrupting individual binding sites and individual nucleotides of predicted causal regulators, revealing their distinct roles in specifying enhancer activity. Our results suggest a general framework for integrating multi-cell functional genomics and epigenomics information to decipher cis-regulatory connections in complex disease.

MAPPING SNPS AND ONCOGENIC MUTATIONS ONTO STRUCTURAL PATHWAYS OF CANCER AND INFLAMMATION

Ruth Nussinov

National Cancer Institute, Frederick (MD), USA.
email: nussinor@helix.nih.gov

Structural pathways are important. They are essential to the understanding of how oncogenic mutations work and to figuring out alternative parallel pathways in drug resistant mutants. Structural pathways also help to understand the inter-relationship among linked phenomena, as in the case of inflammation and cancer. Cell biology provides a global overview of the behavior of the cell, tissue and the organism under different sets of conditions; the structures of single proteins and their coherent interactions provide insight into the dynamic changes in the proteins, such as those taking place through post-translational modifications, binding events and mutations, and into their interactions. Nonetheless, beyond the challenging construction of structural pathways, there is also a need to obtain a mechanistic insight into single proteins, their modifications, interactions and broadly, their changing landscapes. Why is insight into the dynamic landscape of single proteins important? Perceiving proteins' behavior can help to forecast allosteric transitions, and regulation, and it can help relate oncogenic mutations to their constitutive consequences. The talk will largely focus on structural pathways related to cancer and show how mapping SNPs and oncogenic mutations onto structural pathways of cancer and inflammation helps in understanding their mechanism.

Selected Presentations

IMPROVING MUTATION-INDUCED STABILITY CHANGES PREDICTION IN UNSEEN NON-HOMOLOGOUS PROTEINS WITH FEATURE-BASED MULTIPLE MODELS

Lukas Folkman*, Bela Stantic* and Abdul Sattar*

*Griffith University, Nathan (QLD), Australia
email: {L.Folkman,b.stantic,a.sattar}@griffith.edu.au

Reliable prediction of stability changes induced by protein mutations is an important aspect of computational protein design. Several machine learning methods capable of predicting stability changes from the protein sequence alone have been introduced. However, their performance on mutations in previously unseen non-homologous proteins is relatively low. Moreover, the performance varies for different types of mutations based on the secondary structure, accessible surface area, or magnitude of the stability change. In this work, we explored how designing multiple models, each trained for a different type of mutations, can be beneficial for prediction. We identified specific features to make each model highly specialised. Our results show that combining such feature-based multiple models increases prediction accuracy when compared to currently available methods. In particular, building three models for mutations in helical, sheet, and coil residues yielded the best performance. We found that each of these models included features describing evolutionary conservation and accessible surface area of the mutated residue. However, the remaining features were model-specific. These results support a presumption that different interactions govern protein stability in residues with different types of secondary structure. The concept of feature-based multiple models could be extended to related areas predicting the impact of a protein mutation.

A GENERAL FRAMEWORK FOR ESTIMATING THE RELATIVE PATHOGENICITY OF HUMAN GENETIC VARIANTS

Martin Kircher*, Daniela Witten, Gregory Cooper and Jay Shendure

*University of Washington, Seattle (WA), USA
email: mkircher@uw.edu

As genetic information is insufficient to unambiguously implicate many disease-causal variants, annotations that enrich for causal variation are essential. Current annotations tend to exploit a single information type (e.g. conservation) and/or are restricted in scope (e.g. to missense changes). A broadly applicable metric that objectively weights and integrates diverse information is needed. Here, we describe Combined Annotation Dependent Depletion (CADD), a framework that integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations. We implement CADD as a support vector machine, trained to use 63 annotations to differentiate 14.7 million variants derived on the human lineage from 14.7 million simulated variants. We pre-compute CADD-based scores (C-scores) for all 8.6 billion possible single nucleotide variants of the reference genome and enable scoring of short insertions/deletions. C-scores strongly correlate with allelic diversity, pathogenicity of both coding and non-coding variants, and experimentally measured regulatory effects, and also highly rank causal variants within individual genome sequences. Finally, C-scores of complex trait-associated variants from genome-wide association studies (GWAS) are significantly higher than matched controls and correlate with study sample size, likely reflecting the increased accuracy of larger GWAS. Thus, the ability of CADD to quantitatively prioritize functional, deleterious, and disease causal variants across a wide range of functional categories, effect sizes and genetic architectures is unmatched by any current annotation and will be widely useful for the identification of causal variation in both research and clinical settings.

VISUALIZATION OF NUCLEOTIDE SUBSTITUTIONS IN THE TRANSCRIPTOME

Ammar Naqvi, Tiange Cui and Andrey Grigoriev*

*Rutgers University, Camden (NJ), USA
email: agrigoriev@camden.rutgers.edu

We present a novel approach for interactive representation of nucleotide substitutions and modifications in the transcribed genome. With the focus on RNA secondary structure in the context of next-generation sequencing data, it provides intuitive visualization of genomic environment, sequenced reads, nucleotide polymorphisms and editing events integrated with the structural and functional elements of RNA molecules encoded in the region of interest. This approach can find application for visually relating substitution, structure and function for both coding and non-coding RNA. We also discuss examples of polymorphisms and editing in the context of the secondary structure of microRNAs.

GWAS AND DRUG TARGETS

Lipika Pal, Chen Cao, Chen-Hsin Yu and John Moult*

*University of Maryland, Rockville (MD), USA
email: jmoult@umd.edu

We address the issue of leveraging genomic data on the relationship between genetic variation and the risk of complex trait disease to discover new therapeutic strategies. To this end, we use a two-stage model to link the presence of a genetic variant to the effect on a disease phenotype. The first stage of the model provides an estimate of the effect of the variant on the level of in vivo activity of the relevant protein, and the second stage provides an estimate of the coupling between the activity of that protein and the disease phenotype. We have used the first stage of this model to investigate the relative roles of genetic variants that have a high and low impact on protein function. The second stage may be used to aid in the identification of potential drug targets among GWAS genes. To test this possibility, we have investigated the extent to which GWAS studies re-discover known drug targets. Surprisingly, there are very few of these directly identified by GWAS. However, further investigation shows that the GWAS associated genes for a particular disease, together with network information, can be used to identify potential drug targets.

CANCER SURVIVAL IS ASSOCIATED WITH THE NUMBER OF PREDICTED FUNCTIONAL MUTATIONS IN A TUMOR

Boris Reva

MSKCC, New York (NY), USA
email: reva@cbio.mskcc.org

Using mutation data and clinical information of ten cancers studied in TCGA project, we tested the association between aggressiveness of cancer and various types of genomic alteration, such as missense and truncating mutations, predicted functional mutations and gene copy number alterations. For prediction of mutation impact we used the score of Mutation Assessor [1], which assess the mutation impact by the value of entropic disordering of the evolutionary conservation patterns in protein sequence alignments. We found that the total number of mutations and, especially, the number of predicted functional mutations is the most discriminative collective marker of outcome for eight of ten studied cancers and especially for kidney, ovarian and acute myeloid leukemia: the more predicted functional mutations, the poorer outcome. This result underscores the carcinogenic role of mutations in rarely mutated genes -- a mutation “long tail” -- and suggests that the aggressiveness of cancer may depend rather on the total number of driver mutations than on the limited sets of the specifically mutated genes. Because numbers of predicted functional mutations in tumors are significantly smaller than total numbers of mutations in tumors, the reduced sets of predicted functional mutations can significantly facilitate determining genomic markers of outcome, identifying the activated pathways and advancing personalized multi-drug therapy.

FUNCTIONAL ANNOTATION OF NON-CODING VARIANTS

Graham Ritchie[*], Paul Flicek and Eleftheria Zeggini

[*]*EMBL-EBI, Hinxton, UK*
email: grsr@ebi.ac.uk

Identifying functionally relevant variants against the background of ubiquitous genetic variation is a major challenge in human genetics. For variants that fall in protein-coding regions our understanding of splicing and the genetic code allow us to readily interpret possible functional effects. There are, however, currently few integrated methods to interpret variants that fall outside of coding regions and yet these are increasingly being identified as causally relevant in human disease. Efforts such as ENCODE and the Roadmap Epigenomics Project are producing a wealth of annotation in non-coding regions but it is not clear how to integrate the wide range of data into variation studies. To establish which of these annotations might be informative when interpreting variants we use a set of annotated regulatory mutations implicated in human disease from the Human Gene Mutation Database and look for overlaps with annotated regions from these projects across multiple cell lines. We compare these results with annotations of several control sets of common variants identified in the 1000 Genomes Project. We show that the disease-implicated loci are enriched for marks of open chromatin, histone modifications, and evidence of binding for RNA polymerase II and several transcription factors. Using these annotations, along with other relevant genomic properties such as genic context, evolutionary conservation and variation in human populations, we build a classifier that can discriminate between the HGMD mutations and our control sets with good accuracy. We discuss several ways in which scores from this classifier can find application in next generation association studies.

SNPACTS: A VERSATILE ANALYSIS TOOL FOR ANNOTATING AND PRIORITIZING SNV DATA SETS

Bjoern Stade*, David Ellinghaus, Britt-Sabina Petersen and Andre Franke*

*Christian-Albrechts-University, Kiel, Germany
email: b.stade@ikmb.uni-kiel.de, a.franke@mucosa.de

The rapidly decreasing prices for sequencing entire human exomes and genomes result in large amounts of variation data. To cope with these data we developed snpActs, a database-driven toolset that allows scientists to annotate single nucleotide variants (SNVs) and categorize them comprehensively. snpActs scans different gene annotations and identifies SNVs in functional elements. Additionally, snpActs utilizes the results of several established mutation effect prediction algorithms, such as SNAP, SIFT, and Polyphen2, to distinguish between deleterious and functionally neutral amino acid changes caused by SNVs. For this, it also checks the Human Gene Mutation Database (HGMD). In comparison to other annotation-programs snpActs can filter SNV lists using certain rules (e.g. coding SNVs), special masks (e.g. cancer regions) or based upon other SNV lists (e.g. presence in relatives).

snpActs further implements a classical and precise linkage analysis to examine regions that are identical-by-descent in data sets from complex pedigrees. Via these functionalities it is possible to identify potential disease causing genes in examined individuals.

DISEASE-PROPENSITY AND NSSNP PHENOTYPE PREDICTION

Christopher Yates* and Michael J.E. Sternberg

*Imperial College, London, UK
email: c.yates11@imperial.ac.uk

We introduce a new classification termed disease-propensity for proteins and domains, based on the relative numbers of disease-associated and neutral nsSNPs (non-synonymous SNPs) they contain. Using a binomial test, we classify 311 proteins and 112 Pfam families as disease-susceptible and 32 proteins and 67 Pfam families as disease-resistant. These groups of proteins and domains differ in a number of features, such as function, sequence conservation, essentiality and centrality in interaction networks. For example, disease-resistant proteins and domains are more likely to be involved in immunity, whereas disease-susceptible proteins and domains are more likely to be encoded by essential genes.

This classification may be useful in prioritising variants discovered using genome-wide association studies and for predicting nsSNP phenotype, so we are currently developing a predictor using this and other sequence and structural features, which we term SuSPect (Disease-Susceptibility and Structure based Predictor). nsSNPs are mapped to experimental protein structures or homology models, if available, and classified using a support vector machine (SVM). In 10-fold cross-validation, we achieve a balanced accuracy of up to 93% and a Matthew's correlation coefficient of up to 0.83. In two blind tests, SuSPect performs at least as well as other widely-used predictors.

WINNOW: A TOOL TO FILTER AND PRIORITIZE EXOME VARIANT DATASETS TO EXTRACT USER-SPECIFIC RESULTS

Ashwini Bhasi, Zachary Wright and James Cavalcoli*

*University of Michigan, Ann Arbor (MI), USA
email: cavalcol@med.umich.edu

We examined the extent to which non-synonymous single nucleotide polymorphisms (nsSNPs) are disease-causing due to effects at protein-protein interfaces. Our method was to integrate a database of 3D structures of human protein/protein complexes and the UNIPROT humsavar nsSNPs database. nsSNPS were analyzed in terms of their location in the protein core, protein-protein interfaces, and on the surface when not at an interface. Within our dataset of 1027 proteins, 537 proteins had at least one nsSNP (median number of nsSNPs=2, range: 1-257). The total number of nsSNPs for our protein dataset was 4315, of which 24.8% occurred in the protein core, 29.3% in interfaces and 45.9% at the surface.

We found that disease-causing nsSNPs, which do not occur in the protein core, were significantly more often located at protein interfaces rather than surface non-interface regions (OR 1.59, $p<0.0001$). The disruption of the protein-protein interaction can be explained by a range of structural effects, including the loss of an electrostatic salt bridge, destabilization due to reduction of the hydrophobic effect, formation of a steric clash, and the introduction of a proline altering the main-chain conformation. Moreover, polymorphisms occurred significantly more often on the surface rather than the protein core (OR 0.67, $p<0.0001$). Many of these nsSNPs, which are considered neutral genetic variations, had very low BLOSUM62 score, thus suggesting a potential mild deleterious effect.

In conclusion, the results of this study suggest that the use of the interactome, can aid in the understanding of the cause-effect relation of disease-associated nsSNPs.

PREDICTSNP: ROBUST AND ACCURATE CONSENSUS CLASSIFIER FOR PREDICTION OF DISEASE- RELATED MUTATIONS

Jaroslav Bendl, Jan Stouraca, Ondrej Salanda, Antonin Pavelka, Eric D. Wieben, Jaroslav Zendulkab, Jan Brezovsky and Jiri Damborsky*

*Masaryk University, Brno, Czech Republic
email: jiri@chemi.muni.cz

Human genetic variations occur primarily as a result of single nucleotide polymorphisms (SNPs). Many methods have been developed to predict SNPs potentially causing genetic diseases. Since these methods have different founding principles and decision boundaries, users can be confused by their conflicting predictions. Comparison of methods and their possible improvement is challenging because of large overlaps between the training datasets and testing datasets. To address this problem, we have constructed three independent datasets: one benchmark and two testing datasets. The benchmark dataset consists of over 43,000 mutations and was used for the assessment of performance of eight well-established prediction tools. This dataset did not contain any duplicities, inconsistencies or mutations previously used for the training of evaluated tools. On the basis of this evaluation, we selected six best performing methods (MAPP, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT and SNAP) and combined them into a consensus classifier PredictSNP. A method for the construction of consensus function was selected from eight common machine learning approaches based on its accuracy and transparency. In consequent evaluation of PredictSNP with the two testing datasets, the method showed better performance than any of the six constituent predictors. Web interface of PredictSNP combines the results of selected computational methods with the experimental data from Protein Mutant Database and Uniprot, and is freely available to the academic community at http://loschmidt.chemi.muni.cz/predictsnp.

META-SNP: META-PREDICTOR OF DISEASE CAUSING NON-SYNONYMOUS VARIANTS

Emidio Capriotti*, and Yana Bromberg

*University of Alabama, Birmingham, USA
email: emidio@uab.edu

In recent years the number of human genetic variants deposited into the publicly available databases has been increasing exponentially. The latest version of dbSNP, for example, contains ~50 million validated Single Nucleotide Variants (SNVs). SNVs make up most of human variation and are often the primary causes of disease. The non-synonymous SNVs (nsSNVs) result in single amino acid substitutions and may affect protein function, often causing disease. Although several methods for the detection of nsSNV effects have already been developed, the consistent increase in annotated data is offering the opportunity to improve prediction accuracy.

Here we present a new approach for the detection of disease-associated nsSNVs (Meta-SNP) that integrates four existing methods: PANTHER, PhD-SNP, SIFT and SNAP. We first tested the accuracy of each method using a dataset of 35,766 disease-annotated mutations from 8,667 proteins extracted from the SwissVar database. The four methods reached overall accuracies of 64%-76% with a Matthew's correlation coefficient (MCC) of 0.38-0.53. We then used (an eight-element vector of) the outputs of these methods to develop a machine learning based approach that discriminates between disease-associated and polymorphic variants. In testing, the combined method reached 79% overall accuracy and 0.59 MCC, ~3% higher accuracy and ~0.05 higher correlation with respect to the best-performing method. Moreover, for the hardest-to-define subset of nsSNVs, i.e. variants for which half of the predictors disagreed with the other half, Meta-SNP attained 8% higher accuracy than the best predictor.

Here we find that the Meta-SNP algorithm achieves better performance than the best single predictor. This result suggests that the methods used for the prediction of variant-disease associations are orthogonal, encoding different biologically relevant relationships. Careful combination of predictions from various resources is therefore a good strategy for the selection of high reliability predictions. Indeed, for the subset of nsSNVs where all predictors were in agreement (46% of all nsSNVs in the set), our method reached 87% overall accuracy and 0.73 MCC.

Meta-SNP server is freely accessible at http://snps.biofold.org/meta-snp.

.

PROTEIN-PROTEIN INTERACTION SITES ARE HOT SPOTS FOR DISEASE NON-SYNONYMOUS SNPS

Alessia David*, Mark N. Wass, and Michael J.E. Sternberg

*Imperial College, London, UK
email: alessia.david09@imperial.ac.uk

Single Nucleotide Polymorphisms are invaluable markers for tracing the genetic basis of inheritable traits and the ability to create marker libraries quickly is vital for timely identification of target genes. Next-generation sequencing makes it possible to sample a genome rapidly, but polymorphism detection relies on having a reference genome to which reads can be aligned and variants detected. We present Bubbleparse, a method for detecting variants directly from next-generation reads without a reference sequence. Bubbleparse uses the de Bruijn graph implementation in the Cortex framework as a basis and allows the user to identify bubbles in these graphs that represent polymorphisms, quickly, easily and sensitively. The Bubbleparse algorithm is sensitive, can detect many polymorphisms quickly and performs well when compared with polymorphism detection methods based on alignment to a reference in Arabidopsis thaliana and found some SNPs not found by the canonical method. The heuristic can be used to maximise the number of true polymorphisms returned and with a proof-of-principle experiment we show that Bubbleparse is very effective on data from unsequenced wild relatives of potato and enabled us to identify disease resistance linked genes quickly and easily. Bubbleparse is a fast and effective tool for detection of polymorphisms in unsequenced genomes and is an excellent addition to the genomics toolbox, it can speed up variant detection and allow for new analyses in organisms that do not as yet have substantial genomic resources.

BOOGIE: BLOOD GROUP PREDICTION FROM NEXT-GENERATION SEQUENCING DATA

Manuel Giollo, Giovanni Minervini, Marta Scalzotto, Emanuela Leonardi, Carlo Ferrari and Silvio C E Tosatto*

*University of Padova, Padova, Italy
email: silvio.tosatto@unipd.it*

The $1.000 genome and $100.000 analysis is one of the most challenging issues of the Next Generation Sequencing (NGS) era, and probably the greatest limit for its use in the clinical practice.

In the current work we propose BOOGIE, a framework for the inference of human physical traits from public database information. Far from a complete genome characterization, we focus on the prediction of the 32 blood group reported in the Blood Group Antigen Gene Mutation Database (BGMUT). The database is maintained by the NCBI, and contains a curated list of alleles and corresponding blood antigens for most of the relevant blood systems. We used this information source for the annotation of genome data. This is clearly relevant for blood transfusions, since most health-care applications consider just three blood systems (mainly ABO, RH and Kell). In addition, BOOGIE can be of interest for the detection of hemolytic disease of the newborn.

Even thought the haplotype phasing problem cannot be solved, we evaluated BOOGIE using the public data of the Personal Genome Project. In particular, almost 200 samples with genetic data and ABO blood group annotation were extracted and preliminary results show that BOOGIE can predict the most important blood groups with more than 92% accuracy in a few seconds on a desktop PC. This efficiency is clearly an important feature in view of future developments with more phenotype predictions, especially when considering scalability of the tool for personalized medicine.

VCF FEATURES TO TRAIN SVM IN GRAPEVINE SNP DETECTION

Lorena Leonardelli*, Alessandro Cestaro, Carmen Maria Livi, Charles Romieu, Patrice This, Enrico Blanzieri and Claudio Moser

*Fondazione Edmund Mach, S. Michele all'Adige, Italy
email: lorena.leonardelli@gmail.com*

Although next generation sequencing (NGS) technologies are increasing genomic information at unprecedented pace, they are prone to an error rate even higher than one each 100 bp. Efficient approaches are thus needed to distinguish real polymorphisms from the abundant sequencing artefacts. Many open-source tools have been recently developed to identify Single Nucleotide Polymorphisms (SNPs) in whole-genome data, the most popular being Samtools and GATK. Still they present an unsatisfactory accuracy due to high false positive polymorphism prediction. SNPs are the most abundant type of DNA sequence mutations and they are efficient markers for several biological applications such as cultivar identification, construction of genetic maps, the assessment of genetic diversity, the detection of genotype/phenotype associations, or marker-assisted breeding. The biological importance of finding only true SNPs is evident, considering the expensive cost of SNP validation through re-sequencing or SNP-chip, not only in terms of money but also of time and, above all, sample's availability. Since these small mutations can be responsible of large changes in the physiology or the evolution of an organism, our interest is to define if this category of polymorphisms is the genetic determinant of the low acidity content in the grapevine (Vitis vinifera L.) cultivar Gora Chirine. To this aim we are investigating the acidity trait in grapevine by comparative analysis of the genome sequences of Gora and Sultanine, is the latter being a normal acidity grapevine cultivar and genetically a close relative to Gora. Malic acid amount in grape berries is an essential parameter in wine fermentation quality and investigation of new genes involved in grapevine acidic metabolism is a common interest for biologists, enologists and wine makers.

The advent of NGS technologies, such as Illumina/Solexa, AB/SOLiD and Roche/454 created new raw sequence types and several new tools for read alignment have been developed generating alignments in different formats. A standard alignment format supporting all sequence types and aligners allows an easier connection between read alignment and downstream analyses, including variant detection, genotyping and assembly. New data formats for aligned sequences are SAM/BAM (Sequence Alignment/Map and Binary Alignment/Map), now adopted by the entire genomics community. Calling SNPs from SAM/BAM files with predictors like SAMtools and GATK (Genome Analysis Toolkit) provides as output a Variant Call Format (VCF) file. VCF file contains a list of candidate SNPs with relative position on contigs, the nucleotide present on the reference genome and on the alternative alleles, SNP call quality, genotype and many other

parameters. It is hard to consider all those values in order to distinguish which SNPs are actually polymorphisms or sequencing errors, but VCF parameters can be a lot more informative if used to train a Support Vector Machine (SVM) approach that classifies the list of candidate SNPs in real SNPs and false positive results. SVM is an efficient and reliable machine learning method to distinguish categorical data; it separates the positive and negative training data by constructing a linear classifier or a non-linear classifier with a kernel function. Based on training features, SVM represents the data as points in space, where the data belong to two categories (positive and negative) divided by a gap that is as wide as possible. The training features were calculated on an experimentally validated set of SNPs (550 positive data set) and on monomorphic SNP positions (300 negative control data set). The SNP predictors, SAMtools and GATK, output approximately 400 of the 520 positive SNPs and 40 of the 300 negative SNPs, compelling us to re-balance the SNP sets with the SMOTE algorithm before the SVM training. The SVM training was validated by the 10-fold cross validation method. The resulting model will be applied on the biological study mentioned above as well as on other data sets.

SVM trained with 21 and 12 VCF parameters for GATK and SAMtools, respectively, as features has reached an average accuracy of 94% with SAMtools data and 91% with GATK data. The SVM performance suggests which VCF parameters are determinant to understand if polymorphic sites are real SNP sites or errors due to sequencing as well as to low quality nucleotide alignment. SVM can efficiently recognize true SNPs from false positive predictions as shown by high sensitivity (GATK 94%, SAMtools 96%), specificity (GATK 63%, SAMtools 65%), and precision (GATK 97%, SAMtools 97%) resulting from the SVM 10-fold cross validation.

IDENTIFYING AND CLASSIFYING TRAIT LINKED SNPS IN NON-REFERENCE SPECIES BY WALKING COLOURED DE BRUIJN GRAPHS

Dan MacLean, Richard Leggett,
Ricardo Ramirez-Gonzalez,
Cintia Kawashima, Walter Verweij,
Jonathan Jones and Mario Caccamo.

[*]The Sainsbury Laboratory, Norwich, UK
email: dan.maclean@tsl.ac.uk

Exome sequencing is becoming a popular method to study causal variants associated with different diseases. Several variant calling software packages are currently available, but sifting through large datasets of potential variants generated by these software is a time-consuming, multi-step process. A tool to easily filter such large datasets and extract causal variants of specific interest to researchers would be a useful way to speed up the analysis process. To meet this need, we have developed Winnow— a flexible platform to dynamically filter variants based on different user-specified criteria. It has the following features:

(1) User-specific project space to upload sample data, filter relevant variants, and store results. (2) Perform drill-down set operations to compare variants across multiple samples and extract unique variants or recurrent variants. (3) Refine candidate variant lists using filters for: common VCF tags (e.g. QUAL, DP, MQ, PL, GT) and variant caller-specific tags (Samtools, GATK and VarScan); common variants (dbSNP, or 1000Genomes); somatic variants (Samtools CLR scores or Varscan somatic P values); germline variants (non-mendelian, dominant and recessive). (4) Prioritize and filter results based on variant functional impact (protein coding, deleterious effect, etc.) (5) Extract variants belonging to a gene set of interest or those located in a specific chromosome region. (6) Explore transcript isoform variations across samples. (7) View sequence reads across the alignment region of filtered variants.

HOPE AND NEWPROT - CONNECTING WORLDS USING PROTEIN STRUCTURES

Hanka Venselaar* and Gert Vriend

*Radbout University, Nijmegen, Netherlands
email: h.venselaar@cmbi.ru.nl

The ever on-going technical developments in Next Generation Sequencing have led to an increase in detected disease related mutations. A large series of webservers that can analyse these variants exists. However, most of these servers classify a mutation as either 'damaging' or 'not damaging', instead of providing a detailed analysis of the structural effects. We developed HOPE, an automatic webserver that can collect and combine structural information from a series of data sources, including the protein's 3D-structure, annotated information and predictions. HOPE's report provides a detailed insight in the effect of the mutation on the protein's structure and function.

To validate our method, we collected a test dataset of well-described mutations in proteins for which 3D-structure information is available. We compared HOPE's analysis to those made by other webservers and our own manual conclusions. This study provides insight in the possibilities and the limitations of methods based on sequence information alone, hybrid methods, machine learning based methods, and structure based methods.

ACKNOWLEDGMENTS

The SNP-SIG meeting organizers would like to acknowledge:

- Steven Brenner, University of California, Berkeley (CA), USA
- Paul Flicek, EMBL European Bioinformatics Institute, Hinxton, UK
- Alon Keinan, Cornell University, Ithaca (NY), USA
- Manolis Kellis, Massachusetts Institute of Technology, Cambridge (MA), USA
- Sean Mooney, Buck Institute, Novato (CA), USA
- Ruth Nussinov, National Cancer Institute, Frederick (MD), USA

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

AUTHOR INDEX