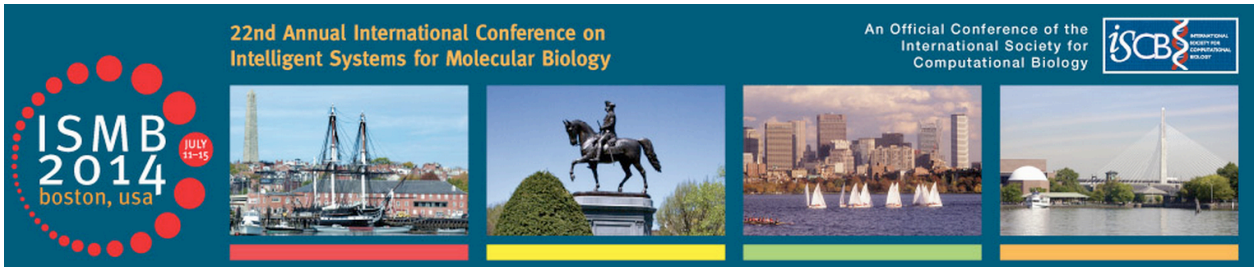
Varl special interest group

Varl-SIG Meeting

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

Room 207
John B. Hynes Memorial Convention Center
July 12th 2014, Boston (MA), USA

http://varisig.biofold.org/



Invited Speakers

George Church
Harvard University, Boston (MA), USA
Testing hypotheses about genomic variation using CRISPR, Organs-on-chips, FISSEQ and OpenHumans.org.

Garry Cutting
John Hopkins University, Baltimore (MD), USA
Interpreting the clinical implications of DNA variants.

Jan Korbel
EMBL, Heidelberg, Germany
Origins and consequences of structural rearrangements.

James Potash
University of Iowa, Iowa City (IA), USA
Sequence in the Service of Psychiatry.

Benjamin Raphael
Brown University, Providence (RI), USA
Algorithms for Cancer Genome Sequencing and Interpretation.

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

Round Table Discussion
Rachel Karchin, John Hopkins University, Baltimore (MD), USA

VarI-SIG Meeting Programme - July 12th 2014, Boston (MA), USA

08:20 – 08:30 Welcome from the committee

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

08:30 – 09:20 **Highlight Speaker: Jan Korbel**, EMBL, Heidelberg (Germany).
Origins and consequences of structural rearrangements

09:20 – 09:45 **Haiming Tang.** University of South California, Los Angeles (CA).
PANTHER-PSEP: predicting disease-causing mutations using position specific evolutionary preservation.

09:45 – 10:10 **Debora S. Marks**. Harvard University, Boston (MA).
Quantitative effects of genetic variants from sequence alone.

10:10 – 10:30 **Coffee Break**

10:30 – 10:55 **Tatiana Tatarinova.** University of South California, Los Angeles (CA).
Geographic Population Structure (GPS): a novel tool to infer biogeographical origins of worldwide human populations.

10:55 – 11:20 **Pietro Di Lena**. University of Bologna, Bologna (Italy).
A new network-based method for gene enrichment analysis: detection of new biological processes associated to OMIM diseases.

11:20 – 12:10 **Highlight Speaker: Benjamin Raphael**, Brown University, Providence (RI).
Algorithms for Cancer Genome Sequencing and Interpretation.

12:10 – 12:25 **Company Presentation: Jennifer Hogan, BIOBASE GmbH**.

12:25 – 13:10 **Lunch Break and Poster Session with the Authors (Hall C)**

Session 2: Genetic variants as effectors of change: disease and evolution

13:20 – 14:10 **Highlight Speaker: George Church,** Harvard University, Boston (MA).
Testing hypotheses about genomic variation using CRISPR, Organs-on-chips, FISSEQ and OpenHumans.org.

14:10 – 14:35 **Yanay Ofran.** Bar Ilan University, Ramat Gan (Israel).
StoP: predicting phenotypes based on SNP-pathway associations.

14:35 – 15:00 **Kristin Ayers.** Mount Sinai School of Medicine, New York (NY).
Assessment of cancer mutational burden from the 1000 genomes sequencing data.

15:00 – 15:25 **Florian Gnad.** Genentech, San Francisco (CA).
Integrated genomic analyses of thousands of tumors reveal cancer-associated epigenetic regulators.

15:25 – 15:45 **Coffee Break**

15:45 – 15:50 **The MD Corner – Presented by Rachel Karchin**

15:50 – 16:35 **Highlight Speaker: Garry Cutting**. John Hopkins University, Baltimore (MD).
Interpreting the clinical implications of DNA variants.

16:35 – 17:20 **Highlight Speaker: James Potash**. University of Iowa, Iowa City (IA).
Sequence in the Service of Psychiatry

17:20 – 18:05 **Round Table Discussion**

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

Invited Presentations

VarI-SIG Meeting – ISMB 2014, July 12th Boston (MA), USA

TESTING HYPOTHESES ABOUT GENOMIC VARIATION USING CRISPR, ORGANS-ON-CHIPS, FISSEQ AND OPENHUMANS.ORG

George M. Church

Harvard University, Boston (MA), USA
email: gmc@harvard.edu

Evaluating variants of unknown significance, especially in noncoding regions, is important for evolutionary studies, clinical analysis and improving predictive algorithms. Four high-throughput components are needed:

1) a way to alter genomes precisely (CRISPR-Cas9),
2) a way to access complex tissues differing by only one or a few mutations (Organs-on-chips),
3) a way to analyze allele-specific expression in a multi-cellular and subcellular context (Fluorescent In Situ Sequencing = FISSEQ),
4) a way to freely share comprehensive big-data including human phenotype and omic information (PersonalGenomes.org, OpenHumans.org).

INTERPRETING THE CLINICAL IMPLICATIONS OF DNA VARIANTS

Garry Cutting

John Hopkins University, Baltimore (MD), USA
email: gcutting@jhmi.edu

We have known for years that most genes associated with Mendelian disorders have a large number of variants. However, the consequence of these variants upon gene function and phenotype are largely unknown. In many cases, only the common variants have been rigorously evaluated. The identification of less common and rare variants by next generation sequencing has created immense translational challenges for diagnostic laboratories. General rules have been applied to interpret variants based on their predicted effect. Variants that introduce premature termination codons (i.e. nonsense frame-shift and canonical splice site) and large scale rearrangements that disrupt coding regions of genes are accepted as having a high likelihood of being deleterious. Variants that cause amino acids substitutions (i.e. missense) and those that alter non-canonical nucleotides in splice sites are challenging to assess. The latter group of variants constitute about half of the variants found in genes associated with recessive disorders. What is needed most urgently are accurate predictive methods for missense and splice site variants. Available algorithms that generate a statistical estimate are useful to some extent in clinical applications. However, diagnostics are performed on individual patients for whom estimates of risk introduce uncertainty that complicates management and treatment decisions. Thus, the clinical field needs algorithms that approach 100% predictive accuracy. In this talk I will provide a synopsis of work from our group in the interpretation of variants in the CFTR gene as an example of the challenges and opportunities in variant annotation.

ORIGINS AND CONSEQUENCES OF STRUCTURAL REARRANGEMENTS

Jan Korbel

EMBL, Heidelberg, Germany
email: korbel@embl.de

My presentation will mainly cover research from our group focused on somatic DNA alterations in the cancer genome. Chromothripsis scars the genome when localized chromosome shattering and repair occurs in a one-off catastrophe. Outcomes of this process are detectable as massive DNA rearrangements affecting one or a few chromosomes. While recent findings suggest a crucial role of chromothripsis in cancer development, the reproducible inference of this process has remained challenging, requiring that cataclysmic one-off rearrangements can be distinguished from localized genetic lesions that occur in a step-wise fashion. We have developed a set of conceptual criteria for the inference of complex DNA rearrangements suitable for rigorous statistical analyses, based on ruling out the alternative hypothesis that DNA rearrangements have occurred in a stepwise (progressive) fashion. In addition to work on chromothripsis I will present data from our group's efforts to decipher the earliest genetic lesions occurring in prostate cancer. In young men prostate cancer appears to initiate through a molecular mechanism involving primarily androgen-driven chromosomal DNA rearrangements, leading to a genomic DNA alteration landscape that is strikingly distinct from prostate cancer genomes of elderly men. We are in the process of using approaches developed in our group to perform an analysis of hundreds of deeply sequenced cancer genomes in the context of the Pan Cancer Analysis of Whole Genomes (PWAG) project organized through the International Cancer Genome Consortium, to search for commonalities and differences in molecular processes leading to cancer in different tumor entities.

SEQUENCE IN THE SERVICE OF PSYCHIATRY

James B. Potash

University of Iowa, Iowa City (IA), USA
email: james-potash@uiowa.edu

It has been 27 years since papers in Nature and Lancet announced linkage between molecular DNA markers and bipolar disorder, or manic-depression. Comparable results were reported in Nature the following year for schizophrenia. Yet relatively little could be demonstrated with confidence over the next 20 years regarding the relationship of sequence to psychiatric disease susceptibility. Since the genome-wide association study (GWAS) era began in our field in 2007, robust associations of genotype and psychiatric phenotype have begun to emerge, though effect sizes are small. This has been particularly true as sample sizes have grown very large. Not only GWAS SNPs, but also copy number variants (CNVs) have been strongly implicated, especially in schizophrenia and autism. Some CNVs have large effect sizes, and, as thesequencing era began, the hope was that rare single nucleotide variants would have large effect sizes too, so that smaller sample sizes would be possible in psychiatric disease variant studies. Studies to date suggest that this is not the case, and that, as with GWAS, large consortia with samples in the thousands to tens of thousands will be required to definitively identify the sought after disease variants. Such consortia have been formed for bipolar disorder and autism sequencing studies. Meanwhile, results to date from GWAS and CNV studies are being moved into the clinic, though the impact on patient care remains uncertain. For example, companies such as AssurexMD and GenoMind are marketing genetic tests to direct antidepressant treatment. Careful study and robust replication should be required before genetic testing can be confidently used in psychiatric care.

ALGORITHMS FOR CANCER GENOME SEQUENCING AND INTERPRETATION

Benjamin J Raphael

Brown University, Providence (RI), USA
email: braphael@brown.edu

The rapidly declining costs of DNA sequencing have enabled large-scale studies of the somatic mutations that drive cancer development. However, both the measurement and interpretation of somatic mutations is complicated by extensive mutational heterogeneity. This heterogeneity is apparent both *across* tumors -- where different individuals with the same cancer type exhibit different combinations of *driver* mutations -- and *within* a tumor, where individual cells in a tumor may possess different complements of somatic mutations. We describe algorithms to address both sources of heterogeneity. In the former case of inter-tumor heterogeneity, we describe the *HotNet* and *Dendrix* algorithms to identify *driver pathways*, groups of genes containing driver mutations, in a large cohort of cancer samples. In the later case of intra-tumor heterogeneity, we present *THetA*, an algorithm that uses convex optimization techniques to infer tumor composition, including the proportions of normal (non-cancerous) cells and one or more populations of tumor cells, in a single sample. We apply these algorithms to genome/exome sequencing and array copy number data from several cancer types in The Cancer Genome Atlas (TCGA).

Selected Presentations

VarI-SIG Meeting – ISMB 2014, July 12th Boston (MA), USA

ASSESSMENT OF CANCER MUTATIONAL BURDEN FROM THE 1000 GENOMES SEQUENCING DATA

Kristin Ayers, Yevgeniy Antipin, Rong Chen* and Boris Reva*

*Mount Sinai School of Medicine, New York (NY), USA
email: {rong.chen, boris.reva}@mssm.edu

Large-scale sequencing of human genomes has revealed many thousands of new natural variants. Many of these variants make significant impact on gene function, predisposing individuals to various diseases and cancer. However, for majority of these variants their role in diseases, and specifically, in cancer is not known. In this work, we nominate new cancer associated non-synonymous variants by determining the functional variants affecting cancer driver genes. The variants and their frequencies were extracted from the 1000 genomes sequencing project. For prediction of mutational impact, we used the score from the Mutation Assessor software. A list of ~1000 cancer genes was formed from annotated cancer genes and genes nominated as cancer drivers from the eleven cancers in The Cancer Genome Project. The predicted impact scores of population variants were compared with the impact scores of somatic mutations in TCGA. The distribution of the impact scores show negative selection for variants and positive selection for somatic mutations. We predicted new cancer associations for variants of genes of known cancer predisposition and determined the population distribution across ethnicities of variants in cancer genes (specifically in tumor suppressors). The population distributions of cancer gene variants are very diverse, on average, every individual is affected by 20 to 60 functional variants in tumor suppressors; the majority of these variants being rarely observed alleles. These numbers are higher for African population. We hypothesize that an individual burden of functional germline variants in cancer genes can be associated with a risk of development cancer and its aggressiveness.

STOP: PREDICTING PHENOTYPES BASED ON SNP-PATHWAY ASSOCIATIONS

Aharon Brodie and Yanay Ofran*

*Bar Ilan University, Ramat Gan, Israel
email: yanay@ofranlab.org

We introduce StoP, (SNPs to Pathways), a web-server that analyzes SNPs in a given genome. Upon submitting the SNPs of a person (including undocumented ones), the server returns a score for over 100 diseases and phenotypes, using pathway-based analysis of the observed genomic variations. We assess these predictions using the CAGI 2013 data and show its potential for correct phenotype prediction. StoP does not assess the molecular effects of individual SNPs, thus it is complementary to most phenotype prediction tools. Gene sets are used to overcome some of the great challenges of interpreting genome-wide association studies (GWAS). The pathway-based approach for gene-set analysis assumes that since function is executed through the interactions of multiple proteins, different perturbations of a pathway could result in a similar phenotype. This assumption, however, was not systemically assessed. To determine whether SNPs associated with a given complex phenotype affect the same pathways more than expected by chance, we analyzed 368 phenotypes that were studied in >5000 GWAS. By linking phenotype-associated SNPs to genes, we mapped phenotypes to known pathways and determined the statistical significance of these associations. Based on this mapping, we developed a scheme for scoring the perturbations in each phenotype-associated pathway, for a given individual. This scheme is the basis of StoP.

We participated in two challenges in the CAGI community experiment, in which participants were tasked with predicting which individuals from a dataset are sick. In the Hypoalphalipoproteinemia (HA) challenge, we correctly predicted not only which family member is sick, but also provided SNP-based scores, which corresponded to the HDL-C levels of each family member. In the second challenge, participants were tasked with differentiating individuals with Crohn's disease from healthy ones. We reached prediction accuracy up to 88% using the StoP approach. Thus, StoP can assist in revealing relevant phenotype-pathway associations as well as in predicting phenotype from genomes.

A NEW NETWORK-BASED METHOD FOR GENE ENRICHMENT ANALYSIS: DETECTION OF NEW BIOLOGICAL PROCESSES ASSOCIATED TO OMIM DISEASES

Pietro Di Lena[*], Pier Luigi Martelli[*], Piero Fariselli[*] and Rita Casadio[*]

[*]*University of Bologna, Bologna, Italy*
email: {pietro.dilena, pierluigi.martelli, piero.fariselli, rita.casadio}@unibo.it

Enrichment analysis is a widely applied tool for determining functional associations between biological processes or pathways and genes or proteins related to the same phenotype. This procedure is useful for shedding light on the molecular mechanisms and functions at the basis of the analysed phenotype, for enlarging the dataset of possibly related genes and proteins and for helping interpretation and prioritisation of new experimentally determined variations.

Standard enrichment methods routinely rely on the only annotations that characterize the genes/proteins included in the input. Rarely they take into consideration the physical and functional relationships among different genes or proteins that can be extracted from the available biological networks of interactions.

Here we describe a method that combines the standard enrichment technique with a new procedure based on the analysis of the sub-networks connecting genes or proteins that share the same functional annotation. These sub-networks, derived from STRING, are generated by finding the minimal graphs that connect the all the proteins that share the same Gene Ontology annotation (Biological Process). We test the ability of the network-based enrichment method in finding annotation terms disregarded by a standard enrichment method and analyse 244 sets of proteins associated to different OMIM diseases. In 149 cases (61%) the network-based procedure extracts GO terms neglected by the standard method and in 79 cases (32%) some of the enriched GO terms are not included in the annotations of the original protein set.

INTEGRATED GENOMIC ANALYSES OF THOUSANDS OF TUMORS REVEAL CANCER-ASSOCIATED EPIGENETIC REGULATORS

Florian Gnad[*] and Zemin Zhang[*]

[*]*Genentech, San Francisco (CA), USA*
email: {gnadf, zhang.zemin}@gene.com

Background: Many cancer cells show distorted epigenetic landscapes characterized by alterations of chromatin-modifying enzymes, disturbed histone modification patterns or changes in DNA methylation. The Cancer Genome Atlas (TCGA) project profiles thousands of tumors from various cancer types, allowing the discovery of somatic alterations in the epigenetic machinery and the identification of potential cancer drivers among members of epigenetic protein families.
Methods: Here, we integrated the TCGA data from seven cancer types to study somatic mutations, altered expression, and aberrations in chromosomal copy number of epigenetic regulators. We defined oncogene- or tumor suppressor-like genomic profiles within epigenetic families, and applied correlation network analysis to identify co-expressed cancer genes.

Results: We uncovered the oncogene-like expression profiles of EZH2, ATAD2, DNMT3B, SUV39H2, KAT2A and other epigenetic proteins. BRDT exhibits a cancer-testis gene signature with exclusive expression in tumors. Genes with the highest mutation rates in tumors, including PBRM1, SETD2, CREBBP, ASH1L and members of the MLL family, were enriched for damaging alterations and commonly down-regulated in our tumor panel, suggesting tumor suppressor characteristics. Correlation network analysis indicated a high rate of co-expression between frequently mutated epigenetic regulators and known cancer genes. EZH2 was the only epigenetic modifier that showed coregulation in a coexpressed cell cycle regulatory gene network.

Conclusions: Combined expression and copy number analysis revealed a few genes with oncogene- or tumor suppressor like characteristics, and some are further supported by co-expression with cancer genes. EZH2 showed the most significant oncogene-like profile, confirmed by its corregulation with a functional cell cycle module. Such complex genomic patterns revealed by the TCGA data will be important for evaluating epigenetic regulators as therapeutic targets.

QUANTITATIVE EFFECTS OF GENETIC VARIANTS FROM SEQUENCE ALONE

Thomas Hopf*, John Ingraham and Debora Marks*

Harvard University, Boston (MA), USA

email: {thomas_hopf, debbie}@hms.harvard.edu

A critical challenge for present day genomics is the accurate prediction of phenotypic consequences of genetic variation. A large number of diverse computational methods predict the potential impact of different kinds of variants, e.g. coding, non-coding, or copy number, on disease likelihood, disease progression, and drug response. However, current methods are typically measured against categorical data parsed from reports of clinical research and have focused on machine learning from known disease causing variants and/or consideration of single position evolutionary conservation. Here, we present a novel probability model-based approach, EVfitness, that we measure against quantitative experimental measures of mutational effects, and that relies exclusively on sequence information, gleaned from the millions of evolutionary experiments that are implicitly recorded in sequence alignments. The model specifically considers dependencies of different genomic positions (since amino acid positions in proteins are not independent - hence epistasis) and estimates the effect of the function of a protein in vivo by considering its evolutionary couplings with other residue positions, reflecting methods in previous work that predicted protein structure. Our method is surprisingly successful in predicting quantitative changes in diverse protein and organism features, including protein stability, enzyme kinetics, antibiotic resistance and organism fitness.

As proof of principle we measure the EVfitness predictions against large scale mutation experiments, providing the biomedical community with comprehensive, specific predictions not biased by previous experiments.

PANTHER-PSEP: PREDICTING DISEASE-CAUSING MUTATIONS USING POSITION SPECIFIC EVOLUTIONARY PRESERVATION

Haiming Tang* and Paul Thomas*

**University of South California,*

Los Angeles (CA), USA

email: haimingt@usc.edu, pdthomas@med.usc.edu

PANTHER-PSEP is a new software tool for predicting deleterious non-synonymous SNPs. PANTHER-PSEP uses a novel methodology to process and interpret homologous alignments called "evolutionary preservation": homologous proteins are used to reconstruct the likely sequences of ancestral proteins at nodes in a phylogenetic tree, and the history of each amino acid can be traced back in time from its current state to estimate how long that state has been preserved in its ancestors. Here we show that the longer a position in a current human protein has been preserved by tracing back to its direct ancestors, the more likely that a mutation at that site will have a deleterious effect, and that this apparently simple metric outperforms a conventional measure of evolutionary conservation in predicting deleterious variants in humans. The method is completely general, and can also be applied to genetic variants in all 81 other species in PANTHER.

GEOGRAPHIC POPULATION STRUCTURE (GPS): A NOVEL TOOL TO INFER BIOGEOGRAPHICAL ORIGINS OF WORLDWIDE HUMAN POPULATIONS

Tatiana Tatarinova* and Eran Elhaik

*University of South California,
Los Angeles (CA), USA.
email: tatarino@usc.edu

The genetic structure of human populations is of fundamental importance to anthropological, evolutionary, and disease studies and has broad applications to forensic and other medical sciences. One of the most important processes that shaped the genetic structure of populations and the risk for disease is inbreeding or the amount of admixture. Knowledge of admixture can be applicable for biogeography, that is, the inference of one's geographical origins from biological data. Although it is widely known that most populations are admixed to various degrees, the analytical methods that are available to study admixture and infer bio-geography on a genome-wide scale are limited. Because measures of admixture and biogeography are routinely used to infer ancestry, correct for population stratification in Mendelian and complex disease studies, and shed light on our recent history, there is a high demand for improved and accurate methods. Absence of accurate tools to infer recent ancestry hinders genetic research in epidemiology and personalized medicine, where ancestry is an important.

To satisfy this demand, we have developed Geographic Population Structure (GPS) tool, which can accurately measure admixture and infer biogeography in complete-genome data sets that are now practical to generate. Our novel algorithm provides biogeographical predictions which, for many of the tested non-admixed individuals, were accurate to the resolution of the home village up to ~1,000 years ago. GPS can help people seeking answers about their past, professionals trying to match cases and controls in disease studies, and scientists seeking accurate answers of identity and origin.

Selected Posters

VarI-SIG Meeting – ISMB 2014, July 12th Boston (MA), USA

MACHINE LEARNING MODELS FOR SURVIVAL PREDICTION IN CANCER USING GENE EXPRESSION AND STRUCTURE-BASED MUTATIONAL DATA

Kanakadurga Addepalli, Gautam Shankar and Iosif Vaisman*

*George Mason University, Fairfax (VA), USA
email: ivaisman@gmu.edu

The rapid progress of molecular sequencing and characterization technologies has generated a torrent of genomic and proteomic information which can now be used to build predictive models for predicting not only the functional effects of missense mutations but integrate the information further and build integrative models for cancer prognosis and survival prediction. Cancer informatics efforts focused on discovery and validation of biomarkers aid in early diagnosis of cancer and faster and more reliable cancer therapies. Patients with identical molecular, histological and clinical diagnoses and given the same treatments, show varied outcomes. Various bioinformatics models have demonstrated the potential of differentiating gene expression profiles in cancer tissues vs. normal tissues when integrated with other genomic data for molecular diagnosis and survival analysis of human cancers. In this work we describe an approach relying on structure-based models of functional impact of somatic mutations to improve feature selection in survival prediction models trained on gene expression profiles from relatively large cohorts of cancer patients with different disease progression patterns.

INTEGRATED DATABASE OF HUMAN CANCER MISSENSE MUTATIONS CROSS-LINKED TO 3D PROTEIN STRUCTURES

Kanakadurga Addepalli* and Iosif Vaisman*

*George Mason University, Fairfax (VA), USA
email: kaddepal@masonlive.gmu.edu, ivaisman@gmu.edu

Missense mutations in oncogenes are being studied widely as they have been implicated in various cancers. Cancer genome sequencing projects have produced very large amounts of data, which is stored in multiple databases. These databases rely on diverse data formats and in some cases significantly overlap. Although many of these databases provide cross-links to other relevant data sources, in most cases they do not allow to obtain information on three-dimensional structure of proteins. Easy access to three-dimensional structure when it is available facilitates structure-based analysis and modeling of effect of mutations on protein function. Here we present an integrated database of missense mutations found in human cancer genes and genomes (IDHCMM) with a web interface for data display and download in different formats. This is an effort to build a unified repository of missense mutations from six different cancer mutation databases, TCGA, ICGC, COSMIC, BIC, IARC TP53 and Cancer Genomics projects at MSKCC (Prostate cancer, Sarcoma). Apart from integrating data from these data sources, a very important feature of IDHCMM is linking these records to the corresponding 3D structures in PDB. IDHCMM combines the molecular, histological, clinical, sequencing and analytical data into one database. Current version of IDHCMM contains more than 1.48 million records, which include 215374 distinct missense mutations. The database is publicly available for download and access through the web interface.

ONCOGENIC DRIVER HOTSPOT IDENTIFICATION THROUGH LARGE-SCALE CLINICAL GENOMICS AND FUNCTIONAL SCREENING PROGRAMS

Tenghui Chen, Yong Mao, Karina Eterovic, Kenna Shaw, Funda Meric-Bernstam, Gordon Mills and Ken Chen*

MD Anderson Cancer Center, Houston (TX), USA
email: kchen3@mdanderson.org

Background: Identifying driver mutations in individual cancer patient is one of the utmost challenges in realizing genomic medicine. At MD Anderson cancer center, we have developed several routine molecular testing platforms at the Institute of Personalized Cancer Therapy (IPCT), to predict driver mutations from targeted sequencing of thousands of cancer patients. Our results were used to inform the decision of clinicians and to guide the cancer target discovery and development (CTD2). We found that current computational approaches are insufficient in modeling the function of individual mutations in specific cancers and have not sufficiently utilized available genomic and functional data.

Results: We developed a novel statistical approach that accounted for variation of mutation rate, mutation type and sequence context in individual genes and cancer types. We identified 1834 oncogenic hotspots from the COSMIC database. The identified mutational hotspots demonstrated specificity in cancer subtypes and enhanced mutual exclusivity in pathways that are known active in cancer development. Hotpots in critical cancer genes such as TP53 and CDKN2A are strongly associated with differential gene expressions in the cancer genome atlas (TCGA) pan-cancer RNA sequencing data. They also received significantly higher scores than non-hotpot mutations from functional prediction tools such as CanDrA that examine alterations in sequence conservation and protein structure.

Conclusion: Our results indicated that identifying mutational hotspots in specific cancer background through knowledge-informed statistical testing can potentially lead to effective identification of driver mutations. Our efforts will focus further on functionally validating our targets through in vitro screening, iterative improvement of methodology, and application to large-scale clinical sequencing.

A STAGEWISE CLOSED TESTING PROCEDURE FOR DISCOVERING GENETIC INTERACTIONS

Mattias Frånberg*, Karl Gertow, Anders Hamsten, Jens Lagergren and Bengt Sennblad

Karolinska Institutet, Stockholm, Sweden
email: mattias.franberg@gmail.com

Despite the success of genome-wide association studies in general medical genetics, the underlying genetics of many complex diseases remains enigmatic. It is widely believed that this lack of understanding can, in part, be attributed to the failure to statistically account for genetic interactions. However, interactions commonly elude thorough treatment because the vast number of possible interactions imposes hard statistical and computational challenges. The discovery of interactions is further complicated by phenomena such as genetic heterogeneity, population substructure and confounding that often require prohibitively slow algorithms to resolve.

We introduce a new strategy where a set of null hypotheses are tested, in the order of their complexity, against a saturated alternative hypothesis according to a statistical principle called closed testing. This has three distinct advantages. It reduces the computational burden of testing refined definitions of interaction. The principle allows precise control of the family-wise error rate. Lastly, we show that it significantly improves the statistical power without relying on a strong marginal effect or prior information. We have applied this methodology on biological data from the PROCARDIS cohort and discover several plausible interactions related to coronary artery disease.

WHY YOUR REFERENCE GENE SET MATTERS

Adam Frankish[*], Jonathan Mudge
and Jennifer Harrow

[*]*Wellcome Trust Sanger Institute, Hinxton, UK*
email: af2@sanger.ac.uk

McCarthy et al. recently demonstrated the strong influence gene set selection has on the prediction of functional effect of variation data. The GENCODE geneset represents the reference human gene annotation for the ENCODE project and is produced by merging manual annotation and automated Ensembl gene predictions with extensive computational and experimental QC and validation. We will highlight some significant differences between the GENCODE and NCBI Reference Sequence Database (RefSeq) gene sets and illustrate GENCODE's larger and more detailed set of annotated transcripts. Specifically, we will discuss divergence in the annotation of alternative splicing (where, for example, GENCODE protein-coding loci have a mean of 7.6 alternatively-spliced transcripts while RefSeq only have 2.1), long non-coding RNAs, pseudogenes, genomic coverage of annotated exons, degree of manual curation, experimental validation, and functionally descriptive biotypes. We will detail the continued extension and refinement of the GENCODE geneset, including the integration of RNAseq, CAGE, polyAseq, ribosome profiling and epigenomic data, to identify novel loci, define 5' and 3' transcript boundaries and identify novel translation initiation sites. Finally, we will describe the filtering options provided to allow the user to reduce complexity of the GENCODE gene set and explain our investigation of RNAseq data to determine the expression level of GENCODE-annotated genes, transcripts and exons, to present a smaller, but biologically meaningful, set of features e.g. only those expressed in a particular tissue.

VARANT: AN OPEN SOURCE TOOL FOR HUMAN GENETIC VARIANT ANNOTATION

Kunal Kundu, Uma Sunderam, Steven E Brenner[*] and
Rajgopal Srinivasan[§]

[*]*University of California, Berkeley (CA) and*
[§]*Innovation Labs, Tata Consultancy Services,*
Hyderabad (AP), India
email: brenner@compbio.berkeley.edu,
raj@atc.tcs.com

Genome sequencing technologies are generating an abundant human genetic variation data. In a clinical study involving such genetic variants, generally the aim is to distinguish causative variants from the millions of variants present in a single human genome. Thus a challenge lies in interpreting the functional relevance of such variations in order to facilitate the distillation of these to a narrower set of more relevent variants for further investigation. Comprehensive annotation of variants is a necessary first step in arriving at a small subset of variants that are most likely to explain the phenotype(s) under investigation. Some of the most frequently employed annotations include: region of occurrence (intron, exon, or intergenic), degree of conservation, inheritance patterns amongst related individuals, minor allele frequency, predicted impact on protein function, and previously established association with phenotype of interest. Several tools already exist for this purpose, each with their strengths and weaknesses in terms of type of variants they annotate, annotation features, type of interface – whether stand alone or web service, and license for using the software. We have developed a comprehensive and extensible open source tool for human genome variation annotation called VARANT, written in the Python programming language. We believe that VARANT distinguishes itself by being fully open source, capable of using multiple threads for swift annotation, and providing extensive annotation of UTR and non-coding regions in addition to the customary annotations of genes. We have carefully benchmarked VARANT to ensure that it has essentially all features present in other widespread tools, and that it makes no errors with respect to the other tools

DISTANCE TO SPLICE SITES IS A DEFINING FEATURE FOR DISEASE-RELATED INTRONIC SNPS

Meng Ma and Rong Chen*

*Mount Sinai School of Medicine,
New York (NY), USA
email: rong.chen@mssm.edu

It is well known that many SNPs have a tight association with diseases. Modifying splicing signals and causing splicing dysfunction is an important pathogenic mechanism for many SNPs. A lot of work has been carried out to study the functional effect of exonic SNPs on splicing, for example Human Splicing Finder. But we are still not clear what genomic features determine the effect of the SNPs in noncoding region on splicing process and how to discriminate the disease-related intronic SNPs from neutral intronic SNPs. To make clear these questions, we collected two SNPs datasets: disease-related intronic SNPs and neutral intronic SNPs. Through statistical analysis, we found that the SNPs from the two datasets are with the similar possibility to modify the intronic splicing enhancers (ISEs) or silencers (ISSs). A rational inference is that the splicing signal change by the disease-related intronic SNPs successfully causes a functional effect on splicing process, but not for neutral intronic SNPs. Intronic splicing signals are abundant where they function positively and infrequent where they are inhibitory, hence naturally inferred the disease-related intronic SNPs locate in the positive ISE/ISS region but neutral intronic SNPs just involve the inhibitory ISE/ISS region. Our experiment shows that ISE/ISS are positional dependent in the proximity of authentic splice sites. Correspondingly, more disease-related intronic SNPs have a shorter distance to authentic spice sites compared with neutral intronic SNPs. Therefore distance to splice sites may be a defining feature to discriminate disease-related intronic SNPs from neutral intronic SNPs.

MODELING PATHWAY-LEVEL EFFECTS: THE IMPACT OF PATHWAY SIZE, VARIABLE SELECTION, AND PERMUTATION TESTING

Michael Mooney*, Shannon McWeeney, Joel Nigg and Beth Wilmot

*Oregon Health & Science University,
Portland (OR), USA
email: mooneymi@ohsu.edu

Pathway representations can provide a functional context for genomic variants associated with complex disease. While multiple approaches for calculating pathway-level association measures exist, each makes different assumptions about the way in which SNP-level effects combine or interact to produce a pathway-level effect. To date there is no consensus about the best way to test for significant associations between a pathway (gene set) and a trait of interest.

The widely used enrichment-type methods are dependent on individual SNPs having independent main effects, because they first calculate individual SNP association measures and then combine these individual effects to calculate a pathway-level association measure. Our hypothesis is that methods utilizing multi-variant models to calculate gene- or pathway-level effects better represent the underlying genetic complexity.

We therefore have examined three methods that utilize the original genotype data as input, rather than individual SNP p-values, to test for association between pathways and ADHD in a large GWAS dataset from the Psychiatric Genomics Consortium. Each method uses a different variable selection and modeling procedure to evaluate pathway association.

Preliminary results from these methods have shown that pathway significance is highly correlated with pathway size (i.e. the number of SNPs mapped to the genes in the pathway), and that permutation tests that randomly permute the phenotype are inadequate to adjust for this bias. We will present the differences in performance and the concordance of results across these methods, as well as the ability of a pathway resampling procedure to correct for pathway size.

COMPUTATIONAL PRIORITIZATION OF PHENOTYPE ASSOCIATED VARIANTS

Kymberleigh Pagel, Yuxiang Jiang, Vikas Pejaver, Sean Mooney and Predrag Radivojac*

*Indiana University Bloomington, Bloomington (IN), USA
email: predrag@indiana.edu

Advances in sequencing technologies have generated a wealth of potentially important variants with uncertain phenotypic implications that have shown promise for the identification of novel variant-phenotype relationships. Traditional methods of identifying relationships between genetic variation and phenotypes have been successful, yet can be time consuming. Genome wide association studies tend to identify common variants with modest contributions to a phenotype and can miss rare variants which are more likely to be causal. The ability of exome sequencing to detect rare causal variants in protein-coding regions has already yielded success in the case of several Mendelian disorders. However, the applications to complex diseases have only recently been realized. A major challenge lies in the prioritization of candidate variants in decreasing order of their putative contribution to a given disease. To address this issue we develop a novel computational framework to systematically combine predictions from MutPred, a tool that predicts the propensity of an amino acid substitution to cause disease, and PhenoPred, a method to infer gene-disease associations using biological networks, for the simultaneous prioritization of putative causal variants and prediction of the resulting phenotype. Specifically, we concentrate on the application of this method to predict status for variants associated with both Mendelian and complex traits with minimal prior knowledge and small sample sizes. The method performed well on exome sequencing data sets for familial combined hyperlipidemia (FCH), hypoalphalipoproteinemia (HA) and Crohn's disease as part of the 2013 Critical Assessment of Genome Interpretation (CAGI) conference.

INSIGHTS FROM GWAS: EMERGING LANDSCAPE OF MECHANISMS UNDERLYING COMPLEX TRAIT DISEASE

Lipika R. Pal, Chen-Hsin Yu, Stephen M. Mount and John Moult*

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

There are now over 1600 loci in the human genome where GWAS has shown the presence of one or more SNPs to be associated with altered risk of a complex trait disease. In each of these loci, there must be some affected molecular level mechanism relevant to the disease. What are these mechanisms and how do they contribute to disease? Here we examine the role of three primary mechanisms: changes to the level of expression of a gene, changed protein function caused by a missense SNP, and effects on message splicing. Associations found in the Wellcome Trust Case Control Consortium (WTCCC1) seven disease GWA study and other related meta analyses and follow-up studies, collected from the GWAS catalog, were used. Hapmap data, 1000genome data, and linkage disequilibrium information were used to identify possible candidate SNPs associated with increased disease risk at each of 356 loci for the seven diseases. We find that 34% of these loci have at least one predicted high impact missense SNP, 60% of have at least one eQTL relationship affecting gene expression level and 37% have at least one splicing effect. Together these results provide candidate mechanisms for 70% of loci involved in these diseases. Each of these putative mechanisms provides a hypothesis for further investigation. Examination of these results shows a wide variety of proteins are implicated, some well established to play a role of in the corresponding diseases, some supporting previous suggestions, and some novel.

VARMOD: MODELING THE FUNCTIONAL EFFECTS OF NON-SYNONYMOUS VARIANTS

Morena Pappalardo* and Mark N. Wass

*University of Kent, Canterbury, UK
email: mp465@kent.ac.uk

Unravelling the genotype-phenotype relationship in human is still one of the most challenging tasks in genomics studies. Recent advances in sequencing technologies have revealed millions of single nucleotide variants (SNVs), whereas to date they only describe a small proportion of heritability. It is now important to develop methods that can identify those variants that are functional and also predict those protein functions that can result in altered phenotypes. In this poster, we present VarMod a novel method for investigating the functional effects of non synonymous single nucleotide variants (nsSNVs) in proteins. VarMod identifies ligand binding and protein-protein interface sites in the query protein and considers the distance of nsSNVs to these functional sites. VarMod combines these features with other widely used features for predicting if nsSNVs alter protein function; these features include: residue conservation, amino acid properties and structural features such as solvent accessibility and secondary structure properties. The features are combined using a support vector machine to make an overall prediction of the nsSNVs that are likely to have an effect on the protein function. VarMod is available as a web server and provides extensive features to visually analyse the protein model and the location of the nsSNVs occurring at ligand binding and/or protein-protein interface sites. In benchmarking on a set of pathogenic and neutral nsSNVs from VariBench, VarMod outperforms PolyPhen.

MUTPRED2: PREDICTING THE PATHOGENICITY, STRUCTURAL AND FUNCTIONAL CONSEQUENCES OF MISSENSE VARIANTS

Vikas Pejaver, Kymberleigh Pagel, Sean Mooney and Predrag Radivojac*

*Indiana University Bloomington,
Bloomington (IN), USA
email: predrag@indiana.edu

The increasing application of high-throughput sequencing to individual-level and population-level studies has resulted in a deluge of genetic variation data. The prioritization of variants that are relevant to and/or cause disease remains a major challenge. This challenge has been partly addressed through the development of a variety of computational prediction methods. However, current methods provide little information on the structural and functional consequences of predicted deleterious variants in coding regions. To this end, we developed MutPred2, a tool for the prediction of pathogenicity of amino acid substitutions and their resulting molecular effects. MutPred2 differs from the original MutPred model in four ways. First, it was trained on a larger set of variants obtained from a wider set of data sources. Second, MutPred2 uses a stack-like model that combines the high overall accuracy of random forests with the robustness of neural networks, resulting in real value predictions that can be better interpreted as posterior probabilities. Third, the repertoire of structural and functional changes covered has been expanded to include several new PTM sites, metal-binding sites and DNA- and protein-binding sites. Finally, the incorporation of in-house predictors for these features allows for the development of an easily installable and usable standalone program with the ability to add more molecular mechanisms in future updates. Preliminary results indicate that MutPred2 achieves an AUC of 88% and can generate testable hypotheses for up to 47 possible structural and functional mechanisms.

A VARIANT AND GENE PRIORITIZATION PIPELINE FOR IDENTIFYING CAUSAL VARIANTS IN INHERITED RARE DISORDERS

Sadhna Rana, Ajithavalli Chellappan, Uma Sunderam, Kunal Kundu, Jennifer M Puck, Steven E. Brenner[*] and Rajgopal Srinivasan[§]

[*]*University of California, Berkeley (CA) and*
[§]*Innovation Labs, Tata Consultancy Services, Hyderabad (AP), India*
email: brenner@compbio.berkeley.edu, raj@atc.tcs.com

We have developed a pipeline for the analysis of genomic variant data, and applied it to numerous clinical cases. The first several steps of the pipeline employ standard tool for mapping and variant calling. However, we have integrated three different callers, each carefully tuned, to yield high quality sets of variants. Quality metrics for mapping, gene coverage in sequencing and called variants are generated throughout the run. Our VARANT tool provides extensive variant annotations. Custom tools and flexible filtering options support interactive analysis of data, enabling users to focus on a set of potential disease causative variants in the study of rare disorders likely to be monogenic. We have assembled a series of modules to further prioritize shortlisted gene variants. The pipeline, written in Python, is easily extensible. Multiple samples can be analyzed in parallel, including multisample calling with a library of genomes. The rich variant annotation, gene prioritization and data visualization modules reduce the millions of variants found in samples down to a manageable short list of possible causative variants for a given phenotype. Three case studies of interesting findings are presented in which the pipeline was applied to exome analysis of families with immune system disorders.

PRIME SUSPECT: DISEASE-SPECIFIC VARIANT PRIORITISATION USING NETWORK INFORMATION

Christopher Yates[*] and Michael Sternberg

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

We have previously developed SuSPect, a method for predicting the phenotypic effects of single amino acid variants (SAVs) in human disease (www.sbg.bio.ic.ac.uk/suspect). SuSPect incorporates sequence-conservation and protein-protein interaction network features to give enhanced prediction compared to other tested methods, with an AUC of up to 0.90 on a test set of over 18,000 SAVs.

Like other SAV scoring methods, SuSPect only gives a score based on whether or not a variant is likely to cause disease, but in many cases a user is only interested in a single disease. PriMe SuSPect (Prioritisation Method using SuSPect) has been developed to meet this need, using a random walk with restart on PPI and domain-domain interaction networks to associate SAVs with specific diseases. The protein- and domain-based scoring is adapted from the PRINCE method developed by Vanunu et al. (2010).

To test this method, known disease-causing SAVs were 'spiked' into exomes from the 1000 Genomes Project. Using just SuSPect scores, in a little over 1% of cases the causative variant was selected as the top candidate. By incorporating protein- and domain-specific features for the disease of interest, PriMe SuSPect was able to select the correct variant up to 66% of the time, with up to 76% of the correct SAVs ranked in the top 10.

PriMe SuSPect will be incorporated into the SuSPect web-server, enabling users to upload data from an exome sequencing project and rank the SAVs identified therein for a specific disease of interest. We consider that by offering disease-specific SAV scoring, PriMe SuSPect could become a valuable tool in the identification of causative variants from exome sequencing studies.

ACKNOWLEDGMENTS

The VarI-SIG meeting organizers would like to acknowledge:

- George Church, Harvard University, Boston (MA), USA
- Garry Cutting**,** John Hopkins University, Baltimore (MD), USA
- Rachel Karchin**,** John Hopkins University, Baltimore (MD), USA
- Jan Korbel, EMBL, Heidelberg, Germany
- James Potash, University of Iowa, Iowa City (IA), USA
- Benjamin Raphael**,** Brown University, Providence (RI), USA

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

AUTHOR INDEX

Addepalli, Kanakadurga 11
Ayers, Kristin 7
Antipin, Yevgeniy 7

Brenner, Steven E 13,17
Brodie, Aharon 7

Casadio, Rita 8
Chellappan, Ajithavalli 17
Chen, Ken 12
Chen, Rong 7,14
Chen, Tenghui 12
Church, George M 4
Cutting, Garry 4

Di Lena, Pietro 8

Elhaik, Eran 10
Eterovic, Karina 12

Fariselli, Piero 8
Frånberg, Mattias 12
Frankish, Adam 13

Gertow, Karl 12
Gnad, Florian 8

Hamsten, Anders 12
Harrow, Jennifer 13
Hopf, Thomas 9

Ingraham, John 9

Jiang, Yuxiang 15

Korbel, Jan 5
Kundu, Kunal 13,17

Lagergren, Jens 12

Ma, Meng 14
Mao, Yong 12
Marks, Debora 9
Martelli, Pier Luigi 8
McWeeney, Shannon 14
Meric-Bernstam, Funda 12
Mills, Gordon 12
Mooney, Michael 14
Mooney, Sean 15,16
Moult, John 15
Mount, Stephen M 15
Mudge, Jonathan 13

Nigg, Joel 14

Pagel, Kymberleigh 15,16
Pal, Lipika 15
Pappalardo, Morena 16
Pejaver, Vikas 15,16
Potash, James B 5
Puck, Jennifer M 17

Ofran, Yanay 7

Radivojac, Predrag 15,16
Rana, Sadhna 13
Raphael, Benjamin J 6
Reva, Boris 7

Sennblad, Bengt 12
Shankar, Gautam 11
Shaw, Kenna 12
Srinivasan, Rajgopal 13,17
Sternberg, Michael JE 17
Sunderam, Uma 13,17

Tang, Haiming 10
Tatarinova, Tatiana 10
Thomas, Paul 10

Vaisman, Iosif 11

Wass, Mark N. 16
Wilmot, Beth 14

Yates, Christopher 17
Yu, Chen-Hsin 15

Zhang, Zemin 8