Varl special interest group

Varl-SIG Meeting

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

ISMB/ECCB 2015
July 11th 2015, Dublin, Ireland

Room Liffey Meeting 2, Convention Center

http://varisig.biofold.org/

Invited Speakers

Søren Brunak
Technical University of Denmark, Lyngby (Denmark).
Multi-morbidities and disease trajectories from data mining of 6 million electronic patient records.

Brendan Frey
University of Toronto, Toronto (Canada)
A new approach to variant interpretation and prioritization.

Nuria Lopez-Bigas
University Pompeu Fabra, Barcelona (Spain)
Analyzing thousands of tumor genomes to identify cancer drivers and their targeted therapeutic opportunities.

Yves Moreau
KU Leuven, Leuven (Belgium).
Variant prioritization by genomic data fusion.

Joris Veltman
Radboud University, Nijmegen (Netherlands).
Large-scale exome and genome sequencing in genetic disease; Impact for research and diagnostics.

VarI-SIG Organizers

Yana Bromberg, Rutgers University, New Brunswick (NJ), USA
Emidio Capriotti, University of Alabama at Birmingham, Birmingham (AL), USA
Hannah Carter, University of California, San Diego, La Jolla (CA), USA.

VarI-SIG Meeting Programme - July 11th 2015, Dublin, Ireland.

08:10 – 08:20 Welcome from the committee

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

08:20 – 09:10 **Highlight Speaker: Nuria Lopez-Bigas**. University Pompeu Fabra, Barcelona (Spain)
Analyzing thousands of tumor genomes to identify cancer drivers and their targeted therapeutic opportunities.

09:10 – 09:35 **Ivan Kulakovskiy.** Vavilov Institute of General Genetics, Moscow (Russia)
Sequence analysis of regulatory variants reveals selection pressure on somatic mutations in breast cancer.

09:35 – 10:00 **Tae Hyun Hwang**. University of Texas Southwestern. Dallas, TX (USA)
iSHEAR: An efficient computational tool to detect genetic variants and estimate intratumor heterogeneity linked with cancer development, progression, and therapy resistance.

10:00 – 10:25 **Jaroslav Bendl**. Masaryk University, Brno (Czech Republic).
PredictSNP 2.0: A unified platform for prediction of disease-related mutations in entire human genome.

10:25 – 10:45 **Coffee Break**

10:45 – 11:00 **Short Talk: Mark Roger.** University of Bristol, Bristol (UK)
Sequential data selection for predicting the pathogenic effects of sequence variation.

11:00 – 11:15 **Short Talk: Ronald Hause.** University of Washington, Seattle, WA (USA).
Envision: leveraging large-scale protein mutagenesis data for variant effect prediction.

11:15 – 12:05 **Highlight Speaker: Søren Brunak**. Technical University of Denmark, Lyngby (Denmark).
Multi-morbidities and disease trajectories from data mining of 6 million electronic patient records.

12:05 – 12:20 **Company Presentation: Alex Kaplun. QIAGEN**.
PGMD: a comprehensive manually curated pharmacogenomic database.

12:20 – 13:00 **Lunch Break and Poster Session with the Authors**

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

13:10 – 14:00 **Highlight Speaker: Brendan Frey.** University of Toronto, Toronto (Canada).
A new approach to variant interpretation and prioritization.

14:00 – 14:25 **Pier Luigi Martelli.** University of Bologna, Bologna (Italy)
The effect of OMIM disease-related variations on protein stability: a large-scale investigation.

14:25 – 14:50 **Alex Cornish.** Imperial College London, London (UK).
Exploring the cellular basis of human disease through a large-scale mapping of deleterious genes to cell types.

14:50 – 15:05 **Short Talk: Eran Elhaik.** University of Sheffield, Sheffild (UK).
Inferring the biogeographical origin of Druze with the Geographic Population Structure (GPS).

15:05 – 15:20 **Short Talk: Steven Brenner.** University of California, Berkeley, CA (USA)
Diagnostic role of exome sequencing in immune deficiency disorders.

15:20 – 15:40 **Coffee Break**

15:40 – 15:45 **Special Session**

15:45 – 16:30 **Highlight Speaker: Joris Veltman**. Radboud University, Nijmegen (Netherlands).
Large-scale exome and genome sequencing in genetic disease; Impact for research and diagnostics.

16:30 – 17:15 **Highlight Speaker: Yves Moreau.** KU Leuven, Leuven (Belgium).
Variant prioritization by genomic data fusion.

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

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

Invited Presentations

MULTI-MORBIDITIES AND DISEASE TRAJECTORIES FROM DATA MINING OF 6 MILLION ELECTRONIC PATIENT RECORDS

Søren Brunak

Technical University of Denmark, Lyngby Denmark.
email: brunak@cbs.dtu.dk

Electronic patient records remain a rather unexplored, but potentially rich data source for discovering correlations between diseases, drugs and genetic information in individual patients. Such data makes it possible to compute fine-grained disease co-occurrence statistics, and to link the comorbidities to the treatment history of the patients. A fundamental issue is to resolve whether specific adverse drug reaction stem from variation in the individual genome of a patient, from drug/environment cocktail effects, or both. Here it is essential to perform temporal analysis of the records for identification of ADRs directly from the free text narratives describing patient disease trajectories over time. ADR profiles of approved drugs can then be constructed using drug-ADR networks, or alternatively patients can be stratified from their ADR profiles and compared. Given the availability of longitudinal data covering long periods of time we can extend the temporal analysis to become more life-course oriented. We describe how the use of an unbiased, national registry covering 6.2 million people from Denmark can be used to construct disease trajectories which describe the relative risk of diseases following one another over time. We show how one can “condense” millions of trajectories into a smaller set which reflect the most frequent and most populated ones. This set of trajectories then represent a temporal diseaseome as opposed to a static one computed from non-directional comorbidities only.

A NEW APPROACH TO VARIANT INTERPRETATION AND PRIORITIZATION.

Brendan Frey

University of Toronto, Toronto, Canada
email: frey@psi.toronto.edu

Abstract not available.

ANALYZING THOUSANDS OF TUMOR GENOMES TO IDENTIFY CANCER DRIVERS AND THEIR TARGETED THERAPEUTIC OPPORTUNITIES

Nuria Lopez-Bigas

University Pompeu Fabra, Barcelona, Spain
email: nuria.lopez@upf.edu

Large efforts dedicated to sequence thousands of tumor genome/exomes are expected to lead to significant improvements of precision cancer medicine. However, high inter-tumor heterogeneity is a major obstacle in the road to develop an arsenal of targeted cancer drugs to treat most cancer patients. Therefore, it is critical to understand the current scope of anti-cancer targeted drugs for different tumor types in order to use them with the highest efficacy, and to define priorities for the development of new ones. We have developed a novel methodology to interpret the genomes of a cohort of tumor samples and to assess their therapeutic opportunities. Starting with somatic mutations detected across the cohort, the methodology identifies the driver genes, highlights those that dominate the clonal landscape of the tumors and determines their mode of action. It then does an in-silico prescription of approved and candidate targeted drugs to each patient in the cohort. The application of this approach to a cohort of 6795 cancer samples of 28 different tumor types showed that the fraction of patients that could benefit from prescribed FDA-approved drugs is strikingly small. Nevertheless, it improves significantly if repurposing opportunities are taken into consideration, with large differences between tumor types. In addition, we identify 80 therapeutically unexploited cancer genes, tightly bound by pre-clinical small molecules or potentially suitable for molecule binding. The resource created with this analysis is also intended to provide interpretation of newly sequenced cancer genomes and to design pan-cancer and tumor type specific sequencing panels for efficient early cancer detection and clinical insight.

VARIANT PRIORITIZATION BY GENOMIC DATA FUSION

Yves Moreau

KU Leuven, Leuven, Belgium
email: moreau@esat.kuleuven.be

NGS has rapidly increased our ability to discover the cause of many previously unresolved rare monogenic disorders by sequencing rare exomic variation. However, after standard filtering against nonsynonymous single nucleotide variants (nSNVs) and loss-of-function mutations that are not present in healthy populations or unaffected samples, many potential candidate mutations are often retained and we need predictive methods to prioritize variants for further validation. Several computational methods have been proposed that take into account biochemical, evolutionary and structural properties of mutations to assess their potential deleteriousness. However, most of these methods suffer from high false positive rates when predicting the impact of rare nSNVs. A plausible explanation for this poor performance is that many of these predicted variants are mildly deleterious, but in no way specific to the disease of interest. We therefore propose a genomic data fusion methodology that integrates multiple strategies to detect deleteriousness of mutations and prioritizes them in a phenotype-specific manner. A key innovation is that we incorporate into our strategy a computational method for gene prioritization, which scores mutated genes based on their similarity to known disease genes by fusing heterogeneous genomic information. We also integrate haploinsufficiency prediction scores that predict the probability that the function of a gene is affected if present in a functionally haploid state. To integrate or fuse these data sources, we develop a machine-learning model using the Human Genome Mutation Database (HGMD) of human disease-causing mutations compared to three control sets: common polymorphisms and two independent sets of rare variation. Benchmarking on HGMD demonstrates that this integrative phenotype-specific variant prioritization significantly outperforms state-of-the-art predictors, such as SIFT or PolyPhen-2.

LARGE-SCALE EXOME AND GENOME SEQUENCING IN GENETIC DISEASE; IMPACT FOR RESEARCH AND DIAGNOSTICS

Joris Veltman

Radboud University, Nijmegen, Netherlands
email: Joris.Veltman@radboudumc.nl

Rapid developments in genomics technologies now allow us to sequence all genes (the exome) or even the entire genome of thousands of patients in research and diagnostics. This is completely changing the way genetics studies are done, taking away the major bottleneck of genomic variation detection (although as I will show there are still some major challenges in this for different types of variation). In this presentation I will explain our research on and diagnostics of genetic disorders and focus on the major remaining bottleneck; Interpretation of the enormous amount of variation present in individual genomes in the context of a clinically heterogeneous phenotype. Solving this will require a concerted clinical, biological and bioinformatics approach, resulting amongst others in international agreement on phenotype ontologies, sharing of clinical and genomic data, optimization of variant interpretation tools and the validation of these using relevant biological models. I will illustrate all of this using severe intellectual disability as a model, for which we are making rapid progress and now have the opportunity to provide medically relevant information to the majority of patients and families involved.

Selected Presentations

VarI-SIG Meeting – ISMB/ECCB 2015, July 11th Dublin, Ireland

PREDICTSNP 2.0: A UNIFIED PLATFORM FOR PREDICTION OF DISEASE-RELATED MUTATIONS IN ENTIRE HUMAN GENOME

Jaroslav Bendl, Jan Štourač, Miloš Musil, Jaroslav Zendulka, Jiří Damborský and Jan Brezovský*

*Masaryk University, Brno
New York (NY), Czech Republic.
email: brezovsky@mail.muni.cz*

Understanding the functional impact of genetic variations that play a key role in the disease development is one of the objectives of the human genetics and personalized medicine. The difficulty of interpretation of noncoding variants caused that the main attention has been paid to protein coding regions constituting only 1-2% of the genome. Although several tools aiming at predictions on whole genome have been published recently, they report performance metrics measured on inhomogeneous datasets compiled from highly unequal number of mutations from each genome region. As a result, the performance is skewed by the regions with the highest number of representatives in variation databases.

In this study, we constructed seven balanced datasets composed of mutations covering similar genomic regions. These datasets were used for the evaluation of five tools: CADD, GWAVA, FATHMM, MutationTaster2, and SIFT-DNA. This analysis revealed large differences in the accuracy of prediction tools in different regions. Developed classifier PredictSNP 2.0 combines these results into a single consensus predictions trained independently for each genomic region. Special attention was paid to non-synonymous mutations in coding regions, where consensus is enhanced by predictions of state-of-the-art tools on amino acid level: MAPP, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT, and SNAP. This led to improved performance in comparison to the best individual tools. Web interface of PredictSNP 2.0 will enable an easy access to the consensus and all integrated tools providing their unified confidences derived from their performance on the datasets. Upon completion, it will be embedded into previously developed platform available at http://loschmidt.chemi.muni.cz/predictsnp.

EXPLORING THE CELLULAR BASIS OF HUMAN DISEASE THROUGH A LARGE-SCALE MAPPING OF DELETERIOUS GENES TO CELL TYPES

Alex Cornish*, Ioannis Filippis, Alessia David and Michael Sternberg

*Imperial College London, London, UK.
email: a.cornish12@imperial.ac.uk*

While the majority of diseases are manifested within a specific anatomical structure, known disease-associated alleles are often inherited and therefore present throughout the body. Understanding how these ubiquitous alleles produce localized disease is key to understanding the mechanisms that drive disease. We have developed a novel approach, called gene set compactness (GSC), that contrasts the relative positions of disease-associated genes on cell type-specific interactomes to identify the cell types most likely to be affected by the alleles. Cell type-specific interactomes were created through the integration of protein-protein interaction (PPI) data and cell type-specific expression data from the FANTOM5 project. We conducted text-mining of the PubMed database to produce an independent map of disease-associated cell types, which we used to validate our method. Our method identifies previously-suggested associations, along with associations that warrant further study. This includes mast cells and multiple sclerosis (MS); a population of cells that is currently being targeted in an MS phase 2 clinical trial. Furthermore, we used the associations identified by our method to construct a pathogenic cell type-based diseasome, offering insight into diseases linked by common etiology. The dataset produced represents the first large-scale mapping of diseases to their pathogenic cell types. Overall, we demonstrate that the GSC method links disease-associated genes to the phenotypes they produce; one of the key goals of systems biology.

ISHEAR: AN EFFICIENT COMPUTATIONAL TOOL TO DETECT GENETIC VARIANTS AND ESTIMATE INTRATUMOR HETEROGENEITY LINKED WITH CANCER DEVELOPMENT, PROGRESSION, AND THERAPY RESISTANCE

Sean Landman, Michael Steinbach, Vipin Kumar, Scott Dehm, Kevin Silverstein and Tae Hyun Hwang*

*University of Texas Southwestern, Dallas (TX), USA.
email: hwang071@umn.edu

Recent genomic sequencing of cancer genomes has revealed extensive heterogeneity within individual tumors and reported that subclones carrying distinct genetic alterations are present at low frequencies within a tumor and often drive tumor progression and therapy resistance. Therefore, identification of the critical subclonal populations that contribute to tumor initiation, tumor progression and response to therapy is urgently needed. Such studies will facilitate the development of novel diagnostic strategies to stratify patients into subgroups based on their likelihood of responding to therapy and can improve understanding of the role of tumor heterogeneity and develop effective therapy in selected patients. However, a challenge remains in accurately identifying subclones, due to the diverse pattern of genetic heterogeneity in a tumor. Here we developed an efficient computational tool for precise identification of somatic alterations in subclones that are present at low frequencies in the tumor. Our approach is the first tool to simultaneously identify and estimate heterogeneity of subclones based on single nucleotide variants (SNVs), small indels, and structural variations (SVs), and provides a personalized reference genome. We applied our method to more than 300 lung, prostate, gastric, kidney cancer cell lines, and one matched normal, primary and liver metastatic lung cancer, one matched normal, primary, secondary, tertiary osteosarcoma. We found and validated genetic alterations present in subclones that are responsible for tumor progression and therapy resistance. These findings will delineate the mechanisms underlying tumor progression and therapy resistance, thus could lead to designing optimal therapy by predicting and preventing resistance mechanisms.

THE EFFECT OF OMIM DISEASE-RELATED VARIATIONS ON PROTEIN STABILITY: A LARGE-SCALE INVESTIGATION

Pier Luigi Martelli*, Piero Fariselli, Giulia Babbi and Rita Casadio*.

*University of Bologna, Bologna, Italy.
email: {gigi,casadio}@biocomp.unibo.it

Modern genomic techniques allow the association of several Mendelian human diseases to single residue variations in proteins. However, molecular mechanisms explaining the relationship among genotype and phenotype, such as the role of protein variant in/stability, are still under debate.

Here we implement a consensus predictor of our two state of the art methods (I-Mutant3 and INPS) predicting the Gibbs free energy change of a protein sequence upon variation and classify variations as perturbing or non-perturbing protein stability based on a threshold DDG absolute value ≤1kcal/mole. We then filter some 25831 OMIM-related single residue variations (OMIM set) and 8089 polymorphisms without clinical consequence (POLY set) derived from 2255 disease-involved proteins. We find that the fraction of perturbing variations significantly differs in the two sets (58% and 34% in the OMIM and POLY sets, respectively). Considering that our combined predictor scores with a false positive rate for the non-perturbing class of 14%, we find that a significant fraction of OMIM disease-related variations are labeled as non-perturbing. We analyze the OMIM set with respect to the disease class (MEDGEN, Disease Ontology and MalaCards), the protein function (GO, KEGG, REACTOME) and the tissue-specific protein/RNA expression and we detect some significant over-representation of non-perturbing variations in extracellular proteins related to bone and cardiovascular diseases and highly expressed in muscle and skin.

SEQUENCE ANALYSIS OF REGULATORY VARIANTS REVEALS SELECTION PRESSURE ON SOMATIC MUTATIONS IN BREAST CANCER

Ilya E. Vorontsov, Ivan V. Kulakovskiy*, Grigory Khimulya, Darya Nikolaeva and Vsevolod J. Makeev.

*Engelhardt Institute of Molecular Biology, Moscow, Russian Federation.

email: ivan.kulakovskiy@gmail.com

Using wide collection of transcription factor binding sites models, we predicted transcription factor binding sites overlapping somatic mutations in breast cancer. By careful statistical assessment, we found that for a number of transcription factors the observed frequencies of mutations in binding sites significantly differed from the expected. Among transcription factors with binding sites significantly affected by somatic mutations we found several regulators of adipogenesis. Furthermore, somatic mutations in transcription factor binding sites were predominantly found at particular positions of the respective binding motifs. We believe this demonstrates that transcription factor binding sites are an important factor of selection pressure in cancer cell lineages.

Short Presentations

VarI-SIG Meeting – ISMB/ECCB 2015, July 11th Dublin, Ireland

DIAGNOSTIC ROLE OF EXOME SEQUENCING IN IMMUNE DEFICIENCY DISORDERS

Steven Brenner*

*University of California, Berkeley, Berkeley (CA), USA.
email: brenner@compbio.berkeley.edu

To interpret genomic variant data, we developed an analysis protocol whose distinctive features enabled solving numerous clinical cases. The first steps were mapping and variant calling. To yield high quality sets of variants, we used multiple callers and employed multisample calling. The pipeline integrates variant annotation, variant filtering, and gene prioritization to prioritize the millions of called variants to a manageable shortlist of possible causative variants. We applied our analysis protocol to exome sequences from patients with undiagnosed primary immune disorders. A particular focus has been infants who screened positive for absent or low T cell receptor excision circles (TRECs), a marker for T-cells, at birth. We discuss cases with immune deficiencies including severe combined immunodeficiency syndrome (SCID), Nijmegen breakage syndrome (NBS), and ataxia telangiectasia (AT). Early detection provides information to offer patient guidance, family genetic counseling, as well as avoiding the diagnostic odyssey. We also discuss an autoimmune syndrome in which immunological studies were unrevealing, but exome analysis revealed compound heterozygosity for novel hypomorphic and activating mutations of ZAP70. All the variants identified by our analysis protocol were confirmed with Sanger sequencing and validated by immunological studies. These case studies highlight unique features of the analysis framework that facilitate genetic discovery using deep sequencing.

INFERRING THE BIOGEOGRAPHICAL ORIGIN OF DRUZE WITH THE GEOGRAPHIC POPULATION STRUCTURE (GPS)

Eran Elhaik*

*The University of Sheffield, Sheffild, UK.
email: e.elhaik@sheffield.ac.uk

The Druze are an ethic minority in the Levant region. According to the Druze doctrine, intermarriage with non-Druze people or conversion into or away from their religion are not allowed. Because consanguineous marriage is relatively common within the Druze, there have increased risks of autosomal recessive and genetic disorders. The Druze oral tradition described that their adherents came from various regions in 11 A.D, however, the historical records of Druze implies multiple origins, including Egypt, Iran, and the Caucasus. A runs of homozygosity (ROH) analysis confirmed that the Druze are highly related, suggesting a single origin for the community. To reconcile the different hypotheses, we adopted a genetic admixture-based approach to identify the geographical origins of Druze. A proof of concept analysis showed that 91% and 77% of Eurasian populations were predicted within 500km and 200km from their countries of origin, respectively. Applied to 67 Israeli and Lebanese Druze the Geographic Population Structure (GPS) tool indicated that Druze have originated along the northern shores of Lake Van between what was Persia and Mesopotamia, ~1,000 km away from their modern-day residence. A northwest to southeast gradient from Armenia to Turkey indicates the direction of the 11th century migration. Our results suggest that Druze have originated from Irano-Turkish tribes residing in Armenia during the early Middle Age period.

ENVISION: LEVERAGING LARGE-SCALE PROTEIN MUTAGENESIS DATA FOR VARIANT EFFECT PREDICTION

Ronald Hause, Vanessa Gray, Jens Luebeck, Jay Shendure and Douglas Fowler*

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

Despite the current ease of ascertaining genetic information with next-generation sequencing, the task of interpreting the functional effects of coding genetic variation remains challenging. Many computational tools exist to predict the effects of amino acid substitutions on protein function; however, most of these methods rely on evolutionary, biochemical, and structural information without leveraging experimental results. Where experimental data is used, it tends to be sparse and outdated. Deep mutational scanning measures the functional effects of hundreds of thousands of protein variants simultaneously. Because deep mutational scans survey the sequence-function landscape of proteins with much larger numbers of mutations than what has been available to date, these datasets may improve the performance of models for predicting mutational consequences. We here use 19,493 mutations from nine deep mutational scans to train Envision, an ensemble classifier that uses evolutionary, physicochemical, and structural features to predict protein functionality across species. Features of solvent accessibility, conservation, or structural rigidity were most important for predicting mutational effects. Using cross-validation and external validation, we demonstrate that Envision achieves high accuracy at predicting damaging from neutral effects and outperforms other available algorithms at predicting molecular functional effects. Envision's predictions of molecular effect also improve our ability to distinguish common, non-synonymous variants from the 1000 Genomes Project from rare, pathogenic non-synonymous variants in Clinvar. These results highlight the power of incorporating available experimental data from DMS into variant effect prediction. As additional DMSes are published and incorporated, we anticipate the global applicability and accuracy of Envision to increase.

SEQUENTIAL DATA SELECTION FOR PREDICTING THE PATHOGENIC EFFECTS OF SEQUENCE VARIATION

Mark Rogers*, Hashem Shihab, Tom Gaunt and Colin Campbell.

**University of Bristol, Bristol, UK.*
email: Mark.Rogers@bristol.ac.uk

Recent improvements in sequencing technologies provide unprecedented opportunities to investigate the role of genetic variation in human disease. Accordingly, we present a machine learning approach to predicting whether single nucleotide variants (SNVs) are functional or neutral in human disease. Many data sources from the Encyclopaedia of DNA Elements (ENCODE) may be relevant to this problem, so in previous work we applied integrative multiple kernel learning (MKL) that weights each source according to its relevance. Our predictor outperformed state-of-the-art methods for SNVs in non-coding regions, and matched their performance in coding regions. However, when selecting from a wide assortment of data sources, MKL may not be an efficient method for isolating the most informative sources. Thus we introduce a greedy sequential selection method that incorporates data sources in a structured fashion prior to MKL. When we apply the method to our coding-region predictor, results indicate that it now outperforms other methods. In addition we have used this method to devise an cancer-specific SNV predictor that promises to outperform the best known competitors by a significant margin.

Selected Posters

VarI-SIG Meeting – ISMB/ECCB 2015, July 11th Dublin, Ireland

FINDINGS FROM THE CRITICAL ASSESSMENT OF GENOME INTERPRETATION, A COMMUNITY EXPERIMENT TO EVALUATE PHENOTYPE PREDICTION

Steven Brenner*.

*University of California, Berkeley, Berkeley (CA), USA.
email: brenner@compbio.berkeley.ed.edu

The Critical Assessment of Genome Interpretation (CAGI, \'kā-jē\) is a community experiment to objectively assess computational methods for predicting the phenotypic impacts of genomic variation. In the experiment, participants are provided genetic variants and make predictions of resulting phenotype, for ten challenges. These predictions are evaluated against experimental characterizations by independent assessors.

For example, in a challenge to predict Crohn's disease from exomes, several groups performed remarkably well, with one group achieving a ROC AUC of 0.94. The experiment also revealed important population structure to Crohn's disease in Germany. In another challenge, two groups were able to successfully map a significant number of Personal Genome Project complete genomes to corresponding trait profiles.

Other challenges were to predict which variants of BRCA1, BRCA2, and the MRN complex are associated with increased risk of breast cancer; to associate exomes, variants, and disease in lipid diseases; to predict how variants in p53 gene exons affect mRNA splicing; to predict how variants of p16 tumor suppressor protein inhibit cell proliferation; and to identify potential causative SNPs in disease-associated loci.

Overall, CAGI revealed that the phenotype prediction methods embody a rich representation of biological knowledge, making statistically significant predictions. However, the accuracy of prediction on the phenotypic impact of any specific variant was unsatisfactory and of questionable clinical utility. The most effective predictions came from methods honed to the precise challenge. Prediction methods are clearly growing in sophistication, yet there are extensive opportunities for further progress.

Complete information about CAGI may be found at http://genomeinterpretation.org.

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

Tenghui Chen, Wanding Zhou and Ken Chen*.

*MD Anderson Cancer Center, Houston (TX), USA
email: {tchen1,kchen3}@mdanderson.org

One DNA sequence can code for multiple different mRNAs, and therefore many different proteins. Conversely, a variant identified at the protein or transcript level may have non-unique genomic origins. One-to-many, many-to-one and many-to-many relationships among sequence variants at the genomic level and those at transcript and protein levels introduce frequent inconsistencies in current practice. We have designed a novel variant annotator, TransVar, to perform three main functions (Figure 1A): (i) forward annotation; (ii) reverse annotation; and (iii) equivalence annotation.

We analyzed unique single-nucleotide substitutions (SNS), multi-nucleotide substitutions (MNS), insertions (INS), deletions (DEL) and block substitutions (BLS) in COSMIC using TransVar, ANNOVAR1, VEP2, snpEff3, and Oncotator4, and observed comparable accuracy in SNS and MNS among all annotators but dramatically higher accuracy in INS, DEL and BLS from TransVar (Figure 1B). TransVar's novel reverse annotation can be used to ascertain if two protein variants have identical genomic origin, thus reducing inconsistency in annotation data. It can also reveal whether or not a protein variant has non-unique genomic origin and requires caution in genetic and clinical interpretation. We reverse-annotated the protein level variants in COSMIC and found that a sizeable fraction (e.g., 11.9% of single-AA substitutions) were not associated with a single genomic variant (Table 1). Among the 537 variants that were cited as clinically actionable at PersonalizedCancerTherapy.org, 78 (14.5%) could be mapped to multiple genomic locations.

Our investigation highlighted the importance of standardizing variant annotation and revealed frequent inconsistencies in current data and systems. With both forward and reverse annotation enabled in TransVar, we can reduce such inconsistency and improve the precision of translational and clinical genomics. The source code of TransVar is available at http://bioinformatics.mdanderson.org/main/transvar and a web interface is at http://transvar.net.

STUDY OF IN-SILICO PREDICTORS TO ASSESS THE PATHOGENICITY OF VARIANTS IN CLINICAL DIAGNOSIS

Angela Del Pozo[*], Kristina Ibañez, Juan Carlos Silla-Castro and Pablo Lapunzina[*].

[*]*Hospital La Paz, Madrid, Spain*
email: {angela.pozo, pablo.lapunzina}@salud.madrid.org

Next Generation Sequencing (NGS) allows full and simultaneous study of all mutations present in a set of genes. This advantageous feature becomes a difficult issue if it is unable to select those variants that may be causally associated with the phenotype under study.

This task must be done ultimately by functional assays but in initial steps of the analysis the variants should be ranked according to the deleterious effect of them based of evidences as the phylogenetic conservation, allelic frequency and in-silico predictions.

Hence, it is essential being able to interpret in clinical terms the numerical outcome of the prediction programs integrated in the analysis of NGS data.

In this study we have performed a comparative analysis of a selected set of prediction tools aiming to first, assign labels and thresholds to each program based on a unified criteria and then, assess their predictive value using a benchmark set of around 20,000 variants assembled from literature, a filtered version of humpvar from SWISS-PROT and in-house set of mutations.

Finally, we propose guidelines for the use of in-silico prediction tools depending on to the genetic scenario of the analysis, that is, pattern of inheritance, haploinsufficiency of the gene and penetrance.

NPMIX: NONPARAMETRIC METHOD FOR ADMIXTURE ANALYSIS

Alona Kryschenko[*] and Tatiana Tatarinova[*].

[*]*University of Southern California*
email: aryshchenko@chla.usc.edu, tatarino@usc.edu

We present a novel algorithm for analysis of individuals of mixed origin. Our computationally efficient method is based on Nonparametric Maximum Likelihood strategy. We demonstrate performance of NPMIX for isolated communities in Siberia and show the method's utility for pharmacokinetics and clinical trials.

THE EUROPEAN VARIATION ARCHIVE

Francisco J Lopez*, Tom Smith, Dylan Spalding, Gary Saunders, Ignacio Medina, Cristina Y Gonzalez, Ilkka Lappalainen, Jacobo Coll, Jose M Mut, Jag Kandasamy and Justin Paschall.

The EMBL-European Bioinformatics Institute, Hinxton, UK
email: fjlopez@ebi.ac.uk

The European Variation Archive (EVA) hosted at the European Bioinformatics Institute (EBI) is an open access genetic variation resource that integrates all variant types, from SNVs and short INDELS to structural variants. EVA currently hosts 25 studies which represent 12 different species, describing more than 300 million unique variants.

All EVA variants are annotated with sequence ontology terms using Ensembl's Variant Effect Predictor (VEP) and our value-added annotation pipelines provide protein substitution scores (SIFT, PolyPhen), conservation scores (PhastCons, PhyloP), population frequencies and HGVS notation. We shall present how EVA provides a comprehensive view of these data via a web-based interface, which allows queries and filters based any of these annotation fields, and how these data can be downloaded at the study, file or variant level.

The latest addition to EVA is our clinical browser that incorporates data from ClinVar, GWAS and Cosmic. Our clinically relevant variant set is enriched with SO term annotations from the tailored GENCODE-basic geneset and is thus targeted at a non-genomics specialist audience. We shall present EVA Clinical and describe how our efforts here are aligned with ClinVar, as well as the Center for Therapeutic and Target Validation (CTTV), and is thus an integrated view of the largest publically available clinically relevant variant set.

Finally, as an active partner of the Global Alliance for Genomics Health (GA4GH), we have implemented a RESTful interface compliant with the GA4GH API to enable standardized programatic access to all EVA data.

PON-P2 AND OTHER TOOLS FOR FAST AND RELIABLE VARIANT PRIORITIZATION

Abhishek Niroula*, Siddhaling Urolagin and Mauno Vihinen.

Lund University, Lund Sweden.
email: abhishek.niroula@med.lu.se

Reliable interpretation of variants requires robust computational tools. We developed PON-P2 for classification of amino acid substitutions. It is a machine learning-based method trained and tested on benchmark data obtained from VariBench. Extensive feature selection was applied to select informative features including evolutionary information, GO annotations, physical and biochemical properties of amino acids and functional and structural annotations, if available. It classifies the variants into three classes. The highly reliable cases are grouped as pathogenic or neutral and the unreliable cases are grouped as unknown. The accuracy and MCC of PON-P2 are 0.90 and 0.80 in cross-validation and 0.86 and 0.71 in independent test dataset. It is clearly better than for state-of-the-art tools also when tested on additional datasets. It is very fast and is freely available at http://structure.bmc.lu.se/PON-P2/.

We have developed other more specific variant predictors. These include PON-BTK for variants in the kinase domain of Bruton tyrosine kinase, PON-Diso for protein disorder affecting variants, PON-MMR2 for mismatch repair system variants and PPSC for protein stability affecting variants.

MYVARIANT.INFO: COMMUNITY-AGGREGATED VARIANT ANNOTATIONS AS A SERVICE

Chunlei Wu, Adam Mark, Jiwen Xin, Sean Mooney, Ben Ainscough, Ali Torkamani and Andrew I. Su*

*The Scripps Research Institute, La Jolla (CA), USA

email: {cwu,asu}@scripps.edu

The accumulation of genetic variant annotations has been increasing explosively with the recent technological advances. However, the fragmentation across many data silos is often frustrating and inefficient. We created a platform, called MyVariant.info (http://myvariant.info), to aggregate variant-specific annotations from community resources and provide high-performance programmatic access. Annotations from each resource are first converted into JSON-based objects with their id fields as the canonical names following HGVS nomenclature (genomic DNA based). This scheme allows merging of all annotations relevant to a unique variant into a single annotation object. A high-performance and scalable query engine was built to index the merged annotation objects and provides programmatic access to the developers. As of today, MyVariant.info is serving >280M variants in total and we are actively expanding the coverage by engaging community efforts. MyVariant.info decouples two fundamental steps in management of variant annotations: the creation and maintenance of centralized web services (which requires deep software-engineering expertise), and the task of structuring biological annotations (which requires broad community effort). Annotation providers from the community can provide data parsers to convert their raw data into JSON-compatible objects. The only requirement is that a valid HGVS name is used as the id field for each object. These data can then be queryable through the query engine we built. The data provider doesn't have to worry about building their own query infrastructure. And the research community doesn't have to learn another query interface in order to access new annotations.

Company Presentation

VarI-SIG Meeting – ISMB/ECCB 2015, July 11th Dublin, Ireland

PGMD: A COMPREHENSIVE MANUALLY CURATED PHARMACOGENOMIC DATABASE

Alex Kaplun*

*QIAGEN

email: Alex.Kaplun@qiagen.com

The PharmacoGenomic Mutation Database (PGMD) is a comprehensive manually curated pharmacogenomics database. Two major sources of PGMD data are peer reviewed literature and FDA drug labels. PGMD curators capture information on exact genomic location and sequence changes, resulting phenotype, drugs administered, patient population, study design, disease context, statistical significance and other properties of reported pharmacogenomic variants. Variants are annotated into functional categories basing on their influence on pharmacokinetics, pharmacodynamics, efficacy or clinical outcome. The current release of PGMD includes over 117000 unique pharmacogenomic observations, covering all 24 disease super classes and nearly 1400 drugs. Over 2800 genes have associated pharmacogenomic variants, including genes in proximity to intergenic variants. PGMD is optimized for use in annotating next generation sequencing data by providing genomic coordinates for all covered variants, including SNPs, insertions, deletions, haplotypes, diplotypes, VNTRs, copy number variations, and structural variations.

ACKNOWLEDGMENTS

The VarI-SIG meeting organizers would like to acknowledge:

- Søren Brunak, Technical University of Denmark, Lyngby, Denmark.
- Nuria Lopez-Bigas, University Pompeu Fabra, Barcelona, Spain.
- Yves Moreau, KU Leuven, Leuven, Belgium.
- Peter Robinson, Charité University, Berlin, Germany.
- Joris Veltman, Radboud University, Nijmegen, Netherlands.

The organizers also acknowledge **QIAGEN** (https://www.qiagen.com/) for its financial support.

AUTHOR INDEX