Genomic Med. (2008) 2:159–161
DOI 10.1007/s11568-009-9082-9

ABSTRACTS

Bioinformatics and statistical genomics

001: GEN2PHEN: an international effort to harmonise and optimise the databasing of gene-disease relationships

[1]**A. J. Brookes**, [2]D. Atlan, [3]C. Beroud, [4]E. Birney, [5]S. Brahmachari, [6]A. Cambon-Thomsen, [1]R. Dalgleish, [7]J. den Dunnen, [8]A. Devereau, [9]C. Diaz, [4]P. Flicek, [10]H. Gudbjartsson, [11]I. Gut, [12]T. Kanninen, [13]H. Lehvaslaiho, [14]J. Litton, [15]J. Muilu, [16]J. Oliveira, [4]H. Parkinson, [17]G. Patrinos, [18]G. Potamias, [19]E. Wingender, [20]L. Yip

[1]University of Leicester, Leicester, United Kingdom, [2]PhenoSystems S A, Lillois, Belgium, [3]Inserm, Montpellier, France, [4]European Bioinformatics Inst., EMBL, Hinxton, United Kingdom, [5]Council of Scientific and Industrial Research, Delhi, India, [6]Inserm, Toulouse, France, [7]Leiden University Medical Center, Leiden, Netherlands, [8]University of Manchester, Manchester, United Kingdom, [9]Fundació IMIM, Barcelona, Spain, [10]deCode Genetics, Reykjavik, Iceland, [11]Commissariat à l'Energie Atomique, Paris, France, [12]Biocomputing Platforms Ltd, Espoo, Finland, [13]University of Western Cape, Cape Town, South Africa, [14]Karolinska Institute, Stockholm, Sweden, [15]University of Helsinki, Helsinki, Finland, [16]University of Aveiro, Aveiro, Portugal, [17]Erasmus University Medical Center, Rotterdam, Netherlands, [18]Foundation for Research and Technology, Crete, Greece, [19]BioBase GmbH, Wolfenbuettel, Germany, [20]Swiss Institute of Bioinformatics, Geneva, Switzerland

With disease studies and genomics research producing ever more and ever larger datasets that connect genotypes and phenotypes, there is an urgent need for advanced informatics solutions that can handle this extensive and diverse information. Launched in January 2008, the GEN2PHEN project (Genotype-To-Phenotype Databases: A Holistic Solution) aims to help address this need.

The GEN2PHEN consortium (http://www.gen2phen.org/) involves 19 research and company partners; including 17 from Europe, one from India, and one from South Africa. Funding of 12 Million Euro from the European Commission (7th Framework Programme) is bolstered by additional funds provided by the partner institutions. Being a key European program, GEN2PHEN is intimately connected with other major related projects such as ENGAGE, CASIMIR, EUROGENTEST, BBMRI, ELEXIR, as well as the Human Variome Project (HVP).

The main objective of GEN2PHEN is to establish the technological building-blocks needed to evolve today's diverse databases into a seamless genotype-to-phenotype (G2P) biomedical knowledge environment, tied into genome browsers like Ensembl.

The project's specific objectives include:
(1) Analysis of the current G2P informatics (2) Analysis of ethical aspects that need to be addressed (3) Development of key standards (4) Creation of generic database components and integration solutions (5) Creation of search modalities and data presentation solutions (6) Facilitation of data flows into G2P databases (7) Creation of a 'G2P Knowledge Centre' providing information exchange solutions, search/analysis tools, and support for primary data and comment deposition (8) Deployment of GEN2PHEN solutions to the community (9) Addressing questions of system durability and long-term financing.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 200754.

002: Genome annotation of *Anopheles gambiae* mosquito using tandem mass spectrometry-derived data

[4]**Akhilesh Pandey**, [1]Kumaran Kandasamy, [1]Dhanashree Kelkar, [1]Santosh Renuse, [1]Sutopa Banerjee, [1]Beema Shafreen, [1]Shivakumar Keerthikumar, [2]Ajeet Kumar Mohanty, [2]Ashwani Kumar, [3]Aditya Prasad Dash, [1]Pradip Acharya, [1]T. S. Keshava Prasad, [1]Nandini Patankar, [1]Raghothama Chaerkady

[1]Institute of Bioinformatics, Bangalore 560066, India, [2]National Institute of Malaria Research, Field Station, Goa 403 001, India, [3]National Institute of Malaria Research (ICMR), Delhi 110 054, India, [4]The Johns Hopkins University School of Medicine, Baltimore, MD 21205, United States of America

With the advent in genome sequencing technology, many genomes have recently been completely sequenced. However, a deeper understanding of genome organization including prediction of protein-coding genes remains a major challenge. We present proteomics as a robust complementary approach to annotate sequenced genomes. Here, we present genome annotation of *Anopheles gambiae*, a major sub Saharan African vector of malaria using exhaustive proteomics analysis, using mass spectrometry-derived data. We carried out a comprehensive mass spectrometry analysis of the proteome of *A. gambiae* mosquito, including its larval stages. The samples were

homogenized and digested using trypsin and the extracted proteins fractionated using strong cation exchange chromatography. Each fraction was then subjected to liquid chromatography tandem mass spectrometry (LC-MS/MS) using a quadrupole time-of-flight mass spectrometer. The MS/MS data was searched against non-redundant protein database of all species of Anopheles, Aedes and Drosophila genera. This approach allowed us to validate a number of proteins that were labelled as 'hypothetical' in the *A. gambiae* databases. We were also able to identify proteins that were missed in *A. gambiae* protein databases but were either known or predicted in related species. An alternative approach that we took was to search the MS/MS data against a six frame translation of the genome of *A. gambiae*. Any peptides that were identified based on genomic sequence but were absent in protein databases were further investigated. Using this 'genome search' strategy, we identified a number of novel genes for which there was no evidence either from gene prediction programs or from transcriptomic studies. We are in the process of validating our findings using RT-PCR assays. Overall our studies show that proteomics is a good complement to transcriptomic and gene prediction approaches to annotate genomes accurately and should be used routinely.

003: Prediction of protein–protein interactions between a malarial parasite and human

Srinivasan Narayanaswamy, Krishnadev Oruganty

Molecular Biophysics Unit, Indian Institute of Science, India

Lack of large-scale efforts aimed at recognizing interactions between host and pathogens limits our understanding of many diseases. We present a simple and generally applicable bioinformatics approach for the analysis of possible interactions between the proteins of a parasite, *Plasmodium falciparum*, and human host. In the first step, the physically compatible interactions between the parasite and human proteins are recognized using homology detection. This dataset of putative in vitro interactions is combined with large-scale datasets of expression and sub-cellular localization. This integrated approach reduces drastically the number of false positives and hence can be used for generating testable hypotheses. We could recognize known interactions previously suggested in the literature. We also propose new predictions which involve interactions of some of the parasite proteins of yet unknown function. The method described is generally applicable to any host-pathogen pair and can thus be of general value to studies of host-pathogen protein–protein interactions.

004: Expression divergence during human evolution is shaped by change in genomic neighbourhood of genes

M. Madan Babu, Sarah Teichmann, Subhajyoti De

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB20QH, United Kingdom

Mutations that alter the expression level or expression pattern of genes can contribute to the evolution of new species. Such changes may arise due to small-scale mutations that affect a single or a few base pairs, or due to large-scale events such as segmental duplication or chromosomal re-arrangements. Although the contribution of such mutations to the evolution of gene expression pattern has been well studied, the role of intermediate scale mutations that alter the genomic neighbourhood of a gene, i.e., position effect, remains largely unexplored. Such mutations which affect the neighbourhood of a gene may be a result of (1) recombination or duplication event that resulted in the incorporation of the gene into a completely different region or (2) insertion or deletion of genetic material around the gene.

In this work, we investigate if a change in the genomic neighbourhood of orthologous genes between human and chimp has resulted in expression divergence. By comparing genomic neighborhood of orthologous genes from human and chimp with the expression levels of transcripts from several equivalent tissues, we demonstrate that genes with altered neighborhood are likely to undergo significant expression divergence. In addition, a detailed analysis revealed that the genes which are highly expressed in the prefrontal cortex and several other brain tissues show a high incidence of altered gene neighborhood when compared to its chimpanzee ortholog. Although some of these genes were precisely characterized in the context of disease associated with human cognition, an excess of cognition related genes with altered neighborhood suggests that such a mechanism might have played an important role in the evolutionary path leading from the primate ancestor to human. Taken together, our findings suggest that, in addition to other molecular processes, position effect induced gene expression change is a fundamental mechanism that contributes to transcriptome evolution and ultimately to the evolution of new species.

005: Flow of information in the *M. tuberculosis* interactome network: pathways to drug resistance

Nagasuma Chandra, Karthik Raman

Bioinformatics Centre, IISc, Indian Institute of Science, Bangalore, India

The global burden of tuberculosis has taken a new dimension in the recent years due to the emergence of drug resistant varieties of *Mycobacterium tuberculosis* MDR and XDR-TB, posing a major threat to TB eradication. Our ability to counter resistance is limited by a lack of understanding of how resistance emerges in bacteria. It is essential to understand the ways by which resistance can emerge upon exposure to a given drug. The reductionist approach of understanding proteins individually is obviously not sufficient, even at atomistic levels, making systems biology approaches essential to gain holistic insights. To a protein–protein interactome of *M. tuberculosis*, drug-induced expression data from literature were incorporated. The resulting network was analysed using computational approaches, to identify high propensity routes that would be traversed to bring about drug resistance. These routes form pathways from the drug targets to the proteins involved in extrinsic and intrinsic resistance mechanisms. Identification of these pathways forms the basis for a novel rational way to counter drug resistance. Our analysis shows that different targets are prone to resistance to different extents through different mechanisms. We introduce the concept of 'co-targets', which when simultaneously inhibited with the intended target, is likely to help in combating drug resistance. Different target-'co-target' pairs are identified in the study, which are expected to be useful in the design of new antitubercular drugs and to render existing drugs more useful. This approach is inherently generic, likely to significantly impact drug discovery.

006: Prediction of deleterious human membrane transporter polymorphisms

Vishal Acharya, H. A. Nagarajaram

Center for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad, India

Human membrane transporters play direct roles in the absorption, disruption and elimination of nutrients, ions and many drugs. Human membrane transporters have been implicated in genetic disorders

caused by transporter malfunction such as glucose malabsorption and insulin-resistant glucose transport. It is said that transporters play a second important role in pharmacology in that about 30% of the commonly used prescription drugs target transporters. Variations in membrane transporters contribute to inter-individual differences in the responses to various drugs. Variations are mostly due to non-synonymous single nucleotide polymorphisms (nsSNPs) some of which have been associated to certain diseases. However, large number of nsSNPs have not been associated to any of the known diseases and have remained uncharacterized. In addition to these, on-going genetic studies and human population based studies have also been leading to the discovery of new SNPs. It is important to predict the effect of nsSNPs at the molecular level as well as at the physiological level. In this study we have focused on the nsSNPs which occur on the transmembrane part of the human membrane transport proteins and try to develop a tool which can classify uncharacterized nsSNPs into benign and disease nsSNPs. In order to develop such a tool we have investigated and discovered a number of sequence-based properties and features such as burial status, evolutionary characteristics etc. at SNP sites as well as their flanking regions and used those features to develop a SVM based classification method. In this presentation we report the details of our studies as well as the results obtained.

007: A novel transmission disequilibrium test for quantitative traits

Saurabh Ghosh

Indian Statistical Institute, Human Genetics Unit, 203 BT Road, Kolkata 700108, India

The classical Transmission Disequilibrium Test (TDT) for binary traits proposed by Spielman et al. (1993) is a family-based alternative to population-based case-control studies and circumvents the problem of population stratification as it tests for allelic association in the presence of linkage. However, since the clinical end-point traits are often defined by quantitative precursors, it has been argued that it may be a more prudent strategy to analyze the quantitative phenotypes without dichotomizing them into binary traits. The paradigm of linkage disequilibrium in the context of quantitative traits generally considers the intuitive concept of differences in allelic frequencies between individuals having high values of the quantitative trait and those with low values of the trait as evidence of linkage disequilibrium between the marker locus and the QTL. While Analysis of Variance (ANOVA) is a popular approach for association analyses of population-based quantitative trait data, it suffers from the inherent problem of population stratification, and hence, it is of interest to explore for family-based association methodologies using transmission patterns. Although some methods have been developed for testing transmission disequilibrium in the context of quantitative traits, these are not direct extensions of the classical TDT. We propose a simple logistic regression based test that can be analytically shown to be statistically equivalent to the TDT for binary traits, and hence is not susceptible to the presence of population stratification in the data. We perform Monte-Carlo simulations under a wide spectrum of disease models and varying parameter values of linkage disequilibrium to evaluate the power of the proposed procedure. We find that similar to the binary TDT, the power decreases with increase of dominance and decrease of heterozygosity at the QTL. The proposed method can be easily extended to incorporate multivariate phenotypes. We apply our method to analyze externalizing symptoms, an alcoholism related endophenotype from the Collaborative Study on the Genetics Of Alcohism (COGA) project.