HOME
CAREER FAIR
CONFERENCE ACCOMMODATION
GENERAL INFO
PROGRAM
EXHIBITORS & SPONSORS
SUBMIT
COMMITTEES
CONTACT
PROMOTION
PRESS PASS

Posters

Poster numbers will be assigned May 30th.
If you can not find your poster below that probably means you have not yet confirmed you will be attending ISMB/ECCB 2015. To confirm your poster find the poster acceptence email there will be a confirmation link. Click on it and follow the instructions.

If you need further assistance please contact submissions@iscb.org and provide your poster title or submission ID.

Category A - 'Bioinformatics of Disease and Treatment'

A001 - Metagenomic and Metatranscriptomic analyses of the hepatocellular carcinoma-associated microbial communities and the potential role of microbial communities in liver cancer

Mahmoud ElHefnawi, National Research Centre, Egypt

Short Abstract: Human microbiota is the collection of microbes that inhabit different sites of the human body and recently its alterations were related to different human diseases especially cancers. Liver cancer incidence is continually increasing in Egypt with a high mortality rate. This study aimed to identify the abundant microbial communities that inhabit the liver of the hepatocellular carcinoma patient and may be associated with disease incidence or at least disease progression.
Methods
Fresh liver biopsy samples of two hepatocellular carcinoma Egyptian patients were obtained. DNA from one sample and RNA from other sample were extracted followed by Illumina sequencing. Taxonomic and functional analyses were performed using the MG-RAST server.
Results
Proteobacteria was the dominant phylum followed by Firmicutes and Actinobacteria in both DNA and RNA samples but it was noted that the bacterial diversity and presence of useful bacteria in sample 2 of grade 1 disease (RNA sample) were more than it in sample 1 of grade 2 disease (DNA sample). Also, infectious diseases pathways analysis showed the enrichment of infectious diseases pathways of Staphylococcus aureus infection, Vibrio cholera infection, pathogenic Escherichia coli infection, Hepatitis c, Tuberculosis, Epithelial cell signalling in Helicobacter pylori infection, Bacterial invasion of epithelial cells, and salmonella infection.

Conclusions
This study is a preliminary study that shed a light on the question of the relation between the gut microbiota and liver cancer. Further studies to confirm the conclusions of the paper are needed in the future.

TOP

A002 - Estimating zoonotic risk of influenza A viruses from host tropism protein signature

Christine Eng, National University of Singapore, Singapore
Joo Chuan Tong, Institute of High Performance Computing, Singapore
Tin Wee Tan, National University of Singapore, Singapore

Short Abstract: Influenza A viruses constantly generate novel strains which threaten humans with severe illnesses, bringing about outbreaks and pandemics with devastating consequences. While the viruses naturally reside in avian populations, the rapid evolutionary rate occasionally produces zoonotic strains with the capability to escape their primary host, cross host-species barriers and infect humans. We have previously constructed an influenza host tropism prediction system with individual models to predict avian or human tropism of influenza proteins. This led us to the discovery of distinct host tropism protein signatures of zoonotic strains as compared to typical avian and human strains, with zoonotic strains showing a pattern of mixed avian and human proteins. We next utilized the host tropism protein signatures to construct a zoonotic meta-predictor consisting of machine learning algorithms Naïve Bayes, artificial neural networks (ANN), support vector machine (SVM), k-Nearest Neighbour (kNN) and random forest to classify avian, human, and zoonotic strains. Given an influenza virus with a complete proteome, the meta-predictor is capable of estimating the zoonotic probability with minimum 93.79% accuracy and 0.953 weighted area under the curve (AUC). As such, this meta-predictor could potentially address the limitation of current influenza surveillance technologies by providing a platform to rapidly screen influenza viruses, distinguishing harmless avian strains from more deadly zoonotic strains. This would allow for swift action to be taken upon identification of potential zoonotic strains to prevent the deadly strains from spreading to the human population.

TOP

A003 - Prediction Of Protein Interaction Networks Based On Structural Information Of Predicted Proteins In Genomes Of Leishmania.

Crhisllane Rafaele Santos Vasconcelos, Universidade Federal de Pernambuco, Brazil
Antonio Mauro Rezende, Fundação Oswaldo Cruz (Fiocruz) - Pernambuco - Centro de Pesquisas Aggeu Magalhães (CPqAM), Brazil

Short Abstract: In according to the World Health Organization, 1-2 million new cases of leishmaniasis occur each year. Available drugs for treatment have serious drawbacks and no effective vaccine has been developed. Thus, applications of systemic approaches to discovery of new drug/vaccine targets are needed. One of these approaches is the study of protein interaction networks. Therefore, the main goal of this work is to model protein interaction networks for Leishmania braziliensis and Leishmania infantum, two important causing agents of leishmaniasis, from their predicted proteomes based on structural information. In order to reach this aim, protein sequences of both proteomes were obtained from the TritrypDB. With the BLAST tool, these sequences were aligned against the PDB proteins, and templates were selected according to their identity and the alignment coverage. The leishmania protein and PDB template were then used as input for MODELLER. Proteins, that templates were not found by BLAST search, were submitted to MHOLline, ModPipe and Phyre server. All generated models were assessed by PROCHECK. A total of 8357 and 8239 proteins of L. braziliensis and L. infantum, respectively, were obtained from TritrypDB. From these, 2212 and 2225 proteins were modelling. All models showed less than 1% of their torsion angles in not allowed regions in Ramanchandran Plot. Our next steps will be evaluation of the models through G-factor, DOPE and GA341. Later, all qualified models will be submitted to molecular docking using Hex and Zdock tools, and template-based protein complex structure prediction tools for construction of protein interaction networks.

TOP

A004 - Systematic characterization of the disease and tissue distributions for identification of novel drug targets

David Westergaard, The Novo Nordisk Foundation Center for Protein Research, Denmark
Alberto Santos, The Novo Nordisk Foundation Center for Protein Research, Denmark
Kalliopi Tsafou, The Novo Nordisk Foundation Center for Protein Research, Denmark
Christian Stolte, Digital Productivity, Commonwealth Scientific and Industrial Research Organisation, Australia
Sune Frankild, The Novo Nordisk Foundation Center for Protein Research, Denmark
Albert Pallejà, The Novo Nordisk Foundation Center for Protein Research, Denmark
Janos X Binder, Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), Germany
Seán O\'Donoghue, Digital Productivity, Commonwealth Scientific and Industrial Research Organisation, Australia
Søren Brunak, The Novo Nordisk Foundation Center for Protein Research, Denmark
Lars Juhl Jensen, The Novo Nordisk Foundation Center for Protein Research, Denmark

Short Abstract: The identification and validation of drug targets remains a major obstacle in drug development. To date, the majority of drug targets fall into four classes: G protein-coupled receptors, nuclear receptors, ion channels, and kinases (Overington et al., 2006). Illuminating the Druggable Genome (IDG) is an NIH initiative that will aid the discovery of novel targets by integrating heterogeneous methods and data sources. To this end, we have developed two novel resources, DISEASES and TISSUES, which project evidence onto proteins from the STRING database and two controlled vocabularies, namely Disease Ontology and the BRENDA Tissue Ontology. The use of controlled vocabularies ensures a perfect translatability between the two resources.

The DISEASES and TISSUES resources both integrate heterogeneous evidence from manually curated databases, high-throughput experiments, and automatic literature mining. DISEASES integrates disease-gene associations from Genetic Home Reference, UniProt, DistiLD, COSMIC. TISSUES is a database of gene expression in human tissues according to publicly available data from microarrays, RNA sequencing, mass spectrometry and immunohistochemical staining. Both resources also contain evidence from comentioning in Medline abstracts. Using gold standards, we calibrate quality scores across evidence types and estimate a confidence level for each association.

TOP

A005 - Analysis of intron retention and cryptic splice site activation in human splicing diseases affecting U12-dependent spliceosome

Ali Oghabian, RNA Splicing laboratory, Finland
Mikko J. Frilander, Institute of Biotechnology, Finland

Short Abstract: Pre-mRNA splicing is a nuclear process where noncoding intron sequences are removed from the pre-mRNA molecules to form mature mRNAs. Most multicellular organisms contain two different types of introns, termed as U2- and U12-type introns that are removed by separate splicesomes, respectively. U12-type introns constitutes less than 0.5%, ~700 introns in humans. Previously, using RNAseq analysis we showed that these introns are globally spliced less efficiently than the U2-type introns and that mRNAs containing unspliced U12-type introns are targeted by the nuclear quality control mechanisms.

Here, we analyze splicing defects in human diseases which are caused by mutations in the components of the U12-dependent spliceosome. We analyze intron retention (splicing efficiency) and cryptic splicing events (i.e. activation of incorrect U2-type splice sites near the U12-type intron) in all the known spliceosomal diseases affecting U12-dependent spliceosome, i.e. Isolated Growth Hormone Deficiency (IGHD1), Taybi-Lindren syndrome (MOPD1/TALS), Myelodysplastic Syndrome (MDS) and various experimental conditions (U6atac and Rnpc48 knockdown), using ENCODE as an additional ctrl. These mutations can potentially disturb the normal splicing of all the U12-type introns in human genome. We find that the majority of upregulated cryptic splicing events are shared between the diseases and experimental conditions, but we also observed disease-specific events. We hypothesize that Cryptic splicing and splicing efficiency could provide further understanding of the molecular basis of the disease and provide novel biomarkers for the identification of human diseases with a deficiency in the splicing of U12-type introns.

TOP

A006 - Varicella-zoster virus-derived HLA peptide affinity is a determining factor to develop postherpetic neuralgia

Pieter Meysman, University of Antwerp, Belgium
Benson Ogunjimi, University of Antwerp, Belgium
Stefan Naulaerts, University of Antwerp, Belgium
Philippe Beutels, University of Antwerp, Belgium
Viggo Van Tendeloo, University of Antwerp, Belgium
Kris Laukens, University of Antwerp, Belgium

Short Abstract: The varicella-zoster virus (VZV) can cause two distinct diseases: chickenpox (varicella) and shingles (herpes zoster). Varicella is a common disease in young children, while herpes zoster is more frequent in older individuals. A common complication of herpes zoster is postherpetic neuralgia (PHN), a persistent and debilitating pain that can remain months up to years after the resolution of the rash. The human leukocyte antigen (HLA) genes are responsible for presenting viral peptides to the adaptive immune system and are highly variable between individuals. In this study, we show that the affinity for varicella-zoster peptides of HLA variants associated with higher PHN risk have a lower relative affinity for varicella-zoster peptides than variants with a lower risk. This was accomplished by collecting previous HLA-genotyping results in a meta-analysis to identify those HLA alleles that increase or decrease the risk of VZV and PHN. Prediction of the VZV peptide affinity for these HLA variants using two existing and one in-house-developed state-of-the-art HLA ligand prediction methods revealed that the relative affinity for VZV peptides of HLA variants with increased PHN risk are significantly lower than those with lower PHN risk. These results provide new insight into the development of postherpetic neuralgia and strongly support the hypothesis that one of its possible underlying causes is a suboptimal anti-VZV immune response due to weak HLA-binding peptide affinity. These findings can therefore help steer vaccine development against VZV and allow estimation of the chance that an individual will develop PHN simply by determining their HLA type.

TOP

A007 - Next generation sequencing of human tumor xenografts is significantly improved by prior depletion of mouse cells

Stefan Tomiuk, Miltenyi Biotec GmbH, Germany
David Agorku, Miltenyi Biotec GmbH, Germany
Kerstin Klingner, Oncotest GmbH, Germany
Stefan Wild, Miltenyi Biotec GmbH, Germany
Silvia Rüberg, Miltenyi Biotec GmbH, Germany
Lisa Zatrieb, Miltenyi Biotec GmbH, Germany
Andreas Bosio, Miltenyi Biotec GmbH, Germany
Julia Schueler, Oncotest GmbH, Germany
Olaf Hardt, Miltenyi Biotec GmbH, Germany

Short Abstract: Human tumor xenografts represent the gold standard method for research areas such as drug discovery, cancer stem cell biology, and metastasis prediction.
During the growth phase in vivo, xenografted tissue is vascularized and infiltrated by cells of mouse origin. Due to this, a strong impact of mouse-derived reads on downstream NGS analyses can be expected.
To overcome these limitations, we have developed a fast and easy method (MCD) allowing for the comprehensive depletion of mouse cells by using automated tissue dissociation and magnetic cell sorting (MACS). We have performed whole exome sequencing of bulk human tumor xenografts from lung, bladder, and kidney cancer, and compared the results to samples depleted of mouse cells. A significant increase in read counts (33%) was observed after MCD, indicating improved sample quality.
We mapped the reads of all samples against human and mouse genomes and determined their putative origin. An average of 12% of reads derived from non-depleted samples was assigned to mouse cells. This amount could be reduced to 0.3% by MCD.
A strong impact of MCD was observed on SNP calling: 63+/-10% of all SNPs predicted for the non-depleted samples could no longer be detected after

MCD, 18+/-1% were specific for the depleted xenograft samples, probably due to higher coverage.
Taken together, MCD significantly improves the analysis of human tumor xenografts by NGS. As this effect was observed although a human sequence specific selection has been carried out during exome enrichment, the influence on whole genome and transcriptome sequencing is expected to be even more prominent.

TOP

A008 - Integrating -omic subtypes to personalize cancer treatment

Natalie Fox, Ontario Institute for Cancer Research, Canada
Emilie Lalonde, Ontario Institute for Cancer Research, Canada
Julie Livingstone, Ontario Institute for Cancer Research, Canada
Vincent Huang, Ontario Institute for Cancer Research, Canada
Takafumi Yamaguchi, Ontario Institute for Cancer Research, Canada
Lawrence Heisler, Ontario Institute for Cancer Research, Canada
Richard de Borja, Ontario Institute for Cancer Research, Canada
Andre Masella, Ontario Institute for Cancer Research, Canada
Michael Xie, Ontario Institute for Cancer Research, Canada
Haiying Kong, Ontario Institute for Cancer Research, Canada
Nicholas Harding, Ontario Institute for Cancer Research, Canada
Michael Fraser, Princess Margaret Cancer Centre - University Health Network, Canada
Theodorus van der Kwast, University of Toronto, Canada
John McPherson, Ontario Institute for Cancer Research, Canada
Robert Bristow, Princess Margaret Cancer Centre - University Health Network, Canada
Paul Boutros, Ontario Institute for Cancer Research, Canada

Short Abstract: Current clinical risk groups for localized prostate cancer (PCa) are imprecise with 20-40% of intermediate-risk patients relapsing after treatment with surgery or radiotherapy, while another third are being over-treated. This demonstrates an urgent need for improved clinical classification of patients. I hypothesize that DNA and RNA based profiles synergize to prognose intermediate-risk PCa, improving upon existing tools. To assess the statistical interactions between DNA and RNA profiles, I first evaluated patient subtypes based on different types of molecular aberrations to determine what type of -omics profiles are most useful for clinical classification. I applied unsupervised machine learning to the Canadian Prostate Cancer Genome Network (CPC-GENE) dataset, which includes 250 surgical samples with both whole genome sequencing and RNA-sequencing, to determine whether robust subtypes exist in PCa, and to what extent these co-classify patients. I created four subtypes in both copy number alteration-space (CNAs) and genomic rearrangement-space (GRs). The CNA results largely recapitulate prior literature, with the exception of a novel subtype characterized by subtelomeric gene gains (i.e. within 5 Mbp of the chromosome ends). GR subtypes are defined by breakpoints in specific genes including a subtype with the TMPRSS2-ERG fusion. Unexpectedly, CNA and GR subtypes, which are both caused by double-stranded DNA breaks, do not co-classify patients. In fact, the CNA subtypes carry prognostic information, while the GR subtypes do not. Ongoing analyses are incorporating data from simple somatic mutations, methylation and RNA aberrations.

TOP

A009 - Tumor Stratification with Four Somatic Mutation Profiles

Lee Sael, State University of New York Korea, Korea, Rep
Sungchul Kim , POSTECH, Korea, Rep
Hwanjo Yu, POSTECH, Korea, Rep

Short Abstract: Tumor stratification based on genomic contents provides understanding of the tumor heterogeneity and thus lead to better targeted treatments. In order to stratify tumor, a tumor profile that accounts for the somatic mutation content is needed. However, defining an efficient somatic mutation profile is challenging due to heterogeneous and sparse characteristics of the mutation data as well as their high dimensionality. To solve this problem, we first conduct an analysis of the somatic mutation profile with different similarity measures. The results show that the somatic mutation data has to be converted into appropriate representations to use Euclidean distance, and with the binary somatic mutation profile Jaccard distance is the best choice to accurately capture the relationships among patients. To obtain more accurate and efficient somatic mutation profiles, we introduce a compact representation and search strategy based on Gene-Ontology and orthogonal non-negative matrix factorization. Statistical significance between the identified cancer subtypes and their clinical features are computed for validation. According to the evaluation results of the newly proposed profiles, our method can identify and characterize clinically meaningful tumor subtypes comparable to the recently introduced Network Based Stratification method while enabling a realtime computation.

TOP

A010 - Structural analysis of B-cell epitopes to guide vaccine design

Saba Ferdous, University College London, United Kingdom

Short Abstract: Peptide vaccines have many potential advantages including low cost, lack of need for cold-chain storage and safety. B-cell epitopes (BCEs) are regions of antigens that are bound by antibodies. Approximately 90% of BCEs are discontinuous in nature making it difficult to mimic them for creating peptide vaccines. We have analyzed the composition of B-cell epitopes in terms of extended 'regions' (R, consisting of at least 3 antibody-contacting residues separated by <= 3 residues) and small fragments (F, contacting residues that do not satisfy these requirements). A purely linear epitope would be R1F0. Second, we have developed an automated classification of region shape: linear, curved or folded. It is expected that isolated peptides of linear and folded epitopes will better maintain their conformation than curved peptides. Linkers, cyclized peptides and stapling techniques would allow small numbers of regions to be coupled, or help stabilize folded peptides.
We have investigated the structures of 463 unique B-cell epitopes from which 962 regions and 470 fragments have been identified. Most frequently, epitopes have compositions R1F0, R2F1, R3F1 or R2F2. The 962 regions are classified as 363 linear, 477 curved and 122 folded. These analyses and classifications are now guiding further analyses of conformational stability on native and modified peptides by using molecular dynamics simulations.

TOP

A011 - Non-invasive multivariate prediction model for myocardial infarction

Ursula Neumann, Department of Bioinformatics, Straubing Center of Science, Germany
Ali Canbay, Department of Gastroenterology, University Hospital, University Duisburg-Essen, Germany
Jan-Peter Sowa, Department of Gastroenterology, University Hospital, University Duisburg-Essen, Germany
Theodor Baars, Clinic of Cardiology, University Hospital, University Duisburg-Essen, Germany
Dominik Heider, Department of Bioinformatics, Straubing Center of Science, Germany

Short Abstract: Cardiovascular diseases (CVDs) were estimated by the World Health Organization (WHO) to be responsible for about 17.5 million of worldwide deaths (from which 7.5 million died due to ischaemic heart disease), and thus being the leading cause of death in high- or middle-income countries.
One of the most known types of ischaemic heart disease is the myocardial infarction, which is a common result of a stenosis. The stenosis diameter is typically determined by percutaneous transluminal coronary angioplasty (PTCA), a technique where an X-ray contrast agent is injected via a catheter into the coronary vessels.
In our study, we aimed at predicting the risk for myocardial infarction with minimally invasive measures. Therefore, we developed a non-invasive computational model that uses serum parameters to predict the percentage of stenosis diameter.
We performed a single-center study with a cohort of 437 patients with a myocardial infarction. The patients were separated into two groups, namely those with a high risk and those with a low risk for further myocardial infarctions.
Using the random forest importance analysis, we identified a subset of serum parameters that are able to predict the percentage of stenosis diameter with an area under the curve (AUC) of 0.97 in a subsequent classification model.

TOP

A012 - Multi-label classification for HIV-1 drug resistance prediction

Dominik Heider, Straubing Center of Science, Germany
Mona Riemenschneider, Straubing Center of Science, Germany
Robin Senge, University of Paderborn, Germany
Ursula Neumann, Straubing Center of Science, Germany
Eyke Hüllermeier, University of Paderborn, Germany

Short Abstract: Antiretroviral therapy is essential for human immunodeficiency virus (HIV) infected patients. In general, combination treatments with agents of several drug classes are administered to inhibit viral replication and therewith to slow progression of disease and prolong a patient's life. However, drug resistance is a major problem in therapy as the high mutation rate of HIV can lead to a fast adaptation under drug pressure, resulting in resistant strains and eventually in therapy failure. Additionally, another major concern is the emergence of cross-resistance among drug classes, which might lead to resistance against drugs that have not yet been applied. Therefore, the study of cross-resistance is highly important in HIV research. Nowadays, computational models have been employed for rapid detection of resistance. However, cross-resistance information has not been widely taken into account yet. In our study, we applied multi-label classification to improve drug resistance predictions for the major drug classes in antiretroviral therapy for HIV-1, namely protease inhibitors (PI), nucleoside and non-nucleoside reverse transcriptase inhibitor (NRTIs and NNRTIs, respectively) by using cross-resistance information. By means of multi-label learning, we were able to significantly improve overall prediction accuracy of drug resistance for all drug classes compared to hitherto applied binary classification models.

TOP

A013 - QuantumClone: clonal reconstrucion from HTS data

Paul Deveau, , France
Valentina Boeva, Institut Curie, France
Gudrun Schleiermacher, Institut Curie, France

Short Abstract: Cancer is driven by somatic mutations and copy-number alterations. Many efforts have focused on the identification of driver mutations; nonetheless passenger mutations, although they are not directly linked to the disease, can provide useful evidence on the phylogeny of the tumor and so help uncover reasons for the proliferative activity of a tumor.
We define a clone as a cell population that harbors a unique pattern of mutations and structural variations (SV). A hierarchical tree represents the ancestry of subclones and reflects the order of appearance of new sets of mutations in each subclone. In other words, a set of mutations present in an ancestral clone will also be present in its progeny; each set containing at least one new mutation or SV giving a selective advantage to the subclone compared to its ancestry.
The reconstruction of clonal structure is usually based on the observed variant allele frequency (VAF) for each mutation in one or multiple samples; however, this reconstruction can be ambiguous for mutations present in regions with copy number changes or loss of heterozygosity (LOH). Such changes are common in neuroblastoma, the pediatric cancer we study in order to understand the mechanisms triggering relapse. To address the specificity of such cancers, we developed an algorithm to uncover the cellular prevalence of subclones relying on observed VAF, the genotype at the mutation locus and the level of contamination by normal cells. We solve the non-uniqueness of the cellular prevalence of mutations located in gained, amplified or LOH sites.

TOP

A014 - LigQ: An open-access protein & ligand structural similarity based tool for the enrichment of Virtual Screening compound sets

Leandro Radusky, UBA, Argentina
Adrián G. Turjanski, UBA, Argentina
Xavier Barril, Universidad de Barcelona, Spain
Javier Luque, Universidad de Barcelona, Spain
Marcelo A Martí, UBA, Argentina

Short Abstract: The discovery of small molecules with biological activity, is of great interest for the development of therapeutic agents. One problem is how to narrow the search of possibly active compounds from large ligand databases.
Starting from a selected protein target, our aim is to select a small set of potential drug-like ligands adequate for performing a Virtual Screening(VS) procedure. In this work we developed a bioinformatic pipeline which involves all steps from protein structure determination to ligand docking (of the selected compounds) against the protein target. Protein structures are derived either directly from the Protein Data Bank or corresponds to homology models of the target (automatically generated by the tool). Initial ligand set is determined from those shown to bind or inhibit the selected target family of proteins. The set is enriched by searching similar compounds (shape and properties) in large compound databases (ZINC, PubChem, ChEMBL). Finally, three dimensional valid structures of the enriched compound set are determined (considering protonation, isomers, tautomers, etc.) and passed to the VS protocol.
Proof of concept of the proposed method was performed using targets of Database of Useful Decoys. In example, for Adenosylhomocysteinase protein, for which 33 inhibitors are known, our results show that protocol yields significant enrichment, retaining 30 of the active compounds reducing the set for VS from ~300k compounds to only 11165, which represents a ~3% of the database to be docked.

TOP

A015 - Role of the DPAGT1/β-catenin/YAP signaling network in oral squamous cell carcinoma

Vinay Kartha, Boston University, United States
Liye Zhang, Boston University, United States
Samantha Hiemer, Boston University, United States
Maria Kukuruzinska, Boston University School of Medicine, United States
Xaralabos Varelas, Boston University, United States
Stefano Monti, Boston University School of Medicine, United States

Short Abstract: Progression of oral squamous cell carcinoma (OSCC) to metastasis involves complex changes in epithelial cell growth, survival and migration. While the roles of protein N-glycosylation, Wnt/β-catenin and Hippo pathways in cancer have been independently highlighted, the interplay between these pathways in promoting tumor metastasis is less understood. Prior studies have identified this co-dependent homeostatic pathway network to be deregulated in OSCC, playing a vital role in its tumorigenesis. However, identifying exact mediators of these changes still remains a challenging task and is crucial to the discovery of novel and lasting OSCC therapeutics. Here, we apply a multi-omic profiling approach to identify potential regulators of OSCC pathogenic pathway activity using a combination of OSCC cell line gene expression profiles and massive public genomic data. Gene expression signatures pertaining to genetic knockdowns of DPAGT1 - a gene crucial to protein N-glycosylation, and TAZ and YAP - two transcriptional activators involved in the Hippo pathway were derived using SCC2 cells. Primary human OSCC high-throughput gene expression data obtained from The Cancer Genome Atlas (TCGA) was then projected onto these signatures and analyzed for their association with clinical features including tumor grade and stage. By scoring samples based on their level of pathway deregulation, and additionally leveraging TCGA Copy Number Alteration (CNA), DNA methylation and somatic mutation data, we are able to identify potential genetic and epigenetic regulators of human OSCC development in the context of the DPAGT1/β-catenin/YAP signaling network, paving the way to discovering targets of OSCC therapy.

TOP

A016 - Rapid fetal aneuploidy detection using hardware accelerated alignment

Tim Burcham, Sequenom, United States
Cooper Roddey, Edico Genome, United States
Michael Ruehle, Edico Genome, United States
Robert McMillen, Edico Genome, United States

Michael Sykes, Sequenom, Inc, United States
Penn Whitley, Sequenom, Inc, United States

Short Abstract: Non-invasive prenatal testing can be performed by massively parallel sequencing of DNA from maternal plasma. This method has been shown effective in the detection of fetal aneuploidies of chromosomes 13, 18, 21 and the sex chromosomes. Accurate classification of these aneuploidies requires, in part, alignment of sequencing reads to the human genome, calculation of chromosome fractions based on these alignments and calculation of z-scores for each chromosome based on these fractions. The success of these steps relies upon the choice of aligner and algorithm used to determine the chromosome fractions.
Here we present reclassification of a dataset of 1269 samples previously analyzed using bowtie 2 as the aligner. In this study alignments are generated by the DRAGEN processor, a hardware-accelerated sequencing analysis system developed by Edico Genome. We report systematic differences between the two aligners but equivalent performance in terms of chromosome fraction variability and thus relative chromosome quantification.
Both the bowtie 2 and DRAGEN based analyses successfully identified all known T13, T18 and T21 cases in the dataset. The accuracy was > 99.8% in each classification. At the same time the DRAGEN system provides speed increases of greater than thirty-fold relative to bowtie2 running on a 3.5 GHz CPU, allowing a single computer to replace the efforts of a small cluster.
These results demonstrate that the classification algorithm for fetal aneuploidy is robust and resistant to localized changes in the alignment profile. The DRAGEN system provides equivalent performance to bowtie2 with a significant increase in speed.

TOP

A017 - Perturbed pathway communities for the characterization of putative disease modules

Daniele Pepe, University of Liege, Belgium
Sheeba Santhini Basil, University of Liege, Belgium
Fernando Palluzzi, Politecnico of Milano, Italy

Short Abstract: Background
Functional interdependencies among genes, transcription factors, and other molecular effectors, define networks characterized by small-word properties and a strong community structure. Gene regulation is influenced by the activity of interacting partners and their possible perturbation in diseases. Therefore, detecting disease-associated sub-networks is essential to understand complex diseases. We propose a novel algorithm to find perturbed functional communities, exploiting gene-expression and the information enclosed in biological pathways, to unravel disease modules.

Results
The procedure was tested using an expression microarray experiment on Discoid Lupus Erythematosus (DLE), including 13 controls and 7 cases. An initial network of 389 nodes and 1014 edges was generated by merging 25 significantly perturbed pathways. Four differentially regulated communities were detected by applying the walktrap algorithm and testing them for differentially regulation using Structural Equation Modeling (SEM). The majority of genes in perturbed communities were associated with literature to DLE, including TNF, IL1B, and RAC-1. DO enrichment analysis confirmed the goodness of the method as Lupus erythematous was significantly enriched.

Conclusions
Our method offers a valuable way to study deregulated genes and molecular pathways in human diseases, considering their disease-community membership as part of the perturbation process. This enables an integrated context-based analysis including regulatory, functional, and topological information. Hence, we propose an important contribution to focus research towards communities driving the pathogenesis mechanisms. Our method can be applied on every kind of expression data, including microarrays and RNA-seq.

TOP

A018 - Computational drug target prediction and validation in PI3K/AKT pathway

Tunca Dogan, EMBL_EBI, United Kingdom
Tulin Ersahin, METU, Turkey
Ahmet Rifaioglu, METU, Turkey
Diego Poggioli, EMBL-EBI, United Kingdom
Andrew Nightingale, EMBL-EBI, United Kingdom
Maria J. Martin, EMBL-EBI, United Kingdom
Rengul Cetin-Atalay, METU, Turkey

Short Abstract: One of the emerging topics in the field of drug discovery, is the employment of computational methods in order to reduce the amount of labour and high costs associated with the process. The first step of computational drug discovery is the prediction of novel bioactive drug-target protein pairs using data mining and machine learning techniques, and statistical methods.

Here we propose a novel computational method to provide novel drug-target pair predictions (using the verified bioactive drug-target association information in ChEMBL database) with the combination of two steps: (1) Mapping of small molecule drugs to the structural domains in the target proteins where their region of interaction resides. This way, the other proteins containing the mapped domains becomes novel candidate targets for the corresponding small molecules. (2) Classification of the small molecules considering their molecular fingerprint similarities in order to predict new compounds for the mapped domains. In the end, predictions from two steps are combined with each other to extend the number of bioactive pair predictions.

We selected PI3K/AKT signalling pathway for the experimental validation of our computational approach. Amplification, mutation, translocation and epigenetic modification of PI3K/AKT pathway genes are frequently reported in cancer samples, as a result, its utilization is of utmost importance for the development of novel therapeutic strategies. Initially known PI3K and AKT isoform specific inhibitors will be used for training and testing the method. Next, we plan to obtain alternative inhibitor combination predictions to block the pathway effectively.

TOP

A019 - Transcriptomics of rare cell populations in the aging neural stem cell lineage

Katja Hebestreit, Stanford University School of Medicine, United States
Dena Leeman, Stanford University School of Medicine, United States
Anne Brunet, Stanford University School of Medicine, United States

Short Abstract: Neural stem cell niches in the adult brain are the locations where neural stem cells produce new neurons necessary for the maintenance and plasticity of brain function. With age, neural stem cell niches deteriorate, with a decline in neural stem cell proliferation and production of new neurons. To examine the transcriptional landscape in neural stem cells during aging we obtained RNA-seq data from freshly isolated cells along the neural stem cell lineage from young and old mice. Because of very low cell numbers per replicate and because differences with age were expected to be subtle, we captured unwanted variance in the data using surrogate variable analysis. We used limma to detect differentially expressed genes between cell types and between old and young samples for each cell type. We found strong gene expression differences between the cell types, especially between quiescent and activated cell types. Intriguingly, we found that quiescent neural stem cells show transcriptional changes with age, whereas activated neural stem cells do not seem to have an aging signature. Using pathway enrichment analysis we found that quiescent and activated neural stem cells use different primary pathways to carry out different modes of proteostasis with quiescent neural stem cells favoring autophagy and activated neural stem cells using the proteasome pathway. As defective proteostasis is a hallmark of aging, it represents an interesting candidate of further investigation to understand why activated neural stem cells are protected from transcriptional aging.

TOP

A020 - R2: Accessible online genomics analysis and visualization platform for biomedical researchers

Jan Koster, AMC, Netherlands
Richard Volckmann, AMC, Netherlands

Danny Zwijnenburg, AMC, Netherlands
Piet Molenaar, AMC, Netherlands
Rogier Versteeg, AMC, Netherlands

Short Abstract: In this era of explosive genomics data generation, there is a growing need for accessible software solutions that can help unlock biological/clinical characteristics from such data. With the biomedical researcher in mind, we developed a comprehensive open access academic web-based system called R2 (http://r2.amc.nl).
The R2 platform consists of a database storing both publicly accessible as well as shielded datasets with unified gene annotation, supplemented with a large suite of tools and visualizations that can be used on these data and their associated annotation. As such the user experiences the same look & feel throughout the mining process.
In the public section, R2 hosts over 60,000 samples. Besides gene expression, the platform is also being employed in the integration, analysis and visualization of aCGH, SNP, ChIP, methylation, miRNA, and whole genome sequencing data.
R2 contains a set of interactive inter-connected analyses, allowing users to quickly hop from one view to another. Analyses include, correlation, differential expression, gene sets, gene ontology, transcription factor binding sites, PCA, k-means, Kaplan Meier scans, signature creation etc.
Visualizations include, various gene oriented plots, heatmaps, circos, genome browser, Venn, etc.
Many parts of the R2 platform are publicly accessible through the portal. The gene expression analysis tools have thus far been used in more than 210 peer-reviewed scientific publications. R2 is also used in many international collaborative efforts involving unpublished datasets. The webservers have been serving over 950.000 pages over the past 12 months (feb 2015).

TOP

A021 - RAMBO: Really Accessible Management of Bioinformatics Outcome in massive parallel sequencing for clinical studies.

Juan Carlos Silla-Castro, Bioinformatic Section. INGEMM, Spain
Angela del Pozo, Bioinformatic Section. INGEMM, Spain
Kristina Ibañez, Bioinformatic Section. INGEMM, Spain

Short Abstract: Clinical diagnosis of genetics diseases using massively parallel DNA resequencing (NGS) has become a common practice. In this scenario tasks as management of data deluge from sequencing platforms, analysis tracking, and results distribution are not trivial processes. There are wide-scope tools like Mercury, Galaxy, and cloud computing solutions although they are focused mainly on analysis pipeline execution or distributed as proprietary software. In this work, we present RAMBO (Really Accessible Management of Bioinformatics Outcome) a full stack web application offering users with an authenticated and encrypted interface for sample upload, analysis, and results visualization of NGS data through web browser for a clinical diagnosis.

RAMBO is developed in Scala language with Play Framework as back-end. Its data model is implemented by a graph database, enabling an ease extension and integration with other information sources available elsewhere as graphs. Front-end uses AngularJS and offers an intuitive interface for common analysis in NGS such as coverage of region of interest (ROI), ranking and curation of detected variants, and report generation of studies.

It requires sample characterization together with phenotype description (HPO). The analysis is then automatically performed over the ROI generated by the clinical suspicion. ROI manual setting is also allowed. RAMBO hides storage and computational resources required NGS studies, avoiding transmission and duplication of data generated during analysis.

Clinicians are not always familiar with management and interpretation of NGS data. RAMBO helps them by providing support in the identification of the genetic causes of human genetics conditions.

TOP

A022 - Toward the characterization of patterns in dravet syndrome combining next generation sequencing, clinical data and machine learning techniques

Kristina Ibáñez, Bioinformatics Section. INGEMM, Spain
Juan Carlos Silla-Castro, Bioinformatics Section. INGEMM, Spain
Rubén Martín-Arenas, INGEMM, Spain
Eva Barroso, INGEMM, Spain
Ángela del Pozo, Bioinformatics Section. INGEMM, Spain

Short Abstract: Dravet syndrome (DS) is a severe epilepsy with onset in infancy which includes severe myoclonic epilepsy of infancy and infancy-borderland. DS is characterized by onset of recurrent febrile/afebrile hemiclonic or generalized seizures, followed by appearance of multiple seizure types, generally resistant to antiepileptic drugs. Mutations in SCN1A encoding the channel Nav1.1 are the most common genetic cause of epilepsy demonstrating haploinsuficciency of SCN1A. But the extreme heterogeneity of mutations in SCN1A is remarkable and the 'common disease/common variant' hypothesis cannot apply to SCN1A and epilepsy. Thus, the study of the mutations in the most relevant epilepsy genes result in abnormal firing patterns in epilepsy might be an interesting avenue to uncover the genetic basis in DS, and can be valuable for guiding treatment and estimating recurrence risks.

We composed a targeted panel of 303 epilepsy-related genes for Next Generation Sequencing (NGS) covering epilepsy phenotypes known so far. We evaluated this panel on a cohort of 65 patients with DS (with/without mutations in SCN1A), including the clinical characteristics (Human Phenotype Onthology, HPO terms).

We focused the study in different perspectives: we analyzed the correlation of each mutation localization on different sodium channel structures with the HPO terms, and pathogenicity predictors, and we analyzed all the variants in search of additional modifiers that could be associated with the phenotype.

We identified disease-causing mutations in all the SCN1A positive patients previously validated by Sanger sequencing. We also propose here supplemental modifiers in epilepsy-related genes that could explain different phenotypes relying on clinical features.

TOP

A023 - An Automated Workflow for Analyzing Human Immune System States Using ImmunExplorer

Susanne Schaller, University of Applied Sciences, Austria
Johannes Weinberger, Red Cross Transfusion Service of Upper Austria, Austria
Raul Jimenez-Heredia, Red Cross Transfusion Service of Upper Austria, Austria
Martin Danzer, Red Cross Transfusion Service of Upper Austria, Austria
Stephan M. Winkler, University of Applied Sciences Upper Austria, Hagenberg Campus, Austria

Short Abstract: Advances in high-throughput sequencing enable competitive analyses of the T and B cell receptors, the most essential molecules of the human adaptive immune system. Using the software ImmunExplorer (IMEX), NGS data of T and B cells of various samples can be analyzed regarding clonality, diversity, sequence functionality, and V(D)J gene arrangements. We here present a new machine learning approach for an automated workflow that has the potential to identify immune repertoire related pattern that are specific for certain kidney diseases and abnormalities in human health.
We have recently extended the functionality of IMEX and implemented new algorithms for calculating a set of features mainly ensuing from clonality and diversity data. Amongst others, features such as area under the clonality and diversity progress curves, mean and standard deviation of the frequencies of the top distinct clonotypes, and the declination of the number of frequencies of the distinct clonotypes can be determined.
Those features that have been derived from a broad range of samples can be used as input for various machine learning approaches to learn models that are targeting a classification for the state of health of new patient data. We have integrated machine learning techniques from the open source framework HeuristicLab into IMEX. Thereby, IMEX offers a user friendly workflow to generate a comprehensive analysis of the state of the human adaptive immune system.

TOP

A024 - Quantifying the contribution of various data types in predicting clinical drug response

Nanne Aben, NKI, Netherlands
Magali Michaut, Netherlands Cancer Institute, Netherlands
Daniel Vis, Netherlands Cancer Institute, Netherlands
Lodewyk Wessels, Netherlands Cancer Institute, Netherlands

Short Abstract: Patient response to anti-cancer therapies is highly heterogeneous. It is believed that a large portion of this heterogeneity can be attributed to variation in tumor characteristics. These characteristics are captured by various data types such as point mutations, copy number alterations, methylation status, tissue of origin and gene expression profiles. However, the relative importance of these data types for predicting drug response is largely unknown. This is mainly due to the high degree of redundancy between the different data types and the inability of prediction algorithms to deal with it.

To address this question, we developed a multi-stage regression algorithm that predicts response to a given drug and takes the redundancy between different input data types into account. We show that a large part of the information present in various data types is contained in the gene expression profiles. Applying our algorithm to a panel of hundreds of cell lines and drugs, we effectively adjust for this redundancy and more precisely assess the contribution of each data type to the response prediction.

Our results identify drugs for which the variation in sensitivity is mostly explained by a single data type. For example, we show that response to MAPK pathway targeting drugs can be explained by mutation data, while predicting response to DNA-synthesis targeting drugs requires gene expression data. The work presented here facilitates the identification of the minimal set of data types required for predicting drug response.

TOP

A025 - Bioinformatic analysis of total internal reflection fluorescence microscopy (TIRFM) data in context of type 2 diabetes and cancer signaling

Daniela Borgmann, University of Applied Sciences Upper Austria, Austria
Peter Lanzerstorfer, University of Applied Sciences Upper Austria, Wels Campus, Austria
Verena Stadlbauer, University of Applied Sciences Upper Austria, Wels Campus, Austria
Ulrike Müller, University of Applied Sciences Upper Austria, Wels Campus, Austria
Stephan Winkler, University of Applied Sciences Upper Austria, Hagenberg Campus, Austria
Julian Weghuber, University of Applied Sciences Upper Austria, Wels Campus, Austria

Short Abstract: The plasma-membrane of living cells is a major organelle for cellular signalling cascades. To warrant the diverse functions of a cell, communication between the cytoplasm and the extracellular space is crucial. Thus, membrane-localized proteins activated by messenger molecules are essential to transmit signals into the cell. Therefore, a defective regulation may lead to various diseases.

Here we use a quantitative approach to characterize protein-protein interactions in a live cell context by using total internal reflection fluorescence microscopy (TIRFM). The acquired data are analysed using Spotty, an analysis framework for TIRFM data. We have already successfully used this assay to identify novel substances that function as insulin mimetics and trigger GLUT4 translocation in a CHO-K1 cell line overexpressing the human insulin receptor. Furthermore, the framework has been applied to a novel method called micropatterning assay that was used to identify and quantify the effects of EGFR and insulin receptor modulating substances in the context of oncogenic and diabetic research.

The Spotty framework comprises different features, such as identification of grid structures (e.g. for micropatterned surfaces), automatic cell detection, descriptive statistics, and visualization tools. As a result, a composition of statistical features is generated that can be used for the characterization of the strength of an interaction.

In summary, the combination of the TIRFM methodology and an appropriate analysis framework is here successfully used in order to get new insights into molecular mechanisms and protein-protein interactions, and to provide a better understanding of important biological processes in the context of diseases.

TOP

A026 - Sample Size Requirements for Molecular Subtyping of Breast Cancer

Arvind Mer, Karolinska Institute, Sweden
Daniel Klevebring, Karolinska Institute, Sweden
Mattias Rantalainen, Karolinska Institute, Sweden

Short Abstract: Sequencing-based technology for primary tumor characterization has substantially increased the potential for individualized treatment of patients. Stratification of breast cancer patients using gene expression-based- subtyping into the intrinsic subtypes (basal-like, HER2-enriched, LuminalA, LuminalB and normal-like) has been shown to provide improved prognostication compared to routine pathology and is likely to eventually supersede current pathology. Establishment of prediction models with sufficiently high sensitivity and specificity is key for the adaption of RNAseq-based subtyping of tumors in the clinic. Due to inter-lab and inter-protocol variability of RNAseq data, fitted prediction models are, however, not directly translatable between labs. Therefore, in-house data sets and models will have to be established.

Here we assess the effects of training sample size on breast cancer subtype prediction accuracy through a combination of sub-sampling and Monte Carlo cross validation of the TCGA breast cancer RNAseq data set for multiple prediction models. We show that the sample size effect on prediction accuracy is specific for each subtype, and that doubling the training size from N=300 to N=600 results in a modest, but clinically relevant increase in accuracy. Our results provide information relevant for study design when translating breast cancer subtyping to the clinic.

TOP

A027 - Predicting the Pathogenic Effects of Sequence Variation

Mark Rogers, University of Bristol, United Kingdom

Short Abstract: Recent improvements in sequencing technologies provide unprecedented opportunities for investigating 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. Predictions in non-coding regions are of particular interest, since an estimated 88% of SNV-trait associations reside in these regions.

We assembled a pathogenic dataset using germ-line mutations from the Human Gene Mutation Database, and a control dataset from the 1,000 Genomes Project. Many data sources from the Encyclopaedia of DNA Elements (ENCODE) may be relevant to this problem, so our method integrates different data types and identifies the most informative sources. Starting with 10 appropriate data sources, we applied integrative multiple kernel learning (MKL) which weights each data source according to its relevance. Different sources were informative for non-coding versus coding regions, so we constructed two distinct predictors. Our non-coding predictor significantly outperformed other known methods, while our coding model equaled their performance. With each prediction we provide a confidence measure. Restricting to the highest-confidence predictions we show that the top 16% highest-confidence examples yield a test accuracy of 98%.

Recently we assembled additional ENCODE data sources and results suggest our coding model now outperforms all other methods. We have also devised cancer-specific predictors that outperform the best known alternative cancer predictors, based on our preliminary experiments. In summmary, our results demonstrate that an MKL approach can effectively integrate information from diverse sources to make powerful prediction models.

TOP

A028 - Seq2Res: Facilitating large-scale, low-cost HIV drug resistance genotyping using high-throughput sequencing technologies.

Simon Travers, South African National Bioinformatics Institute, South Africa
Ram Krishna Shrestha, South African National Bioinformatics Institute, University of the Western Cape, South Africa
Baruch Lubinsky, South African National Bioinformatics Institute, University of the Western Cape, South Africa
Imogen Wright, South African National Bioinformatics Institute, University of the Western Cape, South Africa
Natasha Wood, South African National Bioinformatics Institute, University of the Western Cape, South Africa
Miguel Lacerda, Department of Statistical Sciences, University of Cape Town, South Africa
Irene Ketseoglou, Department of Molecular Medicine and Haematology, University of the Witwatersrand Medical School, South Africa
Maria Papathanasopoulos, Department of Molecular Medicine and Haematology, University of the Witwatersrand Medical School, South Africa

Short Abstract: The development of high-throughput, sensitive and cost-effective HIV drug resistance genotyping approaches is a critical priority for the continued success of treatment programmes in resource-limited countries that suffer a high burden of HIV. Conventional resistance genotyping using Sanger sequencing is too expensive (~US$500 per test) for routine use in such settings. High-throughput sequencing (HTS) approaches, coupled with the pooling of samples from multiple patients, can substantially reduce the cost of sequence data generation. The complexities involved in the interpretation of the resulting data, however, currently preclude the use of HTS approaches for routine resistance genotyping.

We have recently developed Seq2Res, a high-throughput, sensitive, web-based tool that interprets HIV drug resistance data generated using the Roche/454, Illumina and Ion Torrent HTS platforms. Seq2Res utilises hidden Markov models to map sequence reads to a reference sequence in codon space, thereby enabling the identification and correction of sequencing and PCR induced errors, thus ensuring a high level of confidence in reported drug resistance mutations. Seq2Res is fully integrated in that it receives the sequencing data directly from the sequencing platform and returns an easily-interpreted table detailing the resistance profile for each patient.

Using Seq2Res we have evaluated the performance of the Roche/454 GS Junior, Illumina MiSeq and Ion Torrent PGM platforms in the HIV drug resistance genotyping of 40 patients previously exposed to various treatment regimens. We find that, while the coverage depth varies significantly depending on the platform used, all HTS results correlate well with those of the current Sanger-based “gold standard”.

TOP

A029 - A non-invasive bioinformatic method for analysis of fetal aneuploidy in maternal blood by NGS.

Angela del Pozo, Bioinformatics section. INGEMM. CIBERER. Hospital Universitario La Paz, Madrid, Spain
Kristina Ibañez, Bioinformatics section. INGEMM. Hospital Universitario La Paz, Madrid, Spain
Juan Carlos Castro-Silla, Bioinformatics section. INGEMM. Hospital Universitario La Paz, Madrid, Spain
Daniel Prieto, INGEMM. CIBERER. IdiPaz. Hospital Universitario La Paz, Madrid, Spain
Victoria F. Montaño, Structural and Functional Genomics section. INGEMM. Hospital Universitario La Paz, Madrid, Spain
Elena Vallespín, Structural and Functional Genomics section. INGEMM. Hospital Universitario La Paz, Madrid, Spain
Elena Mansilla, INGEMM. CIBERER. IdiPaz. Hospital Universitario La Paz, Madrid, Spain
Maria Angeles Mori, INGEMM. CIBERER. IdiPaz. Hospital Universitario La Paz, Madrid, Spain
Roberto Rodríguez, Department of Gynecology & Obstetrics. Hospital Universitario La Paz, Madrid., Spain
Maria Luisa de Torres, INGEMM. CIBERER. IdiPaz. Hospital Universitario La Paz, Madrid, Spain
Pablo Lapunzina, INGEMM. CIBERER. IdiPaz. Hospital Universitario La Paz, Madrid, Spain
Fe García-Santiago, INGEMM. CIBERER. IdiPaz. Hospital Universitario La Paz, Madrid, Spain

Short Abstract: Fetal aneuploidies constitute over 80% of chromosomal abnormalities at birth. Currently, the majority are detected using invasive screening tests, which involve both risks to the mother and fetus.

Recently, research has focused on the development of non-invasive methods to detect fetal chromosomal aneuploidoies in maternal plasma using next generation sequencing (NGS) technologies. Two common approaches have been employed: Whole Genome Sequencing (WGS) or the selection of various chromosomal loci using a proprietary methodology (DANSRTM), not available to the general scientific community.

We present a novel custom bioinformatic NGS design for non-invasive prenatal testing that employs the dosage detection of an unique set of markers along chromosomes 13, 18, 21 and X, compared to a control, in this case chromosome 1. The design also incorporates chromosome Y markers and numerous SNPs for sexing and fingerprinting, respectively.

Approximately 1,500 very reliable markers have been selected, with lengths ranging from 100 to 300 bp. The sequences are captured using Roche NimbleGen EZ Library and sequenced on a MiSeq platform. A total of 50 samples, including trisomies and normal controls, have been used to validate the assay. The results demonstrate a very good trade-off between sensitivity and specificity.

In conclusion, this NGS design would provide a sensitive and robust method for the detection of fetal aneuploidies in maternal blood and would avoid the necessity to undertake a large number of invasive tests.

TOP

A030 - Bioinformatic Analysis of Long Non-coding RNAs in Neuroblastoma

Kate Killick, University College Dublin, Ireland
Markus Schroder, University College Dublin, Ireland
Sarah-Jane Lennon, University College Dublin, Ireland
Thomas Schwarz, European Molecular Biology Laboratory (EMBL), Germany
Walter Kolch, University College Dublin, Ireland
Desmond Higgins, University College Dublin, Ireland
David Duffy, University College Dublin, Ireland

Short Abstract: Neuroblastoma is an embryonic childhood cancer arising from the neural crest progenitor cells of the sympathetic nervous system. It is the most commonly found extra cranial pediatric tumor accounting for approximately for 15% of all childhood cancer deaths. Amplification of the MYCN gene is found in 25% of neuroblastoma tumors and the degree of amplification is correlated with patient outcome. Non-coding RNAs have no protein coding potential yet have been shown to play a role in a diverse range of cellular functions including cell differentiation and embryonic development. In particular, over the last several years a large body of literature has emerged supporting a role for long non-coding RNAs (lncRNAs) in many types of cancer. Identification of novel lncRNAs has the potential to serve as diagnostic markers and therapeutic targets in this complex disease. Coupled with this, improved methods of examining the transcriptome have enabled advances in identifying and understanding non-coding RNAs. Here bioinformatic analyses were used to identify lncRNAs from RNAseq data taken from a range of MYCN amplified neuroblastoma cell lines. Time course data from a MYCN over-expressed cell line was also examined as well as data from a neuroblastoma cell line treated with a retinoid compound known to induce differentiation of neuroblastoma tumors into mature neurons, rendering them benign. Collectively these results demonstrate the induction of lncRNAs by MYCN in neuroblastoma and identify a subset lncRNAs involved in neuroblastoma cell fate and offer a new perspective for neuroblastoma research.

TOP

A031 - Discovering Relationships between Patient Phenotypes and Gene Functionality in large-scale Transcriptomics Data

Lara Urban, Julius-Maximilians-University of Würzburg, Germany
Christian Remmele, Department of Bioinformatics, Biocenter, Julius-Maximilians-University of Würzburg, Würzburg, Germany, Germany
Thomas Dandekar, Department of Bioinformatics, Biocenter, Julius-Maximilians-University of Würzburg, Würzburg, Germany, Germany
Marcus Dittrich, Department of Bioinformatics, Biocenter, Julius-Maximilians-University of Würzburg, Würzburg, Germany; Institute of Human Genetics, Julius-Maximilians-University of Würzburg, Würzburg, Germany, Germany

Tobias Müller, Department of Bioinformatics, Biocenter, Julius-Maximilians-University of Würzburg, Würzburg, Germany, Germany

Short Abstract: With the current abundance of gene expression data, a challenge of transcriptomics is the extraction of biological relations, contexts and functions. Here, we propose the adaptation of multivariate ecological methods to transcriptomics analysis, supplementing applications like GO enrichment analysis and gene set enrichment analysis. In ecology, RLQ ordination and fourth-corner analysis are used to assess associations between species traits and environmental variables. To adapt these approaches to transcriptomics data, various data transformation procedures were introduced and assessed. The analyses of gene expression data from acute lymphocytic leukemia patients identified associations between patient phenotypes and gene functionality which were mostly supported by literature research. These associations were visualized by spatial proximity between and within gene and patient covariates in RLQ ordination. The fourth-corner analysis provided statistical permutation tests for significance of the associations. Power and specificity of the methods were assessed and validated by computer simulations. In conclusion, the combination of fourth-corner and RLQ analysis highly improves the functional interpretation of transcriptomic high-throughput data.

TOP

A032 - Segmentum: a fast tool for allele-specific copy number analysis of cancer genome

Ebrahim Afyounian, BioMediTech - University of Tampere, Finland
Matti Annala, BioMediTech - University of Tampere, Finland
Matti Nykter, BioMediTech - University of Tampere, Finland

Short Abstract: Genomic Copy number alterations (CNAs) and loss of heterozygosity (LOH) events have been recognized as two important drivers of genomic instability associated with cancer. Acquisition of these genomic instabilities has been shown to be correlated with the expression level of oncogenes and tumor suppressor genes. Thus, accurate detection of these abnormalities is a crucial step in identifying novel oncogenes and tumor suppressor genes. Whole-genome sequencing of tumor tissue has enabled new, cost-effective opportunities for the detection of such aberrations and the characterization of genomic aberrations in tumor samples.

We introduce Segmentum, a fast tool for the identification of CNAs and LOH in tumor samples using whole-genome sequencing data. Segmentum segments the genome by analyzing the read depth and B-allele fraction profiles using a sliding window method. It requires a matched normal sample to correct for biases such as GC-content and mapability and to discriminate somatic from germline events. We evaluate Segmentum on both simulated and real whole-genome sequencing data against competing, state of the art methods to demonstrate its accuracy in calling the aberrations. The tool, written in the Python programming language, is fast and performs the segmentation of a whole genome in less than two minutes.

TOP

A033 - Relative abundances of transcript isoforms are predictive of tumor staging and survival in 12 cancer types

Juan Luis Trincado Alonso, Universitat Pompeu Fabra, Spain
Eduardo Eyras, Universitat Pompeu Fabra, Spain
Amadis Pagés, Universitat Pompeu Fabra, Spain
Endre Sebestyén, Universitat Pompeu Fabra, Spain

Short Abstract: Establishing the stage of a tumor is crucial to select the appropriate therapeutic strategy and to determine patient prognosis. Molecular signatures that accurately predict the clinical outcome of individuals with cancer are essential for appropriate therapy selection. Alterations in RNA processing are emerging as important novel signatures to understand tumor formation and to develop new therapeutic strategies. However, it is not yet known whether specific patterns of transcript isoform expression in tumors can be associated to clinical stage. Using a machine learning approach, we integrate data from RNA sequencing from 12 cancer types from The Cancer Genome Atlas (TCGA) project and build predictive models of tumor staging and clinical outcome. We show that this models can separate with high accuracy early from late stage cancer patients and show significant difference in survival. Applied to patients of unknown stage, we show a significant difference in survival between late-stage and early-stage predicted patients. We further provide evidence that cancer type specific models have better accuracies than generic models across multiple types. In addition, we compare with gene expression classifiers and show that the accuracy obtained through transcript isoforms is comparable to models based on gene expression. We also find significant isoforms changes that separate different prognosis according to the expression of the estrogen receptor gene in breast cancer samples We conclude that transcripts isoform relative abundances could be used as another useful tool in clinical decision-making.

TOP

A034 - Identification of survival associated network modules from personalized networks

Chengyu Liu, University Of Helsinki, Finland
Sampsa Hautaniemi, University Of Helsinki, Finland
Rainer Lehtonen, University Of Helsinki, Finland

Short Abstract: Background
Histologically similar tumors even from the same anatomical position may still show high variability at molecular level hindering analysis of genome-wide data. Leveling the analysis to a network instead of focusing on single genes has been suggested to overcome the heterogeneity issue. Studying altered network modules in an individual patient level is crucial to understand mechanism and heterogeneity of the cancer, and to apply patient-tailored risk prediction and treatment. However, methods that integrate gene expression data and network information focus on discovering altered pathways between normal and cancer groups. Hence, they are not suitable for characterizing the gene regulation networks at the single patient level and for identifying altered network modules from individual patients.
Methods
We present a novel network method, Differentially Expressed Network Module Analysis (DENMA) that integrates expression data and gene regulation network information at a single patient level and subsequently identifies network modules. Our method derives its power by focusing on network modules, that is, a group of genes that are connected through gene regulations.
Results
We demonstrate how DENMA yields insights into both simulation data and real experiment data. On the simulation data, we evaluated the performance of DENMA and its potential of identifying patient-survival related network modules. DENMA utility was shown on 1,040 breast cancer data from The Cancer Genome Atlas (TCGA) and the results show DENMA is able to identify network modules that predict patient survival ($P<0.000154$). DENMA is embodied into a freely available software package.

TOP

A035 - Correcting for Noise and Batch Effects in High-throughput Cancer Measurements Through Partially Blind Domain Adaptation

Adrin Jalali, Max-Plank Institute, Germany
Nico Pfeifer, Max-Planck Institute, Germany

Short Abstract: Technology has recently transformed biology by providing large scale high-throughput methods (e.g., from Sanger sequencing to deep sequencing) and is more and more targeted towards the analysis of single cells. This opportunity also brings many challenges; measurements produce high dimensional data that can be noisy and have batch effects, which can lead to discovering spurious associations. The problem of how to choose a reliable set of features to understand and predict the outcome has not been solved and our work is a contribution in this area.

In this work we extend our recently proposed method that uses multiple models with exclusively different features based on expression and DNA methylation data from TCGA. It learns the relationship between features of the data, and calculates reliability of the features that the models use per test sample, leading to a personalized prediction. This so-called partially blind domain adaptation can decrease the problem of batch effects and noisy data. It also addresses the challenge of having different underlying causes for the same disease out of which often only one is represented in each test sample.

We present how the interpretation can give new insights into the underlying factors of a disease that has not been explored in the context of cancer. We show that the method outperforms other methods or performs very well compared to other methods in several data-sets with different batches.

TOP

A036 - LAPRAS: An Integrative Model Incorporating Heterogeneous Datasets to Discover Genetic Etiology of Autism Spectrum Disorder

Sumaiya Nazeen, Massachusetts Institute of Technology, United States
Rohit Singh, Computation & Biology Group, CSAIL, MIT, United States
Bonnie Berger, CSAIL, Massachusetts Institute of Technology, United States

Short Abstract: Autism spectrum disorder (ASD), prevalent in 1% of the population, refers to a group of complex neurodevelopmental disorders sharing the common feature of dysfunctional reciprocal social interaction. There is compelling evidence that genetic factors are a predominant cause of ASD; however, the genetic heterogeneity underlying ASD makes it challenging to gain conclusive biological insights into the disease. Most of the general-purpose gene prioritization methods and ASD-specific gene network methods suffer from the limitation of depending just on the protein-protein interaction (PPI) network and/or co-expression network, and do not properly utilize other types of ASD-related information available in literature. We believe understanding the complex genetic background of ASD requires a strategy that can integrate multiple forms of data. To this end, we present a computational method termed LAPRAS (LAsso-Penalized logistic Regression based gene ASsociation) that incorporates ASD-specific DNA copy number variations, PPI network topology, phenotypic similarities of diseases, and pathway knowledge from literature. We provide a rank-list of genes in descending order of their probability of association with ASD. The top-ranked genes are overrepresented in neurological pathways, cell adhesion pathways, and signal transduction pathways pertinent to brain, cellular assembly and communication, synaptic development, and neuronal development. The most significant sub-networks discovered in the top-ranked genes are overrepresented in gastro-intestinal disorders, nervous system development, hereditary developmental disorders, and organismal abnormalities suggesting the existence of subclasses of ASD. This integrative method is novel and outperforms other state-of-the-art gene ranking methods.

TOP

A037 - Bringing protein functional positional annotation knowledge into reference genomes

Andrew Nightingale, European Bioinformatics Institute, United Kingdom
Jie Luo, European Bioinformatics Institute, United Kingdom
Maria-Jesus Martin, European Bioinformatics Institute, United Kingdom
UniProt Consortium, EBI & PIR & SIB, United Kingdom

Short Abstract: Over the last decade, life science research has become a data driven scientific field. The challenge for the next decade is to add biological context to these data, transforming the life sciences into a knowledge driven research field. UniProt acts as the global central hub for protein information by providing comprehensive and high quality manually curated annotations and cross-references from, and connecting to, 144 biological data resources.

In collaboration with Ensembl, we have mapped all human protein sequences in UniProtKB to the human reference GRCh38 genome. The mappings are distributed in a BED file format which can be used as track annotations in Ensembl and UCSC genome browsers. As well as mapppings to genomic coordinates, UniProt's BED files provide important information on functional positional annotations such as active and metal binding sites, Post Translational Modifications (PTMs), disulfide bonds and UniProtKB manually curated variants.

The UniProt BED files are a new valuable resource for translational biology research. By bringing UniProtKB's protein positional functional annotations and disease related variants into the genomic assembly context, UniProt provides additional information for genomic analysis that contributes knowledge on the functional role of a protein, its role within biological processes and how modifications to the gene or protein product can lead to a disease state.

This poster will illustrate the richness of information provided by UniProt and demonstrate how aligning these annotations to genomic annotations provide a powerful way for understanding molecular and cellular processes and there contributions to diseases.

TOP

A038 - Constructing Gene Signatures for Connectivity Mapping to Target Colorectal Cancer

Qing Wen, CCRCB, Queen's University Belfast, United Kingdom
Peter Hamilton, CCRCB, Queen's University Belfast, United Kingdom
Shu-Dong Zhang, CCRCB, Queen's University Belfast, United Kingdom

Short Abstract: A fundamental challenge in biomedicine is to establish the relation between the action of drugs and their effect on diseases. Connectivity mapping is used to identify connections between different biological states and has great potentials in accelerating drug discovery and development. Our research aims to find optimized methods to create high-standard gene expression signatures and to propose a standardized procedure and protocol for connectivity mapping.
We construct gene signature with microarray data through differential gene expression analysis to represent disease phenotypes. In particular, we propose a very useful method to select significant genes across multiple datasets. Using an optimised ranking scheme, all genes have a score in each dataset, and each score contains a sign to indicate the direction of differential expression. Genes are selected according to the absolute value of their total score. Using this method, we created a combined signature of 148 genes from 5 independent colorectal cancer datasets. The signature identified 10 significant candidate compounds via SscMap (statistically significant connections' map), four of which are anti-cancer drugs. The compounds Irinotecan and Etoposide are the most interesting and reassuring discoveries in the result as they are chemotherapy drugs currently used to treat colorectal cancer. Other compounds in the results are under investigation as these might be potential drugs for treating colorectal cancer.
The method we proposed to create gene signatures for multiple datasets for connectivity mapping is effective and efficient; the generated quality gene signatures can greatly utilise the power of connectivity mapping to identify potential drugs for the treatment of diseases.

TOP

A039 - Towards personalised medicine in Cystinuria: insights from structural modelling

Mark Wass, University of Kent, United Kingdom

Short Abstract: Cystinuria is a genetic disease that results in the formation of kidney stones. Mutations in two genes, SLC7A9 and SLC3A1, which form an amino acid transporter, are responsible for causing the disease. Failure to transport cystine into the cell, leads to the build of insoluble extracellular cystine, which forms stones. We have used structural modelling of the transporter proteins to propose how the mutations present in a cohort of individuals with the disease affect amino acid transport. We have linked our analysis with the disease severity of each patient to propose the extent of the effect on transporter function by each mutation with the aim that this information can be used to inform future treatment of patients when diagnosed with the disease.

TOP

A040 - Analysis of Volatile Organic Compounds during Sepsis in Rats

Anne-Christin Hauschild, Max Planck Institute for Informatics, Germany
Felix Maurer, Department of Anesthesiology, Intensive Care and Pain Therapy, Saarland University Medical Center, Homburg (Saar); Germany, Germany
Tobias Fink, Department of Anesthesiology, Intensive Care and Pain Therapy, Saarland University Medical Center, Homburg (Saar); Germany, Germany
Jörg Ingo Baumbach, Faculty Applied Chemistry, Reutlingen University, Reutlingen, Germany, Germany
Sascha Kreuer, Department of Anesthesiology, Intensive Care and Pain Therapy, Saarland University Medical Center, Homburg (Saar); Germany, Germany
Sandrah P. Eckel, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA, United States

Jan Baumbach, Computational Biology group, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark, Denmark

Short Abstract: Sepsis is the leading cause of death in critically ill patients in the USA and Europe and the number of incidences. Moreover, we are lacking a rapid and accurate identification method for sepsis and its causative microorganisms, which is essential for a successful treatment.
Studies indicate that breathomics, the metabolomics studies of exhaled air, provides a solid technique to detect sets of potential biomarkers for various diseases. Furthermore, it is well suited for bed side monitoring that is essential for a fast progressing disease such as sepsis.
Two previous studies on systemic inflammatory response and sepsis investigated the breatholom at a single time point only. Our study, for the first time, incorporates and analyzes the course of the metabolite intensities during disease progression in rats.
The breath of 20 male Sprague-Dawley rats were captured. Sepsis was induced in 10 of these study subjects. On this poster we will present the first results of the longitudinal analysis on the resulting time dependent data.

TOP

A041 - Identifying Allele Specific Expression in Parkinson's Disease Brain Tissue

Rachael Ivison, Boston University, United States
David Jenkins, Boston University, United States
Demetrius Dimucci, Boston University, United States

Short Abstract: Allele specific expression (ASE) occurs when mRNA transcripts from two alleles are expressed at different levels. This disequilibrium is due to differential regulation across the two alleles. Variation in expression due to regulatory elements has been shown to be abundant for both cis and trans regulatory elements. Parkinson's disease (PD) is a neurodegenerative disease that manifests late in life and affects between seven and ten million people worldwide. In this study, we sought to identify genes undergoing ASE in pre-frontal cortex tissue of PD patients, but not in control patients. Our approach combined SNP microarray and RNA-Seq data from 33 control samples and 11 PD patients to identify genes containing at least one heterozygous SNP in any sample. We aligned RNA-Seq reads with GSNAP, a SNP-tolerant aligner, to prevent reference allele mapping bias. The software MBASED was used on the aligned reads of heterozygous SNPs to identify genes that showed evidence of ASE. To facilitate exploration of complex expression patterns and identification of SNPs experiencing ASE, we wrote custom visualization software that displays heatmaps of allele expression ratios. Preliminary results found 6722 genes that undergo significant ASE in at least one individual and 126 genes that undergo significant ASE and are significantly overrepresented in patients with PD ($p < .05$).

TOP

A042 - RNA-Seq analysis in Leishmania infantum reveals differentially expressed genes related to drug resistance

Leilane Gonçalves, Fundação Oswaldo Cruz, Brazil
Juvana Andrade, Fundação Oswaldo Cruz, Brazil
Renato Delfino, Fundação Oswaldo Cruz, Brazil
Daniela Resende, Fundação Oswaldo Cruz, Brazil
Pascale Pescher, Intitute Pasteur, France
Gerald Spaeth, Intitute Pasteur, France
Silvane Murta, Fundação Oswaldo Cruz, Brazil
Jeronimo Ruiz, Fundação Oswaldo Cruz, Brazil

Short Abstract: Leishmania parasites cause a broad spectrum of clinical diseases known as leishmaniasis. Annually, approximately 1.3 million new cases are reported and many therapeutic failures due to development of drug resistance have been observed. We performed a comparative transcriptomics analysis of L. (L.) infantum chagasi strains (MHOM/BR/74/PP75) that are susceptible or resistant to potassium antimonyl tartrate (SbIII) using NGS Illumina Sequencing. In order to investigate the differential gene expression associated with drug-induced stress response and SbIII-resistance mechanisms, we compared SbIII-treated and non-treated samples of each strain. In the analysis process, TopHat 2 was used for mapping the reads against the reference genome, Cufflinks was used for transcripts assembly and Cuffdiff and DESeq 2 were used for statistical analyses. The pipeline applied in the analysis process allowed the identification of 262 differentially expressed genes in the SbIII-resistant strain and 186 genes in the SbIII-susceptible strain. Metabolic pathways associated with the differentially expressed genes were addressed using KEGG and functional enrichment analysis (biological process, molecular function and cellular component) was performed using the Blast2GO software. The most frequent pathways identified are phospholipid glycerol metabolism, aminobenzoate degradation and biosynthesis of fatty acids. Details associated with the analysis process and the results will be presented.

Financial support: CNPq, FAPEMIG, UNICEF/UNDP/World Bank/WHO, PROEP/CNPq/FIOCRUZ, Convênio Instituto Pasteur/FIOCRUZ.

TOP

A043 - Novel Bioinformatics Approaches for Vector Integration Sites Profiling and Clonal Repertoire Study in Gene Therapy

Saira Afzal, , Germany
Raffaele Fronza, German Cancer Research Centre, Germany
Stefan Wilkening, German Cancer Research Centre, Germany
Cynthia Bartholomae, German Cancer Research Centre, Germany
Christof von kalle, German Cancer Research Centre, Germany
Manfred Schmidt, German Cancer Research Centre, Germany

Short Abstract: In the last two decades, gene therapy has shown rapid advancements as a promising approach to treat genetic diseases by introducing corrected genes into patient cells. Viruses are used to deliver therapeutic genes to modify hematopoietic stem cells. However, integration of viral vector at undesirable genomic locations can lead to deleterious effects, e.g. insertional mutagenesis. Therefore, long term monitoring of the distribution pattern of vector integration sites (IS) is the most appropriate strategy to address vector safety concerns.
In recent years, next generation sequencing technologies have dramatically increased the possibility to generate substantial amount of vector-genome sequencing data for comprehensive IS analysis. An efficient downstream analysis of this data requires automated and fast computational methods.
We are presenting here a novel bioinformatics approach for time-efficient and reliable analysis of vector-genome junctions and quantitative assessment of clonal repertoire. This framework is designed to analyze the sequencing data generated from traditional linear amplification mediated PCR (LAM-PCR) based methods and also from new targeted DNA single and paired end sequencing technologies.
We have used state-of-art aligners and soft clipped information from alignment files along with multiple sequential steps to process and extract vector-genome integration distribution. It takes approximately 30 minutes for complete processing, starting from raw reads till obtaining annotated IS, of 10 million paired end sequencing data. We have used this method for monitoring clinical and pre-clinical gene therapy data. It is a highly appropriate bioinformatics method for in-depth quantitative analysis of biosafety and transduction efficiency of viral vectors.

TOP

A044 - Application of Latent Class Analysis to Mutation Profiles of Brain Gliomas Reveals Distinct Molecular Sub-Types

Alex Fichtenholtz, Foundation Medicine, United States
Juliann Chmielecki, Foundation Medicine, United States
Michael Goldberg, Foundation Medicine, United States
Garrett Frampton, Foundation Medicine, United States
Debrah Morisini, Foundation Medicine, United States
Siraj Ali, Foundation Medicine, United States
Doron Lipson, Foundation Medicine, United States
Roman Yelenski, Foundation Medicine, United States
Jeff Ross, Foundation Medicine, United States
Vince Miller, Foundation Medicine, United States

Phil Stevens, Foundation Medicine, United States
Eric Neumann, Foundation Medicine, United States

Short Abstract: Cancer is the uncontrolled proliferation of cells, and while it is widely accepted that tumor growth is driven by mutations in tumor suppressors and oncogenes, diagnosis of cancer is done via histology, and classification of the disease is primarily based on tissue of origin. We hypothesize that disease classification based on molecular profiles will provide better explanatory power for clinical outcomes than what is possible using tissue based classification. FoundationOne is an NGS-based genomic profiling test that detects all mutation types in the coding exons of 287 cancer related genes and 47 introns of 19 genes frequently rearranged in tumors. Using FoundationOne, we sequenced 91, 103, and 539 samples from patients with low-grade astrocytoma (LGA), anaplastic astrocytoma (AA), and glioblastoma (GBM), respectively. Mutation load was 3.92 alterations per sample for LGA, 4.41 for AA, and 5.37 for GBM. Latent class analysis of the aggregate sample space was performed using the R package poLCA. Results show stratification of samples into several mutation combination based sub-types, including classes driven by concurrent TP53/IDH1/ATRX mutations and concurrent CDKN2A/CDKN2B/EGFR mutations. Mutually exclusive and co-occurrence relationships between individual genes are clearly visible. All sub-classes were present in every disease, though at differing proportions. Low grade brain tumor samples were enriched in the TP53/IDH1/ATRX class, suggesting a better prognosis for these patients, whereas high grade tumor samples were enriched in the CDKN2A/CDKN2B/EGFR class, suggesting a poorer prognosis for these patients. These results show concrete progress towards the systematic classification of cancer based on molecular profiling.

TOP

A045 - Detection of genome rearrangements in papillary thyroid cancer

Aleksandra Pfeifer, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Poland
Dagmara Rusinek, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Poland
Jadwiga Zebracka-Gala, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Poland
Tomasz Tyszkiewicz, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Poland
Joanna Polanska, Silesian University of Technology, Poland
Barbara Jarzab, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Poland

Short Abstract: Papillary Thyroid Carcinoma (PTC) is the most common type of thyroid cancers. In about 70% of cases the driver mutation of BRAF, HRAS, NRAS or KRAS gene or the rearrangement of RET is present. In rest of the samples, one of the rare alterations occur, there are also many samples in which the driver alteration is unknown. In our study we aimed to detect novel rearrangements which have not been already described in PTC. We selected 6 PTC samples. Two of them were positive controls and harbored RET/PTC1 and RET/PTC3 rearrangements. Four of them did not harbor none of most frequent driver alterations: BRAF, HRAS, NRAS, KRAS, RET/PTC1 nor RET/PTC3. We performed paired end RNA-seq experiment on Illumina HiSeq for those samples. Further, we performed bioinformatic analysis in order to detect genomic rearrangements. The analysis was performed with three programs: ChimeraScan, TopHat-Fusion and SnowShoes. Own filters were also applied to remove false positive results. In two positive control samples we confirmed the presence of RET/PTC1 and RET/PTC3. In one sample we detected ETV6-NTRK3 rearrangement which have been recently described as an oncogenic driver in PTC. In one sample we detected a novel isoform of RET/PTC1. Both rearrangements were further positively validated with PCR and Sanger sequencing. Summarizing, in our study we detected a novel rearrangement in PTC, which would be useful in understanding its biology.
The Project was financed by the National Science Center Poland based on the decision no. DEC-2011/03/N/NZ2/03495. This research was supported in part by PL-Grid Infrastructure.

TOP

A046 - Computational discovery of plasma microRNA profiles as biomarkers of temporal lobe epilepsy

Catherine Mooney, Royal College of Surgeons in Ireland, Ireland
Rana Raoof, Royal College of Surgeons in Ireland, Ireland
Hany ElNaggar, Beaumont Hospital, Ireland
Amaya Sanz Rodriguez, Royal College of Surgeons in Ireland, Ireland
Eva Jimenez-Mateos, Royal College of Surgeons in Ireland, Ireland
Sebastian Bauer, Philipps-University, Germany
Felix Rosenow, Philipps-University, Germany
Norman Delanty, Beaumont Hospital, Ireland
David Henshall, Royal College of Surgeons in Ireland, Ireland

Short Abstract: Epilepsy is a common neurological disorder affecting approximately 1% of the population and is characterised by recurrent unprovoked seizures. The lack of a clinically accepted biomarker for epilepsy diagnosis as well as the incomplete and vague histories often provided by patients contribute to a 30% misdiagnosis rate. MicroRNAs are a class of small non-coding RNA that regulate gene expression at a post-transcriptional level. MicroRNAs are important contributors to brain function and emerging animal and human data suggest microRNAs control multiple pathways in epilepsy. MicroRNAs are also detectable in various body fluids and their stability as well as link to disease mechanism makes them potentially ideal molecular biomarkers of epilepsy. We determined plasma levels of over 800 microRNAs collected from 20 healthy volunteers and 20 epilepsy patients using high-throughput real-time quantitative reverse transcription PCR. Computational analysis included normalisation, clustering, differential expression analysis, target prediction and pathway analysis. A number of significantly differentially expressed microRNAs were identified between control and epilepsy samples including known brain-expressed microRNAs implicated in epilepsy. Furthermore, we applied feature selection with machine learning algorithms, including support vector machines and bidirectional recurrent neural networks, to build a microRNA-based predictor of epilepsy, validated on an independent test set. This analysis showed that these classifiers may be useful in supporting the existence of a set of microRNAs implicated in disease pathogenesis that may be biomarkers of human epilepsy.

TOP

A047 - A Bayesian Tensor Factorization Method to Predict Drug Response in Cancer Cell Lines

Nathan Lazar, Oregon Health and Science University, United States
Mehmet Gonen, Oregon Health and Science University, United States
Kemal Sonmez, Oregon Health and Science University, United States
Shannon McWeeney, Oregon Health and Science University, United States
Lucia Carbone, Oregon Health and Science University, United States
Adam Margolin, Oregon Health and Science University, United States

Short Abstract: Precision oncology aims to improve cancer patient outcomes by identifying the putative networks that drive a given patient's tumor, and attacking these drivers with combinations of targeted therapies. Cell line screening panels give information on how specific tumors may respond to drugs by measuring the growth of tumor cells after exposure to compounds at varying doses in a high-throughput manner. By combining information from these panels with ‘omic profiles of cell lines as well as structural and target information on drugs we can build models to predict response and gain insight into the mechanisms governing response.

Our method decomposes a central three-dimensional tensor encoding the responses of 70 breast cancer cell lines treated with 90 drugs at 10 doses into latent factors representing characteristics of the cell lines and compounds. By reconstructing the central tensor object from these latent factors, we are able to predict full dose-response curves for missing cell line-drug combinations and extrapolate to unseen cell lines and drugs. The use of Bayesian sparsity-inducing priors while training the model allows for the automatic determination of the most informative cell line and drug features. Our method predicts responses at each dose, circumventing issues that arise from using summary measures of drug response. Lastly, by incorporating all of the data into one model, we are able to leverage information from each cell line and drug for the prediction of the others and examine relationships between cell line and drug features.

TOP

A048 - An approach to identify disease-related genes based on biological and web data

Jeongwoo Kim, Yonsei University, Korea, Rep

Sanghyun Park, Yonsei University, Korea, Rep

Short Abstract: Since the genome project in 1990s, a large number of studies associated with genes have been conducted and researchers have confirmed that genes are involved in disease. For this reason, the identification of the relationships between diseases and genes is important in biology. We propose a method called LGscore, which identifies disease-related genes using Google data and literature data. To implement this method, first, we construct a disease-related gene network using text-mining results. We then calculate the weights of edges in the gene network. The weights contain two values: the frequency and the Google search results. The frequency value is extracted from literature data, and the Google search result is obtained using Google. We assign a score to each gene through a network analysis. We assume that genes with a large number of links and numerous Google search results and frequency values are more likely to be involved in disease. For validation, we investigated the top 20 inferred genes for five different diseases using answer sets. The answer sets comprised six databases that contain information on disease–gene relationships. We identified a significant number of disease-related genes as well as candidate genes for Alzheimer's disease, diabetes, colon cancer, lung cancer, and prostate cancer. Our method was up to 40% more accurate than existing methods.

TOP

A049 - Beegle: From literature mining to disease-gene discovery

Sarah ElShal, KU Leuven, Belgium
Jesse Davis, KU Leuven, Belgium
Yves Moreau, KU Leuven, Belgium
Léon-Charles Tranchevent , KU Leuven, Belgium
Alejandro Sifrim , KU Leuven, Belgium
Amin Ardeshirdavani, KU Leuven, Belgium

Short Abstract: Disease-gene identification is a challenging process that has multiple applications within functional genomics and personalized medicine. Typically, this process involves both finding genes known to be associated with the disease (through literature search) and carrying out preliminary experiments or screens (e.g. linkage or association studies) to determine a set of promising candidates for experimental validation. We would like to present Beegle, an online search and discovery engine that automates this process entirely. Its starts by mining the literature to automatically extract a set of genes known to be linked with a given query, and then it integrates multiple sources of genomic information to generate novel gene hypothesis. Compared to other gene prioritization tools, Beegle shows an improvement of 44% in the average recall at the top 5% prioritized genes. In a study of recent disease-gene discoveries, Beegle proposed true novel genes for 10 diseases of interest among the top 20 prioritized genes out of about 20,000 human genes (top 0.1% of the human genome). Beegle is publicly available at: http://homes.esat.kuleuven.be/~bioiuser/Beegle/

TOP

A050 - Integrated Analysis of Deep Sequencing MicroRNA and MRNA Data in Relapsed Diffuse Large B-Cell Lymphoma

Katherine Icay, University of Helsinki, Finland
Suvi-Katri Leivonen, Department of Oncology, Helsinki University Central Hospital Comprehensive Cancer Center; Genome-Scale Biology Research Program, University of Helsinki, Finland
Ilari Siren, Department of Oncology, Helsinki University Central Hospital Comprehensive Cancer Center; Genome-Scale Biology Research Program, University of Helsinki, Finland
Minna Taskinen, Department of Oncology, Helsinki University Central Hospital Comprehensive Cancer Center; Genome-Scale Biology Research Program, University of Helsinki, Finland
Chengyu Liu, Genome-Scale Biology Research Program, University of Helsinki, Finland
Rainer Lehtonen, Genome-Scale Biology Research Program, University of Helsinki, Finland
Sampsa Hautaniemi, Genome-Scale Biology Research Program, University of Helsinki, Finland
Sirpa Leppä, Department of Oncology, Helsinki University Central Hospital Comprehensive Cancer Center; Genome-Scale Biology Research Program, University of Helsinki, Finland

Short Abstract: Diffuse large B-cell lymphoma (DLBCL) is the most common lymphoma in adults. Although 60-70% of the patients can be cured with standard treatment, a significant portion do relapse with more treatment-resistant tumors. To understand the biological processes underlying relapsed tumors and their poor prognosis, we analyse microRNAs in seven matched, tumor sample pairs taken from patients before immunochemotherapy (RCHOP) treatment and after relapse. Our aim is two-fold: (1) to discover differentially expressed microRNAs and their contribution to more treatment-resistant tumors, and (2) to discover consistently high or low expressed microRNAs and their contribution to initial treatment resistance.

Studying the regulatory roles of small non-coding RNAs (smallRNAs) is challenging because it requires both smallRNA and mRNA data. We thus developed a computational workflow to standardize and automate the processing, integration, and analysis of deep sequenced data for both data types. Our analysis also extends novel miRNA discovery to include other smallRNAs with putative miRNA-like function (e.g. small-nucleolar RNAs).

13 differentially expressed, 24 high-expressed, and 177 low-expressed miRNAs were identified in our dataset. The consistently expressed miRNAs were filtered with a control set of 8 non-malignant cells. One of ten novel miRNA regions found to be differentially expressed annotates to SNORD71, a snoRNA with miRNA-like features and a predicted role in methylation. Computational analysis of the 8 miRNAs downregulated in relapsed DLBCL identified target genes enriched in key, cell survival pathways. Indeed, upregulation of these genes was linked to poor overall survival. Experimental validation further supports these miRNAs as potential novel therapeutic markers.

TOP

A051 - Computational Analysis of the Microenvironment in Glioblastoma

Suvi Luoto, BioMediTech, University of Tampere, Finland
Kirsi Granberg, 1) BioMediTech, University of Tampere 2) Department of Signal Processing, Tampere University of Technology, Finland
Juha Kesseli, BioMediTech, University of Tampere, Finland
Matti Nykter, BioMediTech, University of Tampere, Finland

Short Abstract: Interactions between various components in the tumor microenvironment and dysregulated immune responses are thought to play important roles in cancer development. To better understand the role of immune cells in tumor pathogenesis and destruction, we computationally model the microenvironment of an aggressive brain tumor glioblastoma multiforme (GBM).

We downloaded GBM patient RNA-seq data from the Cancer Genome Atlas (TCGA). Using cluster analysis, we identified 27 clusters, each containing 10 to 933 genes that show a statistical enrichment of immune response related gene ontology terms. Utilizing a panel of RNA-seq data from normal cell types, we constructed regression models to characterize the expression profiles of GBM samples in the clusters of interest as linear combinations of normal cell and reference GBM expression profiles. This was done using non-negative linear regression that results in a sparse solution for the model coefficients. Simulated data was used to validate that the regression model coefficients accurately reflect the contributions of normal cell types to the expression profiles of tumor samples. Based on the regression analysis, we were able to uncover high variability in the composition of microenvironment across the TCGA cohort, suggesting diverse immune responses in tumors. Taken together, our analysis provides a detailed characterization of the microenvironment and enables stratification of the patients based on tumor induced immune responses.

TOP

A052 - Drugging the DNA damage response

Frances Pearl, The University of Sussex, United Kingdom
Amanda Schierz, Bluefool Innovations, United Kingdom
Bissan Al-Lazikani, ICR, United Kingdom
Simon Ward, University of Sussex, United Kingdom
Laurence Pearl, Unversity of Sussex, United Kingdom

Short Abstract: The DNA damage response (DDR) is essential for maintaining the genomic integrity of the cell, and its disruption is one of the classical 'hallmarks of cancer'. Historically in the treatment of cancer, defects in the DDR have been exploited therapeutically using radiation therapies or genotoxic chemotherapies. However recently, protein components of the DDR systems have been identified as promising avenues for targeted cancer therapeutics. Here, we present an in-depth computational analysis of 450 expert-curated human DDR genes.
We have examined the deregulated components of the DDR using multi-platform cancer data and through data-driven evaluation of chemical activities data, 3D structure, and analysis of interaction and genetic networks, we have identified potential novel therapeutic targets.

TOP

A053 - A Systematic Approach to Integrate Gene Expression Profiling and Copy Number Variation Data

Sadaf Mughal, German Cancer Research Center, Germany
Thomas Wolf, German Cancer Research Center, Germany
Christina Geörg, German Cancer Research Center, Germany
Stephan Wolf, German Cancer Research Center, Germany
Marcus Renner, Institute of Pathology, University Hospital , Germany
Benedikt Brors, German Cancer Research Center , Germany
Gunhild Mechtersheimer, Institute of Pathology, University Hospital, Germany

Short Abstract: Genomic instability is common among cancers, and genes present in the aberrant chromosomal regions often display deregulated expression. Genes deregulated by copy number alterations can facilitate the molecular characterization of tumor sub-types and discovery of new drug targets. We present here a novel approach to study concordant genome-wide copy number profiles and gene expression. Genes selected by this method were able to classify tumors into the histologically defined subgroups.

Soft-tissue sarcomas are a heterogeneous, complex group of aggressive mesenchymal tumors of difficult classification. In the present study, 79 untreated, primary soft-tissue sarcoma samples and 15 sarcoma cell lines representing eight different histopathological subtypes were analyzed for copy number variation using array-based comparative genomic hybridization and expression profiling conducted using HumanHT-12 v3 beadarray with genome-wide transcriptional coverage. Probes generated from two different array platforms were matched and features were selected using pair-wise comparisons of fold change, area under curve and tukey HSD test among the eight subtypes. From nearly 22k matching features, we were able to narrow down to list 50 genes having concurrent copy number change and aberrant gene expression profile.

Integration of copy number aberrations with gene-expression profiles in a heterogeneous group of high-grade soft-tissue sarcomas revealed marked patterns among different subtypes. In conclusion, this framework integrates copy number variations and gene expression profiles obtained from different array-based platforms, identifies genes that may serve as prognostic biomarkers and used for molecular characterization of subgroups in heterogeneous samples.

TOP

A054 - Prediction of HIV-2 Coreceptor Usage with Support Vector Machines

Matthias Döring, Max Planck Institute for Informatics, Germany
Nuno Taveira, Instituto Superior de Ciências da Saúde Egas Moniz (ISCSEM) Faculty of Pharmacy, Universidade de Lisboa, Portugal
Pedro Borrego, Instituto Superior de Ciências da Saúde Egas Moniz (ISCSEM) Faculty of Pharmacy, Universidade de Lisboa, Portugal
Nico Pfeifer, Max Planck Institute for Informatics, Germany

Short Abstract: HIV-2 is a type of human immunodeficiency virus that is prevalent in Western Africa and specific countries in Europe. In contrast to HIV-1, fewer treatment options are available for HIV-2. Drugs blocking the CCR5-coreceptor might be a new treatment option for individuals infected with HIV-2. Before prescription of CCR5-antagonists it is necessary to determine whether the viral population can only use the CCR5-coreceptor for cell entry (R5) or also the CXCR4-coreceptor (X4-capable). For HIV-2, X4-capable viruses are less susceptible to antibody neutralization than R5-viruses. Treating viral populations that contain X4-capable and R5 viruses with CCR5-antagonists could lead to a decrease in R5 viruses and promote disease progression.

We developed an SVM-based prediction engine utilizing the amino acid sequence of the V3 region of the gp105 surface glycoprotein to identify whether a viral population is mainly R5 or X4-capable. The data set consists of 89 R5 and 49 X4-capable sequences from the LANL HIV database, literature, and personal communications, which were aligned to the HIV-2 reference MAC239 using the HIV-Align tool. As kernel functions we used linear, polynomial, and RBF kernels, as well as edit kernels based on amino acid substitution matrices.

Using 10 runs of 10-fold cross-validation, we found that there were no considerable differences between the linear (AUC 0.935) and the other kernel functions. This suggests that individual amino acids are already predictive of coreceptor usage in HIV-2, rather than more complex patterns of mutations. In the future, we want to extend the model onto other regions of gp105.

TOP

A055 - Comparative analysis of Huntington's and Parkinson's disease transcriptome in post-mortem human brain identifies putative pan-neurodegenerative disease gene signature

Adam Labadorf, Boston University, United States
Andrew Hoss, Boston University School of Medicine, United States
Tom Beach, Sun Health, United States
Richard Myers, Boston University School of Medicine, United States
Seung Hoan Choi, Boston University School of Public Health, United States

Short Abstract: Huntington's Disease (HD) and Parkinson's Disease (PD) are neurodegenerative diseases that selectively affect different types of neurons in the brain yet share some common symptoms. While changes to the transcriptome in human brain have been characterized in both diseases separately, a comparative study of the two conditions may yield insight into the specific signatures of each disease as well as implicate common genes and pathways in neurodegeneration. In this study, we present a detailed analysis of the transcriptional changes in 29 HD and 29 PD human post-mortem prefrontal cortex compared to 50 neuropathologically normal controls using high-throughput mRNA sequencing. Comparing the differentially expressed genes from the two conditions reveals functional gene signatures unique to each disease, a set of genes common to both, and a group of genes discovered only when the two disease datasets are analyzed together. The disease-specific signatures differ markedly in their function, transcriptional regulation, and miRNA targeting patterns as determined by pathway and geneset enrichment analysis. The common differentially expressed genes between PD and HD are enriched in the NFkB pathway, implicating the immune response in both conditions despite disease-specific differences in the degree of neurodegeneration of this brain area. Differentially expressed genes identified by contrasting controls to both diseases together are highly enriched in ribosomal components, suggesting ribosomal alterations may be a common feature of neurodegenerative disease. Together, these results suggest important similarities and differences between HD and PD that may be useful for characterizing the pan-neurodegenerative disease phenotype.

TOP

A056 - Challenges in Using Gene Homology Data for Identifying Cross Species Disease Models

Susan Bello, The Jackson Laboratory, United States
Li Ni, The Jackson Laboratory, United States
Mary Dolan, The Jackson Laboratory, United States
Richard Baldarelli, The Jackson Laboratory, United States
James Kadin, The Jackson Laboratory, United States
Janan Eppig, The Jackson Laboratory, United States
Judith Blake, The Jackson Laboratory, United States

Short Abstract: A major goal of Mouse Genome Informatics (MGI, www.informatics.jax.org) is to facilitate the use of mouse as a translational model for human disease research. Gene homologies between mouse and human provide immediate model candidates where a human gene - to - disease association is known and mutations in the homologous mouse gene exist. Additionally, mutations in mouse models that phenocopy human diseases suggest new candidate genes for diseases.
Many gene orthology and homology prediction algorithms exist reflecting the current state of knowledge. MGI currently uses HomoloGene as its source of homology data and is working to incorporate orthology predictions from the HUGO Gene Nomenclature Committee (HGNC). In evaluating the intersection of HomoloGene and HGNC representations, over 10,700 homology class sets were found to be identical in both sources. Each source also contained unique sets (HomoloGene 2389, HGNC 15059). Most unique classes from HGNC contain singleton human pseudogenes or RNA genes. Further, sets were identified where one source had classes containing both mouse and human genes while the other source had only single species classes (HomoloGene 5409, HGNC 500). Finally, there were 154 sets where both sources included mouse and human genes but class membership differed. Integration of multiple homology sources presents challenges and opportunities for mapping between human and mouse genomes in the evaluation of potential mouse models for human diseases. We will present our strategy for a combined use of HomoloGene and HGNC human-mouse homology data in presenting mouse models of human disease in the MGI interface.

TOP

A057 - PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins

Hiba ABI HUSSEIN, University Paris Diderot, Sorbonne Paris Cité, UMRS-973, MTi, Paris, France, France
Alexandre BORREL, University Paris Diderot, NSERM, UMRS-973University Paris Diderot, Sorbonne Paris Cité, UMRS-973, MTi, Paris, France, MTi, Paris, France, France
Colette GENEIX, University Paris Diderot, Sorbonne Paris Cité, UMRS-973, MTi, Paris, France, France
Michel PETITJEAN, University Paris Diderot, Sorbonne Paris Cité, UMRS-973, MTi, Paris, France, France
Leslie REGAD, University Paris Diderot, Sorbonne Paris Cité, UMRS-973, MTi, Paris, France, France
Anne-Claude CAMPROUX, University Paris Diderot, Sorbonne Paris Cité, UMRS-973, MTi, Paris, France, France

Short Abstract: Therapeutical molecules bind to preferred active sites, which are mainly pockets located at protein surface. Therefore, pockets estimation and characterization is a major issue in drug target discovery. Predicting protein pocket's ability to bind drug-like molecules with high affinity, i.e., druggability, is a key step of compound clinical progression projects. Indeed, "drug-like molecules" are small molecules with particular properties such as beeing orrally bioavailable. Currently, computational druggability prediction models are attached to one unique pocket estimation method despite pocket estimation uncertainties. Here, we present "PockDrug-Server" that predicts pocket druggability, efficient on both: estimated pockets guided by the ligand proximity (extracted by proximity to a ligand from a holo protein structure using several thresholds) and estimated pockets not guided by the ligand proximity (based on amino atoms that form the surface of potential binding cavities). PockDrug-Server is based on a statistical model corresponding to a combination of 7 linear discriminant analysis model using 9 pocket descriptors to provide a mean druggability. It provides consistent druggability results using different pocket estimation methods. It is robust with respect to pocket boundary and estimation uncertainties, thus efficient using apo pockets that are challenging to estimate. It clearly distinguishes druggable from non druggable pockets using different estimation methods and outperformed recent druggability models for apo pockets. It can be carried out from one or a set of apo/holo proteins. PockDrug-Server is publicly available at: http://pockdrug.rpbs.univ-paris-diderot.fr.

TOP

A058 - Functional annotation of obesity SNPs underlying potential parent of origin effects

Xuanshi Liu, University of Leipzig, Germany
Anke Tönjes, Department of Medicine,University of Leipzig, Germany
Michael Stumvoll, IFB AdiposityDiseases & Department of Medicine,university of Leipzig, Germany
Peter F. Stadler, Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany
Yvonne Böttcher, IFB AdiposityDiseases, University of Leipzig, Germany

Short Abstract: Genome-Wide Association Studies (GWAS) were successfully applied to discover genetic variants associated with obesity, however some heritability were unexplained. Analyses of potential parent of origin effects (POE) may provide further insights into genetic mechanisms of obesity. The aim of study was to identify novel functional SNPs/regions which would contribute to obesity.
Genome-wide genotypes from Sorbs (N=525), a German self-contained population, were phased using AlphaImpute. Three different GWAS were applied in PLINK: (i) standard association, (ii) considering paternal and (iii) maternal alleles. Ten top tagging SNPs from paternal and maternal GWAS were selected respectively. An R package FunciSNP was used to identify correlated SNPs.
We identified 109 SNPs and 180 SNPs correlated with tagging SNPs underlying paternal and maternal POE respectively. Two transcription factors SF1 (steroidogenic factor 1) and LRH1 (liver receptor homolog-1) putatively bound at rs1204880 underlying paternal POE. SF1 involves in determining sex and differentiation while LRH1 affects bile acid metabolism and glucose homeostasis. Rs1204880 was highly correlated (r2 =1) to tagging SNP rs942459 and located within the putative promoter of PADI6 (Peptidyl Arginine Deiminase, Type VI) which may associate with reorganizing cytoskeletal in egg and during early embryo development. MicroRNA binding sites were identified at rs11180547 and rs4562666 underlying maternal POE.
Incorporating high-throughput epigenetic data, variants from 1000 Genomes and various genomic databases into POE specific GWAS may reveal novel putative SNPs located in potentially functional regions and maybe involve in obesity.

TOP

A059 - Galahad: a web server for the analysis of drug-induced gene expression data

Griet Laenen, KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics & iMinds Medical IT Department, Belgium
Amin Ardeshirdavani, KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics & iMinds Medical IT Department, Belgium
Yves Moreau, KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics & iMinds Medical IT Department, Belgium
Lieven Thorrez, KU Leuven Kulak, Interdisciplinary Research Facility Life Sciences, Department of Development and Regeneration, Belgium

Short Abstract: The pharmaceutical industry is facing unprecedented productivity challenges. Attrition rates have risen sharply, especially in late-phase clinical trials. With safety and efficacy being the main bottlenecks, a better knowledge of a candidate drug's mode of action and its off-target effects could be of substantial value to drug development. DNA microarray technology enables us to observe the effect of drug treatment on the activity of all genes simultaneously, and thus can provide valuable information for identifying drug-target interactions and resulting effects prior to clinical trials.
We have developed an easy-to-use web server called Galahad, for the in-depth exploration of a drug's mode of effect based on gene expression changes following treatment. Our software provides multiple tools needed for gaining new insights into the biological effects of a drug by combining Affymetrix gene expression data preprocessing, quality assessment and exploratory analysis, genome-wide drug target prioritization, differential expression analysis and pathway, as well as disease phenotype enrichment. The core of Galahad is a network-based analysis method for drug target prediction, integrating the expression data with functional protein association information. Our website can be accessed at https://galahad.esat.kuleuven.be. An example data set is provided for which the results are directly available.

TOP

A060 - hierGWAS – a Bioconductor package for hierarchical testing in Genome Wide Association Studies

Laura Buzdugan, ETH Zürich, Switzerland
Peter Bühlmann, ETH Zürich, Switzerland

Short Abstract: Even though Genome Wide Association Studies genotype a very large number of single nucleotide polymorphisms (SNPs), the data is often analyzed one SNP at a time by using the Armitage Trend test. Such marginal one SNP at a time methods are often poor in terms of predictive power for prognosis of disease and they are also over-simplistic in terms of interpretation.

We propose a procedure in which all the SNPs are analyzed jointly. As a major novelty, our method yields p-values for assessing significance of single SNPs or groups of SNPs: it controls the Family Wise Error Rate (FWER) in multiple testing and it automatically leads to a data-driven resolution level for refining clusters of SNPs to smaller groups or single markers. The interpretation of single or groups of SNPs in a simultaneous model is different and “stronger” than using a marginal approach, and it leads to better predictions as well.

The method is implemented in the hierGWAS Bioconductor package. It has a modular design with separate functions for each step of the analysis: clustering, regression and statistical testing. This gives the user the opportunity to change particular aspects of the analysis, for example one can choose a different clustering approach, or a different regression model. The final output of the analysis is a list of SNP groups, along with their p-values. Additionally, the user can obtain the amount of variation explained by a group of SNPs or individual markers.

TOP

A062 - Composition and temporal stability of the gut microbiota in older persons

Denise Lynch, University College Cork, Ireland
Ian Jeffery, University College Cork, Ireland
Paul O\'Toole, University College Cork, Ireland

Short Abstract: The composition and function of the human gut microbiota has been linked to health and disease. We previously identified correlations between habitual diet, microbiota composition gradients and health gradients in an un-stratified cohort of 178 elderly subjects. To refine our understanding of diet-microbiota associations and differential taxon abundance, we adapted an iterative bi-clustering algorithm (iBBiG) and applied it to microbiota composition data from 732 faecal samples from 371 ELDERMET cohort subjects including longitudinal samples. We thus identified distinctive microbiota configurations associated with ageing in both community and long-stay residential care elderly subjects. Mixed-taxa populations were identified which had clinically distinct associations. Microbiota temporal instability was observed in both community-dwelling and long-term care subjects, particularly in those with low initial microbiota diversity. However the stability of the microbiota of subjects had little impact on the directional change of the microbiota as observed for long-stay subjects who display a gradual shift away from their initial microbiota. This was not observed in community-dwelling subjects. This directional change was associated with duration in long-stay. Changes in these bacterial populations represent the loss of the health-associated and youth-associated microbiota components and gain of an elderly-associated microbiota. Interestingly community-associated microbiota configurations were impacted more by the use of antibiotics than the microbiota of individuals in long-term care, as the community-associated microbiota showed more loss but also more recovery following antibiotic treatment. This improved definition of gut microbiota composition patterns in the elderly will better inform the design of dietary or antibiotic interventions targeting the gut microbiota.

TOP

A063 - Application of large-scale text-mining and curation for extracting neuronal electrophysiology data

Dmitry Tebaykin, University of British Columbia, Canada
Shreejoy Tripathy, University of British Columbia, Canada
Brenna Li, University of British Columbia, Canada
Kristofer Anderson, University of Aberdeen, United Kingdom
Delaram Abdollahzadeh, University of British Columbia, Canada
Paul Pavlidis, University of British Columbia, Canada

Short Abstract: Recently, there has been a major effort by neuroscientists to systematically organize and integrate vast quantities of brain data. Here, as part of the NeuroElectro project (www.neuroelectro.org), we employ large scale text-mining, supplemented with manual curation, to extract quantitative measurements from >100K neuroscience full-text articles.

Specifically, we use heuristic and machine learning approaches to identify sentences containing electrophysiological data or metadata entries. We then break these sentences down into smaller pieces and collect the desired information using regular expression rule based methods. For example, we identify ion concentrations (e.g., Ca2+, Mg2+, etc.) in chemical solutions with 89% accuracy. While major ion concentrations are known to have specific qualitative effects on neurons, they are not well quantified for various neuron types.

Another domain that we are mining is neuronal synaptic plasticity, which describes the strengthening (or weakening) of connections between neurons. Although, our regular expression based approach can identify this information with some degree of accuracy, we have found that it is absolutely necessary to manually curate the text-mined data. For this purpose we have employed a graphical interface (NLP BRAT) and trained curators who are able to efficiently verify the raw text-mined values.

Ultimately, this extracted data will enhance the existing NeuroElectro database of normalized neuronal electrophysiology values. It would further allow us to link electrophysiological diversity across neuron types to corresponding differences in gene expression levels and disease phenotypes.

TOP

A064 - Rationally designed drug blending as a mechanism to overcome drug resistance in cancer

Francisco Martinez, CRG/CNAG, Spain
John P. Overington, European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, United Kingdom
Bissan Al-Lazikani, Institute of Cancer Research, Cotswold Road, Sutton, Surrey, United Kingdom, United Kingdom
Marc A. Marti-Renom, CNAG/CRG, Spain

Short Abstract: Drug resistance is one of the major problems in cancer treatment. Rapid mutation and selective pressure can efficiently select drug-resistant mutants, although there are many mechanisms for drug resistance, a classic mechanism is due to coding mutations in the drug-target binding site. Some well-characterized examples are the resistance to BRAF inhibitors in the BRAF (V600E) positive melanomas, resistance to gefitinib in non-small-cell lung cancers due to point mutations in epidermal growth factor receptor, or resistance to topotecan caused by topoisomerase I mutations.

To systematically analyse the mutational landscape that can potentially cause resistance to targeted therapies, we have studied the current targets of small-molecule treatments for 30 different classes of cancer. Using the mutational frequencies from Alexandrov et al. (1) we have generated the 3D models for each of most likely mutants for the current drug targets. Next, for all the 3D-generated mutants, we have predicted the resistance potential to the current drug treatments. Finally, for each of the mutants evaluated as likely to confer resistance, we have computationally defined a set of molecules targeting the same protein, which would theoretically overcome drug resistance. This study aims to reduce the difficulties in the choice of the optimal treatment and subsequently, it is a step further in the development of the personalized medicine for the treatment of cancer.

1. Alexandrov, L.B., et al. (2013) Signatures of mutational processes in human cancer. Nature, 500, 415-421

TOP

A065 - A comprehensive database of cell-type specific marker genes for the mammalian brain

Ogan Mancarci, University of British Columbia, Canada
Lilah Toker, University of British Columbia, Canada
Shreejoy Tripathy, University of British Columbia, Canada
Paul Pavlidis, University of British Columbia, Canada

Short Abstract: Large-scale gene expression analyses are often used to study complex neuropsychiatric disorders. Due to technical complications, such analyses are commonly based on bulk brain tissue. However, interpreting changes in gene expression in bulk tissue can be challenging due to the cellular heterogeneity of the tissue. For example, expressional changes can reflect alterations in cellular proportions (e.g. loss of dopaminergic cells in Parkinson's disease). Such changes could potentially be inferred from the expression of cell-type specific marker genes, however, reliable marker genes do not currently exist for the majority of brain cell-types.

Here we present a database of cell-type enriched genes based on cell-type specific expression data from mouse brain. The data was compiled from ~20 sources each including several cell-types. The datasets were manually curated to ensure the quality and the cellular specificity of the data. We developed a clustering-based method to select reliable cell-type marker genes, taking the spatial co-existence of the cell-types into account. We next used eigengene values of the relevant cellular markers to estimate cellular proportions in bulk brain tissues.

We found several cellular markers that were previously believed to be not expressed in the brain. In addition, we show that some of the widely used markers lack specificity and/or sensitivity and should be used with caution.
Using the human homologues of the marker genes, we show they can further be used to inform cellular proportions and cell-type specific expression changes in human-brain expression data such as different diseases and developmental stages.

TOP

A066 - In-silico prediction of Caco-2 monolayer permeability through deep learning

Moonshik Shin, KAIST, Korea, Rep
Dongjin Jang, KAIST, Korea, Rep
Doheon Lee, KAIST, Korea, Rep

Short Abstract: Various in-vitro assays have been developed to measure the absorption of orally administrated drugs. The most commonly used in-vitro model for screening the human intestinal permeability of drugs are Caco-2 cell monolayers, which is derived from human colorectal carcinoma. Caco-2 cell monolayers display many functional properties of the intestinal epithelial cell barrier. Due to these advantages, Caco-2 cells are considered a surrogate for measuring human intestinal permeability of diverse molecules. Construction of in-silico models that predicts Caco-2 cellular permeability of molecules may support the early screening stage of drug development. In our research, Deep Neural Network (DNN) approach based binary classifier (highly and moderate-poor permeable) has been constructed to predict the permeability of molecules using the 674 Caco-2 cell dataset. Dropout regularization is applied to solve the over-fitting problem and for the non-linear activation Rectified linear unit (ReLU) is adopted instead of sigmoid to improve the overall performance. 209 chemical descriptors driven from CDKDescriptor are used. In complete data, DNN based binary classifier provides classification accuracy of 88.14% for test sets (10-fold cross validation). DNN based classifier outperforms the previously developed LDA based permeability prediction model. The results of this study suggest that the constructed DNN based binary classifier is suitable for predicting cellular permeability of diverse chemical molecules in Caco-2 cell lines, which may be a useful prediction model in screening new drugs and other bioactive compounds.

TOP

A067 - Inferring new drug indications using the complementarity between clinical disease signatures and drug effects

Dongjin Jang, KAIST, Korea, Rep
Doheon Lee, KAIST, Korea, Rep

Short Abstract: Drug-repositioning is defined as the process of finding new uses outside the scope of the original indications for existing drugs. Its importance has been dramatically increasing recently due to the enormous increase in new drug discovery cost. However, most of the previous molecular-centered drug-repositioning works have some difficulty in reflecting the end-point physiological activities of drugs because of the inherent complexity of human physiological systems.
Here, we suggest a novel computational framework to infer alternative indications of marketed drugs using Electronic Health Records (EHRs)-based clinical information which reflects the end-point physiological results of drug on human biological activities. In this work, the key concept is the complementarity between clinical disease signatures and clinical drug effects. With this concept, we establish clinical disease signature and clinical drug effect vectors by applying statistical analysis and literature mining. We then assign a repositioning score to each disease-drug pair by the calculation of complementarity between clinical states (“up” or “down”) of disease signatures and clinical effects (“up” or “down”) of drugs.
We examined prediction results by using statistical experiments (enrichment verification, hyper-geometric and permutation test P<0.004) in two benchmark datasets (CTD and Clinical trials) and demonstrated evidences for those with already published literature.
The results show that EHR-based clinical information is a feasible data and that complementarity is a potentially predictive concept in drug-repositioning research. It makes the proposed approach a useful to identity novel disease-drug relationships that have a high probability of being biologically valid.

TOP

A068 - Prediction of drugs having opposite effects on disease genes

Hasun Yu, KAIST, Korea, Rep
Sungji Choo, KAIST, Korea, Rep
Junseok Park, KAIST, Korea, Rep
Jinmyung Jung, KAIST, Korea, Rep
Yeeok Kang, KAIST, Korea, Rep
Doheon Lee, KAIST, Korea, Rep

Short Abstract: Motivation: Developing novel uses of approved drugs, called drug repositioning, can reduce costs and times in traditional drug development. Network based approaches have presented a promising result in this field. However, even though various types of interactions such as activation or inhibition exist in drug-target interactions and molecular pathways, most of previous network based studies did not consider this information.
Results: Here, we propose a computational method for Prediction of Drugs having Opposite effects on Disease genes (PDOD). With the consideration of ‘effect type’ and ‘effect direction’ in paths from a drug to a disease, PDOD discovered drugs likely to restore altered states of disease genes. Our approach not only considered various types of drug-target interactions but also resolved conflicts occurred in directed molecular pathways. Results from a case study show that our approach reflecting ‘effect type’ and ‘effect direction’ information outperformed other methods without this information. In addition, many of new indications of existing drugs we predicted in a case study are supported by previous literature. We provide a web service that researchers can submit genes of interest with their altered states and will obtain a drug set seeming to have opposite effects on input genes.
Availability: PDOD is implemented in online. You can submit a gene list with changed state and receive a candidate drug set for therapies at our web page.

TOP

A069 - Application of an automatic next-generation analysis pipeline for pedigrees identified novel variants for psychiatric disorders

WEI YUN TSAI, National Health Research Institutes, Taiwan
Chen-Yu Kang, National Health Research Institutes, Taiwan
Po-Ju Yao, National Health Research Institutes, Taiwan
Hui-Ju Tsai, National Health Research Institutes, Taiwan
Chia-Hsiang Chen, Chang Gung Memorial Hospital, Taiwan
Ren-Hua Chung, National Health Research Institutes, Taiwan

Short Abstract: For complex disease studies, sequencing a few pedigrees with multiple affected members based on next-generation sequencing (NGS) has become a powerful approach to identifying rare alleles associated with the disease. Moreover, imputing rare variants in families can be more accurate than the imputation for unrelated samples, which is cost-efficient to increase the sample size. Automating the NGS analysis using family data thus becomes important to facilitate the analysis. We developed an NGS analysis pipeline, FamPipe, for family data with complex diseases. It includes commonly used family analysis functions, such as identity-by-descent (IBD) sharing among affected relatives, imputation, linkage, and association analyses. We applied FamPipe to two pedigrees with 17 and 5 individuals, respectively. Specifically, some members in one pedigree were diagnosed with schizophrenia and some members in the other pedigree were diagnosed with bipolar disorders. Five individuals from each family were whole-exome sequenced with Illumina HiSeq and all individuals were genotyped with Affymetrix SNP Array 6.0. Individuals not sequenced were imputed. Variants were filtered based on several criteria: (1) call rate>90%; (2) proportion of IBD sharing>0.8; (3) GERP score≥2; (4) PolyPhen score>0.5; (5) missense or nonsense mutations. Variants were retained if all criteria were met. There were 6,54,518 variants in the raw data, and 57 variants remained after the filtering. We identified six candidate genes among these 57 variants that have been reported for psychiatric disorders.

Interestingly, variants in 3 candidate genes (BIN1,FCRL3, and LRP1B) are novel variants. Follow-up studies are required to verify their roles in the disorders.

TOP

A070 - Building text-mining framework for biological relation extraction using deep learning

Jaehyun Lee, KAIST, Korea, Rep
Dongjin Jang, KAIST, Korea, Rep
Kwangmin Kim, KAIST, Korea, Rep
Doheon Lee, KAIST, Korea, Rep

Short Abstract: The scientific literature is a rich resource for information retrieval on the biological knowledge. Nevertheless, the unstructured textual data in the research articles makes it difficult to access the information with computer-aided systems. Text-mining is one of the solution that can transform unstructured information in the text into database content, and most of the approaches are based on the machine learning models. Since these approaches require high-dimensional features, the performance of the model is heavily dependent on the selection of features. However, it is usually difficult and labor-intensive to choose good features, because feature extraction requires prior knowledge and ingenuity of human experts. Here, we suggest a novel framework to extract biological relations from the texts by using hierarchical text features that enhance the effectiveness of relation extraction model. The proposed framework is composed of two parts, node and edge detection, using deep belief networks. Each part is based on the hierarchical text features learned by Gaussian-Bernoulli restricted Boltzmann machine (GBRBM). In this work, we performed gene-cancer relation extraction task as a pilot study. The classification model was trained based on both GE09 corpus from BioNLP'09 Shared Task and CoMAGC corpus. The results show that our model achieved better performance than other handcrafted feature-based approaches. The evaluation results suggest that deep belief networks offers the optimized and generalized hierarchical text features for the large-scale text mining.

TOP

A071 - Comparative sequencing of renal cancer biopsy pairs reveals co-existing subclones

Ariane Hofmann, ETH Zürich, Switzerland
Christian Beisel, ETH Zürich, Switzerland
Jonas Behr, ETH Zürich, Switzerland
Peter Schraml, Institute for Surgical Pathology, University Hospital Zurich, Switzerland
Holger Moch, Institute for Surgical Pathology, University Hospital Zurich, Switzerland
Niko Beerenwinkel, ETH Zürich, Switzerland

Short Abstract: Due to the process of mutation and selection, a tumor is composed of various subclones with different genotypes and phenotypes. It is crucial to investigate the subclonal structure and to understand the dynamics of their interplay for improving treatment success. We analyzed two biopsies of each renal cell carcinoma (RCC) together with a matched normal sample from the same individual. Our analysis is based on next-generation sequencing data of the exomes of 16 RCCs from 16 patients. We performed variant calling and pairwise comparison of the variations found in the two tumor biopsies. On average two thirds of the mutations in a patient were private to one of the two samples. This finding points towards a high intra-tumor heterogeneity. However, this number might be confounded by the variant calling and filtering process. Indeed, when checking whether the private mutations would have at least one read in the respective other sample which supports the private variant, we obtained a different picture: Counting these private variants with support in the other sample as potentially shared mutations, the fraction of private mutations decreased to less than one third. Pairwise comparison of the frequencies of the SNVs showed that some of them differed remarkably. The results showed that most ancestor clones might still exist at varying frequencies in the two samples. The private mutations represent new clones that emerged in some samples. Ultra-deep sequencing of those genes harboring the potentially shared mutations will enable a more detailed investigation of the subclonal tumor structure.

TOP

A072 - Network roles and sensitivity to germline variants distinguish groups of genes associated to diseases of different nature

Janet Piñero, IMIM-UPF, Spain
Ariel Berenstein, CONICET-UBA, Argentina
Abel Gonzalez-Perez, IMIM-UPF, Spain
Ariel Chernomoretz, UBA-CONICET-FIL, Argentina
Laura I. Furlong, IMIM-UPF, Spain

Short Abstract: The study of network and molecular properties of disease-related genes has served the double purpose of understanding disease mechanisms and identifying novel disease gene candidates. Here, we explored the classic (such as degree and betweenness) and cartographic network properties and the tolerance to deleterious germline variants of groups genes involved in different classifications of diseases across six different protein interaction networks (PINs). We found that the network properties of disease genes are mostly due to cancer genes. We then focused on more homogeneous disease categories, such as autosomal dominant and recessive Mendelian disease genes and cancer driver genes, and show that cancer drivers tend to occupy the most central roles in the PIN, in terms of number of connections and of number of clusters they participate. They are significantly more sensitive than the average to deleterious germline variants. Genes associated to recessive Mendelian diseases tend to be enclosed within modules and are significantly more tolerant to likely deleterious variants than other disease genes and other genes in the network. Genes related to dominant Mendelian diseases occupy intermediate positions in the network and are less sensitive to likely deleterious germline variants, although still more than average. In summary, we show that different classifications of diseases condition the network and molecular properties of disease genes. Computational methods aimed at prioritization of candidate disease genes or at identification of cancer driver genes could benefit from our results to improve their performance.

TOP

A073 - Comparison of indel calling methods for identifying cancer genes from whole exome sequencing data

Jingu Lee, School of Information and Communications, Gwangju Institute of Science and Technology, Korea, Rep
Seungchul Lee, School of Information and Communications, Gwangju Institute of Science and Technology, Korea, Rep
Se-hoon Lee, Division of Hematology/Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Korea, Rep
Hyunju Lee, School of Information and Communications, Gwangju Institute of Science and Technology, Korea, Rep

Short Abstract: As genetic variants are highly associated with cancer, many researches have been conducted to find the accurate variants related with cancer. Identification of structural variants from the whole exome sequencing (WES) data, especially insertion and deletion (indels), is an important research topic. Various methods for indel calling have been developed. One of the challenges in indel identification in cancer is to detect the somatic variants that exist only in tumor tissues. Because some indel calling tools do not support a somatic mode, a statistical test with the number of variants in cancer samples and normal samples are required. Another challenge is the reduction of indels calling errors. Some of the retrieved indels from WES are false positives, which occurred from a false sequence alignment. To reduce the false positives, machine learning techniques can be applied. Several biological features are used for training a machine learning model. In this work, we first compare several indel calling methods in WES data, and then proposed a method to identify somatic indels in cancer genomes.

TOP

A074 - Tumor Induced Reprogramming Drives Lymph Node Stromal Cell Transformation

David Shorthouse, University of Cambridge, United Kingdom
Angela Angela Riedel, University of Cambridge, United Kingdom
Jacqui Shields, University of Cambrdige, United Kingdom
Benjamin Hall, University of Cambrdige, United Kingdom

Short Abstract: Lymph node stromal cells (LNSCs) are heavily involved in lymph node function, including the regulation of immunity, structural support for other cells, and the facilitation of cell-cell interactions within the nodes. Lymph nodes are also the preferential site of metastasis in many cancers, however, the role of LNSCs in this process have not been established. Here we present the first comprehensive gene expression analysis for a single subset of LNSCs, specifically fibroblastic reticular cells (FRCs) from resting vs. tumour draining lymph nodes. By combining expression data at different timepoints prior to metastasis with imaging data from the whole node we propose how transcriptional events drive functional changes in the lymph node. Finally, we attempt to combine these data sources into an executable model describing FRC response over time to tumour challenges.

TOP

A075 - Differential expression among tissues in morbidly obese individuals using a finite mixture model under BLUP approach

Lisette Kogelman, University of Copenhagen, Denmark
Daniah Trabzuni, Department of Molecular Neuroscience, UCL Institute of Neurology, United Kingdom
Marc Jan Bonder, University Medical Center Groningen, Netherlands
Lude Franke, University Medical Center Groningen, Netherlands
Peter C. Thomson, ReproGen - Animal Bioscience Group, Faculty of Veterinary Science, The University of Sydney, Australia
Haja N. Kadarmideen, Department of Veterinary Clinical and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

Short Abstract: Morbid obesity, the excessive accumulation of body fat, has major consequences for human health by its association with several severe diseases, like Type 2 Diabetes. The biological mechanisms behind this association are mostly unclear, however, the biological complexity of morbid obesity indicates an important role for pathogenesis arising from different tissues/organs and interactions among them. We hypothesized that differentially expressed (DE) genes between different organs in morbidly obese individuals result in a better insight in the biological mechanisms explaining the link between obesity and obesity-related diseases. We used whole-transcriptome expression levels of subcutaneous adipose tissue (SAT), visceral adipose tissue (VAT), liver, and muscle of 93 morbidly obese individuals (26 males, 67 females) who were all phenotyped for different metabolic parameters. We estimated genetic random effects of the interactions between tissues and probes using BLUP (Best Linear Unbiased Prediction) linear models correcting for gender, which were subsequently used in a finite mixture model to detect DE genes in each tissue. This approach evades the multiple-testing problem and is able to detect variation among different biological states. Preliminary results show evidence of DE genes operating within tissues, but variation in these effect sizes was similar between the four tissues. We detected 17% of the transcripts to be DE in liver, 16% in muscle, 13% in SAT and 23% in VAT (Probability (DE) > 0.9). Results will be used to gain insight into the underlying genetic and biological mechanisms by functional annotation of DE genes, detection of pathways and construction of gene networks.

TOP

A076 - A Disease Name Normalization Method for Named Entity Recognition in Biomedical Articles

Hyejin Cho, School of Information and Communications, Gwangju Institute of Science and Technology, Korea, Rep
Hyunju Lee, School of Information and Communications, Gwangju Institute of Science and Technology, Korea, Rep

Short Abstract: Biomedical articles contain information about biological entities such as genes, disease names, and chemicals. Hence, a named entity recognition (NER) technique that identifies entity names from the text is an essential component in text mining to extract biological knowledge from articles. A disease name is a basic biological entity, and disease is a major subject of biomedical researches. Identification of disease names in research articles will improve accessibility of information in the text. Hence, several NER methods for disease names have been developed.
After an NER method is applied to articles, a next step is to normalize identified disease names into pre-defined disease concepts (i.e. MeSH disease terms). In this work, we developed a new approach to normalize disease names based on a deep learning algorithm. A disease name is represented as a vector using a neural network and then the similarities between disease names are used to normalize them. We validated the proposed method using gold standard data from DNorm and applied it into abstracts from PubMed.

TOP

A077 - Systematic drug repositioning for a wide range of diseases by integrating phenotypic and molecular data

Yoshihiro Yamanishi, Kyushu University, Japan

Short Abstract: Drug repositioning is a challenging issue in medical and pharmaceutical research. In this study, we developed a new computational method to predict new drug indications for systematic drug repositioning in a framework of supervised network inference. We defined a descriptor for each drug–disease pair based on the phenotypic features of drugs (e.g., medicinal effects and side effects) and various molecular features of diseases (e.g., disease-causing genes, disease-related pathways, diagnostic markers, and environmental factors) and constructed a statistical model to predict new drug–disease associations for a wide range of diseases in the International Classification of Diseases. Our results show that the proposed method outperforms previous methods in terms of accuracy and applicability, and its performance does not depend on drug chemical structure similarity. Finally, we performed a comprehensive prediction of a large-scale drug–disease association network and described biologically meaningful examples of newly predicted drug indications for various diseases.

TOP

A078 - A transcriptomics approach for revealing cell-type proportion changes in psychiatric disorders

Lilah Toker, University of British Columbia, Canada
Ogan Mancarci, University of British Columbia, Canada
Shreejoy Tripathy, University of British Columbia, Canada
Paul Pavlidis, University of British Columbia, Canada

Short Abstract: Increasing evidence has accumulated regarding the involvement of cellular death and neuroinflammation in neuropsychiatric disorders (e.g., schizophrenia). Identifying the affected cell-types is crucial for understanding the pathophysiology of a disorder, providing new directions for future studies, and assisting in the analyses of gene expression data. Nevertheless, lack of reliable markers for majority of the brain cells and scarcity of human brain samples limits the use of direct cell counting for this purpose. Since different cell-types express distinct sets of genes, it is plausible that changes in cellular populations would result in observed transcriptional alterations in the bulk tissue. For example, loss of oligodendrocytes is likely to result in decreased transcript levels of enzymes involved in myelin synthesis. Thus, if the cell-type specific transcripts are known, changes in cellular populations can potentially be inferred from the expression data of bulk tissue. Publicly available expression data from brains of psychiatric and healthy subjects provide the opportunity to accomplish this task without the need for additional experiments.
We used a database of cell-type specific genes compiled by our group to infer changes in subtypes of neuronal and glial cells in brains of psychiatric patients. To test the robustness of our results, we repeated the analysis in four datasets based on a similar cohort of subjects. Using statistical methods, we were able to infer changes in several cellular populations, some of which were previously shown by direct cell counting methods. The inferred changes were similar across the four datasets, supporting the robustness of our method.

TOP

A079 - A new molecular signature approach for prediction of driver cancer pathways from transcriptional data

Boris Reva, Mount Sinai School Of Medicine, United States
Dmitry Rykunov, Mount Sinai School Of Medicine, United States
Andrew Usilov, Mount Sinai School Of Medicine, United States
Hui Li, Mount Sinai School Of Medicine, United States
Eric Schadt, Mount Sinai School Of Medicine, United States

Short Abstract: Assigning cancer patients to the most effective treatments requires an understanding of the molecular basis of their disease. While DNA-based molecular profiling approaches have flourished over the past several years to transform our understanding of driver pathways across a broad range of tumors, a systematic characterization of key driver pathways based on RNA data has not been undertaken.

Here we introduce a new approach to predict the status of driver cancer pathways based on weighted sums of gene expressions or signature functions derived from RNA sequencing data. To identify the driver cancer pathways of interest, we mined DNA variant data from TCGA and nominated driver alterations in seven major cancer pathways in breast, ovarian, and colon cancer tumors. The activation status of these driver pathways were then characterized using RNA sequencing data by constructing signature functions in training datasets and then testing the accuracy of the signatures in test datasets.

The signature functions perform well in separation tumors with nominated active pathways from tumors with no genomic signs of activation (average AUC equals to 0.83) systematically exceeding the accuracies obtained by the SVM method that we employed as a control approach. A typical pathway signature is composed of ~20 biomarker genes that are unique to a given pathway and cancer type. Our results confirm that driver genomic alterations are distinctively displayed at the transcriptional level and that the transcriptional signatures can generally provide an alternative to DNA sequencing methods in detecting specific driver pathways.

TOP

A080 - Combined strategy to detect somatic point mutations from circulating DNA by targeted sequencing

Nicola Casiraghi, Centre for Integrative Biology, University of Trento, Italy
Alessandro Romanel , Centre for Integrative Biology, University of Trento, Italy
Gerhardt Attard, Royal Marsden National Health Service Foundation Trust, United Kingdom
Francesca Demichelis, Centre for Integrative Biology, University of Trento, Italy

Short Abstract: We developed a computational method that combines genetic knowledge and empirical signal to readily detect and quantify somatic point mutations in cell free DNA by fully exploiting single base resolution information from targeted next generation sequencing data using patient's plasma (case) and matched germline sample (control). First, each targeted base is tested both in cases and controls for allelic fraction, local coverage and reads supporting the alternative allele(s). Controls allelic fractions distribution is built to determine the cut-off corresponding to the desired detection specificity. Second, to mitigate the impact of potential strand-bias, we implemented a combination of standard Fisher's and Odds Ratio tests with ad-hoc analysis of study cohort reference/alternative strand proportions distribution. Third, control samples are exploited to build a genomic locus-specific error model to estimate the probability that observed case allelic fraction is indeed evidence of a somatic event. Fourth, comparison of expected versus observed ratios of non-synonymous and synonymous substitution rates in targeted control genes is adopted as additional quality check. Last, if the targeted design allows for case tumor content and local somatic copy number state estimations, the method also controls for point mutation detection suitability stratified by locus coverage (false negative rates). The robustness of our combined strategy was tested across a range of coverage depths by in-silico down-sampling analysis. We will present the strategy efficacy on 46 plasma samples from 15 metastatic patients recently profiled with a targeted panel spanning 40 Kb across eight cancer genes at 1500X mean coverage.

TOP

A081 - A systems approach to unravel host response processes and host-parasite interactions in trypanosome infections

Siddharth Jayaraman, The Roslin Institute, United Kingdom
Sofie Demeyer, The Roslin Institute, United Kingdom
Heli Vaikkinen, University of Glasgow, United Kingdom
Anne Donachie, University of Glasgow, United Kingdom
Annette MacLeod, University of Glasgow, United Kingdom
Darren Creek, University of Melbourne, Australia
Christiane Hertz-Fowler , University of Liverpool, United Kingdom
Michael Barrett, University of Glasgow, United Kingdom
Liam Morrison, The Roslin Institute, United Kingdom
Tom Michoel, The Roslin Institute, United Kingdom

Short Abstract: The relative resistance or susceptibility to parasite infections depends on a complex interplay between host and parasite genotypes. We hypothesize that the dynamically changing levels of disease-perturbed genes, proteins and metabolites during the progress of infection, explain disease mechanisms and that those mechanisms can be learned from the integration of genome-wide data captured at multiple levels like genotype, gene expression and metabolome, across hosts and parasites and over multiple time points. We set up an experiment based on a host-parasite genotype infection matrix using severe and mildly virulent T. brucei strains with susceptible and resistant mice to give much greater information on the relative contribution of host or parasite genotype to disease phenotypes. We have generated data for RNA-seq expression level in the host and parasite during disease progression, as well as host plasma metabolome levels (LC-MS), across four parasite/host genotype combinations at 3, 6, 10 and 12 days post infection. Here we present the findings from the correlation of gene expression changes to host metabolome changes derived from the same sample at different disease stages, which will serve to develop a mathematical model to study the contribution of host and parasite genotype to pathogenesis and the dynamics of systemic host-cell response, in order to generate detailed hypotheses on the processes that determine disease outcome.

TOP

A082 - Illuminating HNSCC: A Platform for Identification of New Therapeutics

Gabrielle Choonoo, OHSU, United States

Short Abstract: We have leveraged the public omics data from the The Cancer Genome Atlas (TCGA) with over 500 head and neck squamous cell carcinoma (HNSCC) patients in conjunction with our own repository of clinically well-annotated OHSU HNSCC patient samples, which have both omic and functional characterization. For the TCGA data, we have identified the most significantly aberrant pathways in HNSCC. These HNSCC-related pathways that we have identified are being examined for known drug targets in order to guide panel development to evaluate potential drug repurposing to HNSCC. In addition to those pathways currently targeted by existing drugs ("light" pathways), we are also able to detect and quantify the number of "dark" pathways to guide future drug development. We hypothesize that molecular therapeutics for HNSCC can be expanded by a rational approach combining in silico evaluation of TCGA genomics data and functional analysis of HNSCC cell response to inhibitor panels.

TOP

A083 - Rescue of gene-expression changes in an induced mouse model of spinal muscular atrophy by an antisense oligonucleotide that promotes inclusion of SMN2 exon 7

Huo Li, Biogen, United States
John Staropoli, Biogen, United States
Seung Chun, ISIS Pharma, United States
Norm Allaire, Biogen, United States
Patrick Cullen, Biogen, United States
Alice Thai, Biogen, United States
Christina Fleet, Biogen, United States
Yimin Hua, ISIS Pharma, United States
Frank Bennett, ISIS Pharma, United States
Adrian Krainer, ISIS Pharma, United States
Xiao Yang, Biogen, United States
Eric Zheng, Biogen, United States
Doug Kerr, Biogen, United States
Alexander McCampbell, Biogen, United States
Frank Rigo, Biogen, United States
John Carulli, Biogen, United States

Short Abstract: Spinal muscular atrophy (SMA), the leading genetic cause of infant mortality, is a neuromuscular disease characterized by progressive loss of α-motor neurons in the anterior horn of the spinal cord. SMA is caused by disruption of the survival motor neuron 1 (SMN1) gene, which is

partly, but insufficiently, compensated for by the neighboring, nearly identical paralogous gene SMN2. Inclusion of exon 7 is critical for production of full-length SMN protein and occurs at a much lower frequency for SMN2 than for SMN1. Antisense oligonucleotide (ASO)-mediated blockade of a splicing silencer in intron 7 was previously shown to promote inclusion of SMN2 exon 7 in multiple mouse models of SMA and mediate phenotypic rescue. It also is the basis of a therapy (ISIS-SMNRx) currently under clinical investigation in children and infants with SMA. However, to date, the downstream molecular consequences of this ASO therapy have not been defined. Here we characterize the gene-expression changes that occur in an induced model of SMA and show substantial prevention or reversal of those changes in central nervous system tissue upon intracerebroventricular administration of an ASO that promotes inclusion of exon 7, with earlier administration promoting greater rescue of the expression profile. The enrichment of cell cycle signaling pathways among rescued transcripts may highlight an emerging role for SMN in DNA replication and repair. This study offers a robust reference set of preclinical pharmacodynamic gene expression effects against which other investigational therapies for SMA can be compared.

TOP

A084 - Evaluation of molecular subtypes and classifications in Breast Cancer

Yu-Jui Ho, Cold Spring Harbor Laboratory, United States
Molly Hammell, Cold Spring Harbor Laboratory, United States

Short Abstract: Outcomes for breast cancer patients vary depending on the cancer types, disease stages, and patients' age. Adequately characterizing breast caner into distinct groups according to their biological function can have a large influence on how physicians treat patients in clinical settings and leads to a direct impact on outcome of the patient.
Here we present an approach for characterizing large cohorts of breast tissue samples collected by The Cancer Genome Atlas (TCGA) through a semi-supervised classification method. Our rationale is to find common gene expression patterns that can be used to classify unknown samples, and compare with existing molecular subtypes. Starting from an expanded 'intrinsic' list consisting of around 2000 genes, we first assess groups that show statistical significance in hierarchical clustering. We further identify genes that are essential for classification and calculate centroids within each group. Orthogonally, we also apply the same analysis pipeline to another gene list that are extracted by decomposition of the expression data from the same samples. This gives us two lists of essential genes that can be used for predictions.
Groups classified by using the two extracted gene lists are compared with the predefined molecular subtypes classified by using PAM50 and shows a high concordance. Venn diagrams show an enrichment of PAM50 genes in the two extracted gene lists, which confirm the important biological functions of these core genes that define each subtype. Moreover, Kaplan-Meier plots are used to demonstrate difference in patient survival status between groups identified using our new essential gene list.

TOP

A085 - Perturbations of PIP3 signalling trigger a global remodelling of gene expression and reveal a transcriptional feedback loop

Vladimir Kiselev, Babraham Insititute, United Kingdom
Veronique Juvin, Babraham Institute, United Kingdom
Mouhannad Malek, Babraham Institute, United Kingdom
Nicholas Luscombe, London Research Institute, United Kingdom
Phillip Hawkins, Babraham Institute, United Kingdom
Nicolas Le Novere, Babraham Institute, United Kingdom
Len Stephens, Babraham Institute, United Kingdom

Short Abstract: PIP3, synthesized by PI3Ks, regulates complex cell responses, such as growth and migration. Signals that drive long-term reshaping of cell phenotypes are difficult to resolve because of complex feedback networks that operate over extended times. PIP3-dependent modulation of mRNA accumulation is clearly important in this process but is poorly understood. We have quantified the genome-wide mRNA-landscape of non-transformed, breast epithelium-derived MCF10a cells and its response to transient (EGF or PI3Ka-selective inhibitor) or chronic (isogenic cells expressing an oncomutant PI3Ka allele or lacking the pip3-phosphatase/tumour-suppressor, PTEN) perturbations of PIP3. These results show that whilst many mRNAs are changed by long-term genetic perturbation of PIP3 signalling ("butterfly effect"), a much smaller number change with a directional logic that aligns with different PIP3 perturbations. Analysis of transcription factor activity revealed the transcription factor binding motifs SRF and PRDM1 as important regulators of PIP3-sensitive mRNAs involved in cell movement. Some of the PRDM1 target genes code for proteins involved in the regulation of the PI3K signalling pathway suggesting a transcriptional feedback loop.

TOP

A086 - Prognostic Long Non-coding RNAs Involved in Prostate Cancer Progression and Treatment Resistance

Varune Rohan Ramnarine, Laboratory of Advanced Genome Analysis, Vancouver Prostate Centre, Canada
Alexander Wyatt, Laboratory of Advanced Genome Analysis, Vancouver Prostate Centre, Canada
Fan Mo, Laboratory of Advanced Genome Analysis, Vancouver Prostate Centre, Canada
Mohammed Alshalalfa, Research and Development, GenomeDx Biosciences, Canada
Elai Davicioni, Research and Development, GenomeDx Biosciences, Canada
Yuzhuo Wang, Laboratory of Advanced Genome Analysis, Vancouver Prostate Centre, Canada
Colin Collins, Laboratory of Advanced Genome Analysis, Vancouver Prostate Centre, Canada

Short Abstract: Neuroendocrine prostate cancer (NEPC) is a treatment-resistant lethal disease where most patients die within 1 year and treatment options are strictly palliative. New therapeutic strategies and a greater understanding of how resistance emerges are urgently required to improve patient outcome.

To comprehensively characterize the disease we have pioneered a high fidelity xenograft model of NEPC. Adenocarcinoma (AD) patient tumors are grafted into mice, exposed to clinically used treatment, and then develop aggressive NEPC. As the tumor progresses, using deep next-generation sequencing, we profile at various time points generating a time series of events that result in terminal NEPC.

To help elucidate novel mechanisms of resistance we apply robust sequence analysis algorithms focusing on underexplored regions of the genome that don't code for protein (98.78%). Within these non-coding RNAs (ncRNAs) are a subclass of long ncRNAs (lncRNAs) – most lacking known function and/or unannotated. A core feature of our sequencing pipeline is the quasi de novo transcriptome assembly, which identifies unannotated transcripts, novel spliced-isoforms, as well as known transcripts. We detect ~90K and ~40K ncRNAs including 13,730 and 20,463 lncRNAs, annotated and novel respectively.

Our results show unique patterns of lncRNA expression along our time series that we have classified into five transcript categories, dissecting the pathogenesis from AD to NEPC. Overlaying these patterns with patient data have identified clinically relevant candidates and prognostic associations to patient outcome. Furthermore our research is the first to provide evidence of lncRNAs involved in NEPC and most importantly insights into novel mechanisms driving the disease.

TOP

A087 - Exploiting large-scale drug-protein interaction information for computational drug repurposing

Jaques Reifman, Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, United States
Anders Wallqvist, Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, United States
Ruifeng Liu, Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, United States
Narender Singh, Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, United States
Gregory Tawa, Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, United States

Jaques Reifman, Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, United States

Short Abstract: In this talk, we will present a computational approach that utilizes genome-wide drug-protein interaction information in the public domain and Bayes' theorem to predict novel therapeutic uses for Food and Drug Administration (FDA)-approved drugs. The underlying hypotheses are that 1) the therapeutic effect of a drug is due to its interactions with human proteins and 2) many drugs interact with multiple human proteins, likely having therapeutic potentials for treating diseases other than those for which they have been approved. We will demonstrate that, by using drug-protein interaction profiles derived from mining >171 million chemical-protein interaction pairs reported in the literature, robust models for identifying drugs with potential to treat a disease can be developed even when only one of the drugs approved for treating the disease is used to train the models. Thus, the method is applicable to diseases for which there are currently very limited pharmaceutical options.

TOP

A088 - Evaluation of T cell epitope prediction tools -- How well do the MHC class II binding prediction programs predict T cell immunogenicity?

Chin-Hsien (Emily) Tai, National Cancer Institute, United States
Ronit Mazor, National Cancer Institute, United States
Byungkook Lee, National Cancer Institute, United States
Ira Pastan, National Cancer Institute, United States

Short Abstract: The ability to identify immunogenic determinants that activate T cells is important for the development of new vaccines, allergy therapy and protein therapeutics. In silico MHC class II binding prediction algorithms are often used for T cell epitope identification. To understand how well those programs predict immunogenicity, we compared the predicted and experimentally identified T cell epitopes on PE38, a fragment of an anti-cancer immunotoxin. We found that the IEDB-recommended method performed best among the seven programs tested. The predictions for individual donors did not correlate well with the experimental data and the AUCs of the ROC curves were lower than 0.70. Although the two strongest epitopes were predicted at multiple cutoffs, overall four out of nine epitopes were missed. In order to predict all frequent epitopes, the binding threshold would need to be lowered to the point where all peptides are considered to be frequent epitopes. The high rates of false positives and false negatives may impede the use of prediction programs. We conclude that MHC class II binding predictions are not sufficient to predict the T cell epitopes in PE38.

TOP

A089 - Investigating evolutionary models of genome structure in aggressive prostate cancer

Marek Cmero, The University of Melbourne, Australia
Natalie Kurganovs, Royal Melbourne Hospital, Australia
Jessica Chung, The Victorian Life Sciences Initiative, Australia
Jan Schröder, The Walter + Eliza Hall Institute, Australia
Kangbo Mo, The University of Melbourne, Australia
Clare Sloggett, The Victorian Life Sciences Initiative, Australia
Niall Corcoran, Royal Melbourne Hospital, Australia
Christopher Hovens, Royal Melbourne Hospital, Australia
Cheng Soon Ong, NICTA, Australia
Geoff Macintyre, Cancer Research UK, United Kingdom

Short Abstract: Tumour evolution is a complex and multifaceted process. Recently, many approaches have arisen for inferring the evolutionary dynamics of tumour cell populations from point-mutation and copy-number data. Studying the role of structural variations (SVs) in cancer evolution however, particularly balanced rearrangements, has been less thoroughly explored. We present a method of reconstructing cancer phylogeny from multiple single-patient samples using large scale genomic aberrations and apply it to prostate cancer, which is particularly rearrangement-driven. We demonstrate that tumour phylogenies are able to be reconstructed using rearrangement data alone, and we further expand our model to characterise subclonal SVs. We demonstrate our methods by applying them to longitudinal samples from patients undergoing second-line anti-hormone therapy to gain insight into the mechanisms of castration resistance.

TOP

A090 - Cancer cell line response to compound dose change

Avid Afzal, Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, United Kingdom
Martin Otava, Interuniversity Institutes for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, Belgium
Ziv Shkedy, Interuniversity Institutes for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, Belgium
Andreas Bender, Interuniversity Institutes for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, United Kingdom

Short Abstract: Dose response models are commonly used to evaluate drug's effects in biochemical and cell-based assays. Transcription-profiling experiments are usually performed at only one dose and effects can be identified by analysis of variance (ANOVA) models. These single dose transcription-profiles are not suitable to distinguish effects with different potencies; hence they limit the utility of expression data in comparison to other bioassays.

In experiments that transcription-profile is measured in a multiple dose, the analysis of dose-response transcriptomics has been shown to be the most powerful data analysis framework to detect transcription effects across doses. One such experiment that has been conducted in large scale is the Library of Integrated Network-Based Cellular Signatures (LINCS).

In this contribution we present the first large scale dose response analysis on this dataset where we have analysed a cancer cell line response to compound treatment at 3 different doses (being DMSO, 5 μM and 10 μM). The response of the prostate cancer cell line (PC3) to a set of 561 small-molecule compounds has been measured in terms of gene expression, after 24 hours.

The aim of this study is to identify genes that change in response to the change of the dose for each compound and to investigate common factors such as structure, protein target(s) or pathway motif similarity among compounds that exhibit similar gene expression patterns in response to dose concentration change.

TOP

A091 - Experimental and computational approaches to enable the discovery of diagnostic small-RNAs in FFPEs and liquid biopsies

Marie Fahey, Asuragen, United States
Stephanie Bridger, Asuragen, United States
Errin Lagow, Asuragen, United States
Diane Isley, Asuragen, United States
Dennis Wylie, University of Texas, United States
Brian Haynes, Asuragen, United States

Short Abstract: RNA-Seq has proven its potential to enable observations of biological variation in RNA processing at an unprecedented level of detail. Variation in microRNA biogenesis manifests through base trimming and additions at the 5' and 3' ends and single base substitutions yielding alternative mature microRNA transcripts termed isomiRs. IsomiR abundance has been shown to depend on genetic background, tissue and disease state and the variation in these sequences has biologically functional consequences.
Compared to microarrays and qPCR measurement platforms, RNA-Seq offers more comprehensive profiling to uncover novel small RNA biomarkers such as isomiRs to help advance the diagnosis and treatment of complex diseases. However, it remains a challenge to apply this technology to clinically relevant sample types such as FFPE, FNA and liquid biopsies that test the assay limits of RNA input and integrity. Moreover, it is an open question the extent to which observed variation in isomiRs is a result of artifacts introduced during NGS library prep and sequencing.
To disentangle the biological and technical contributions of miRNA sequence variation we evaluated the isomiR distributions of biological samples from 4 tissues against a pool of 950 synthetic microRNAs. Next, we applied RNA-Seq to discover diagnostic biomarkers from a set of 20 benign and

malignant FFPE thyroid biopsies and identified a population of isomiRs that were predictive of malignancy. Finally, we demonstrate the potential of RNA-Seq applied to liquid biopsies through the sequencing of low-input (250 µl) serum samples exhibiting technical reproducibility while exceeding the sensitivity of matched qPCR.

TOP

A092 - A Flexible Efficient Statistical Model for Joint Analysis of Differential Gene Expression in Multiple Studies

Yingying Wei, The Chinese University of Hong Kong, Hong Kong
Toyoaki Tenzen, Massachusetts General Hospital, United States

Short Abstract: The exponential accumulation of gene expression data in public data repositories provides an unprecedented opportunity to improve differential gene expression detection by borrowing information across the massive number of datasets. However, so far the standard methods for detecting differential gene expression are mostly designed for analyzing a single gene expression experiment.

When data from multiple related gene expression studies are available, separately analyzing each study is not ideal as it may fail to detect important genes with consistent but relatively weak differential signals in multiple studies. Jointly modeling all data allows one to borrow information across studies to improve the analysis. However, a simple concordance model, in which each gene is assumed to be differential in either all studies or none of the studies, is incapable of handling genes with study-specific differential expression. In contrast, a model that naively enumerates and analyzes all possible differential patterns across studies can deal with study-specificity and allow information pooling, but the complexity of its parameter space grows exponentially as the number of studies increases.

Here, we propose a correlation motif approach to address this dilemma. This approach searches for a small number of latent probability vectors called correlation motifs to capture the major correlation patterns among multiple studies. The motifs provide the basis for sharing information among studies and genes. The approach has flexibility to handle all possible study-specific differential patterns. It significantly improves detection of differential expression and overcomes the barrier of exponential model complexity.

TOP

A093 - Early Detection of Preeclampsia using Circulating small non coding RNA

Liron Yoffe, Tel Aviv University, Israel
Ofer Isakov, Tel Aviv University, Israel
Daphna Weissglas, ,
David Golan, ,
Kypros Nicolaides, ,
Moshe Hod, ,
Noam Shomron, Tel Aviv University,

Short Abstract: Preeclampsia is one of the most dangerous pregnancy complications, and the leading cause of maternal and perinatal mortality and morbidity; yet its cause remains unclear. Although clinical symptoms appear late, early detection of preeclampsia can be feasible at the first trimester. Recent findings suggest that circulating small non-coding RNAs (ncRNAs) in the mother's blood may be effective markers for early diagnosis of preeclampsia, however as of yet such ncRNAs were identified only in late stages of the pregnancy and have not been implemented in clinical practice. Furthermore, mapping ncRNA expression at an early stage of the disease might shed light on the possible mechanisms involved in the disease etiology.
We have compared small ncRNAs in plasma of first trimester pregnant women with and without preeclampsia. To this end, we have performed small ncRNAs Next Generation Sequencing (NGS) of preeclampsia and control samples, and identified several transcripts significantly differentially expressed between the two sets. We further utilized the list of these transcripts and created a pipeline for supervised classification of preeclampsia versus control samples. Our pipeline generates a generalizable logistic regression model using a 5-fold cross validation on numerous random partitions into training and test sets. Using this procedure our classification pipeline obtained high accuracy. Furthermore, applying the procedure on two different ethnic groups resulted in similar accuracy values, which demonstrates our method's generalization capability.
Our findings lay the foundation for an early non-invasive diagnostic tool of preeclampsia based on circulating small ncRNAs, in order to lower the life-threatening risk for the mother and fetus.

TOP

A094 - The IRIDA genomic epidemiology ontology: standards to improve infectious disease outbreak detection and investigation

Melanie Courtot, Simon Fraser University, Canada
Emma Griffiths, Simon Fraser University, Canada
Damion Dooley, BC Public Health Microbiology & Reference Laboratory and University of British Columbia, Canada
Josh Adam, National Microbiology Laboratory, Public Health Agency of Canada, Canada
Franklin Bristow, National Microbiology Laboratory, Public Health Agency of Canada, Canada
João André Carriço, Faculty of Medicine, University of Lisbon, Portugal
Bhavjinder Dhillon, Simon Fraser University, Canada
Matthew Laird, Simon Fraser University, Canada
Raymond Lo, Simon Fraser University, Canada
Thomas Matthews, National Microbiology Laboratory, Public Health Agency of Canada, Canada
Aaron Petkau, National Microbiology Laboratory, Public Health Agency of Canada, Canada
Geoff Winsor, Simon Fraser University, Canada
Lynn Schriml, University of Maryland School of Medicine, United States
Morag Graham, National Microbiology Laboratory, Public Health Agency of Canada, Canada
Gary Van Domselaar, National Microbiology Laboratory, Public Health Agency of Canada, Canada
Fiona Brinkman, Simon Fraser University, Canada
William Hsiao, BC Public Health Microbiology & Reference Laboratory and University of British Columbia, Canada

Short Abstract: Infectious disease outbreak investigations using microbial genomic data are currently hampered by delays incurred when manually integrating essential laboratory and epidemiological data from heterogeneous sources. The Integrated Rapid Infectious Disease Analysis (IRIDA) project, comprised of partners from national and provincial public health organizations and academic labs, is building a bioinformatics platform with a suite of tools needed to support time-sensitive infectious disease outbreak investigations. An ontology-based approach, combined with semantic web technology, will enable robust data integration and more efficient analysis within IRIDA.
We present the development of a new microbial genomic epidemiology ontology and associated standards, based on
(1) comprehensive review of existing relevant metadata standards to identify and cover missing elements,
(2) assessment of how these standards should support IRIDA workflows, and
(3) competency questions our project must address to support epidemiology studies.
By adhering to the best practices of the Open Biomedical and Biological Ontology (OBO) Consortium, our model allows consolidation of various existing ontological efforts into a resource directly compatible with IRIDA. Our modular development approach also ensures that it will be extendable, supporting more comprehensive coverage in the future, e.g., in the domain of food categories.
This research is a key component of the IRIDA platform to allow data integration and processing in a more automated fashion, alleviating the burden of manual analyses. Standardized reporting should also facilitate automated epidemiology and more efficient outbreak detection and mitigation, triggering action for example after auto-detecting deviations above expected biosurveillance baselines.

TOP

A095 - Genome-wide landscape of microsatellite instability in cancer

Ronald Hause, University of Washington, United States
Emily Turner, University of Washington, United States
Mallory Beightol, University of Washington, United States

Colin Pritchard, University of Washington, United States
Jay Shendure, University of Washington, United States
Stephen Salipante, University of Washington, United States

Short Abstract: Microsatellites, repeating 2-5 base pair sequences present throughout the human genome, can abnormally shorten or lengthen because of defects in the DNA mismatch repair (MMR) system, resulting in a “microsatellite instability” (MSI) phenotype. MSI is a key prognostic and diagnostic tumor phenotype that has been well studied by conventional methods. However, both the genomic landscape of MSI events and differences in MSI among cancer types remain poorly illuminated. We here present a comprehensive, genome-wide analysis of the landscape of MSI in cancer exomes. We catalogued MSI events at over 500,000 incidentally sequenced microsatellite loci across 4,478 cancer exomes spanning 18 different cancer types from The Cancer Genome Atlas. We constructed a global classifier for MSI that achieved 93.75% sensitivity and 98.5% specificity, compared to gold-standard MSI calls based on the revised Bethesda guidelines. We observed that MSI-low (MSI-L) samples did not display significant differences from MS-stable (MSS) samples in the number of MSI events and support discontinuation of the use of MSI-L as a distinct classification. Comparative examination of MSI between cancer types revealed both core loci and cancer-specific loci associated with MSI. Lastly, we investigated the mutational spectra of MMR genes in MSS and MSI-high (MSI-H) samples and correlated global MSI status with clinical covariates. Our results provide a comprehensive view of MSI in cancer exomes, highlighting both conserved and cancer-specific MSI properties and identifying candidate genes underlying predisposition to global MSI. Future work will attempt to functionally validate these candidates as causally influencing global MSI.

TOP

A096 - WES reveals mutations in NARS2 and PARS2 in patients with Alpers syndrome

Marcela Davila, University of Gothenburg, Sweden
Jorge Asin Cayuela, University of Gothenburg, Sweden

Short Abstract: Alpers syndrome is a progressive neurodegenerative disorder that presents in infancy or early childhood and is characterized by diffuse degeneration of cerebral gray matter. While mutations in POLG1, the gene encoding the gamma subunit of the mitochondrial DNA polymerase, have been associated with Alpers syndrome with liver failure (Alpers–Huttenlocher syndrome), the genetic cause of Alpers syndrome in most patients remains unidentified. With whole exome sequencing we have identified mutations in NARS2 and PARS2, the genes encoding the mitochondrial asparaginyl- and prolyl-tRNA synthetases, in two patients with Alpers syndrome. One of the patients was homozygous for a missense mutation (c.641C>T, p.P214L) in NARS2. The affected residue is predicted to be located in the stem of a loop that participates in dimer interaction. The other patient was compound heterozygous for a one base insertion (c.1130dupC, p.K378 fs*1) that creates a premature stop codon and a missense mutation (c.836C>T, p.S279L) located in a conserved motif of unknown function in PARS2. This report links for the first time mutations in these genes to human disease in general and to Alpers syndrome in particular.

TOP

A097 - HIV-1 Transmitted Drug Resistance: New Insights Into the Transmissibility of SDRMs

Raf Winand, Katholieke Universiteit Leuven / iMinds Medical IT, Belgium
Kristof Theys, Katholieke Universiteit Leuven, Belgium
Mónica Eusébio, Universidade Nova de Lisboa, Portugal
Jan Aerts, Katholieke Universiteit Leuven / iMinds Medical IT, Belgium
Ricardo Camacho, Katholieke Universiteit Leuven / Centro Hospitalar de Lisboa Ocidental, Belgium
Perpetua Gomes, Instituto Superior de Ciências da Saúde Sul / Centro Hospitalar de Lisboa Ocidental, Portugal
Anne-Mieke Vandamme, Katholieke Universiteit Leuven / Universidade Nova de Lisboa, Belgium
Ana Abecasis, Katholieke Universiteit Leuven / Universidade Nova de Lisboa, Portugal

Short Abstract: Surveillance drug resistance mutations (SDRMs) in drug naïve patients (DN) are typically used to measure HIV-1 Transmitted Drug Resistance (TDR). We tested here how SDRMs in patients failing treatment (TR), the original source of TDR, contribute to assessing TDR, transmissibility and transmission source of SDRMs.

The prevalence of SDRMs in DN and TR patients was retrospectively measured for 3554 HIV-1 subtype B infected patients. The transmission ratio (prevalence in DN/prevalence in TR) of each SDRM was calculated and analyzed by robust linear regression with outlier detection to interpret transmissibility.

Prevalence of SDRMs in DN and TR were linearly correlated, but some SDRMs were classified as outliers – above (protease: D30N, N88D/S, L90M, reverse transcriptase: G190A/S/E) or below (RT: M184I/V) expectations. The normalized regression slope was 0.073 for PI, 0.084 for NRTI, and 0.116 for NNRTI.

We present an innovative and simple approach to investigate the transmissibility of SDRMs by determining individual mutation transmission ratios and using linear regression to describe the relationship between their prevalence in TR and DN. The significant linear correlation between prevalence of SDRMs in DN and in TR indicates that the latter can be useful to predict levels of TDR. Higher transmission ratios and outliers above the regression line indicate more onwards transmission among DN and/or higher persistence of such SDRMs, while the opposite indicates lower transmission among DN and/or lower persistence. Our results emphasize the importance of monitoring SDRMs in TR in order to gain further insight into the dynamics of the transmission of SDRMs.

TOP

A098 - A New Efficient and Accurate Bioinformatics Tool for Vector Integration Site Analysis in Hematopoietic Stem Cell Gene Therapy Trials

Andrea Calabria, San Raffaele Telethon Institute for Gene Therapy, Italy
Giulio Spinozzi, San Raffaele Telethon Institute for Gene Therapy, Italy
Stefano Brasca, San Raffaele Telethon Institute for Gene Therapy, Italy
Fabrizio Benedicenti, San Raffaele Telethon Institute for Gene Therapy, Italy
Erika Tenderini, San Raffaele Telethon Institute for Gene Therapy, Italy
Alessandra Biffi, San Raffaele Telethon Institute for Gene Therapy, Italy
Eugenio Montini, San Raffaele Telethon Institute for Gene Therapy, Italy

Short Abstract: The molecular analysis of viral vector genomic integration sites (IS) is a key step in hematopoietic stem cell-based gene therapy (GT) applications, allowing to assess both the safety and the efficacy of the treatment and to study the basic aspects of hematopoiesis and stem cell biology. The increasing number of proviral/host genomic junctions obtained through recent next generation sequencing (NGS) platforms requires efficient and accurate bioinformatics pipelines for IS analysis. Several tools are available to study IS from raw NGS reads with good level of precision and recall but their common drawback is on computational requirements or low scalability. Here we present the first bioinformatics IS analysis pipeline build on BWA and latest filtering tools and on custom Python software designed to efficiently parallelize the computational load. We optimized the parameters of the pipeline exploiting simulated IS data based on empirical distributions acquired from our clinical and preclinical datasets, and we obtained sensitivity and specificity scores >0.95. Our tools have been successfully applied in our ongoing clinical trial for metachromatic leukodystrophy with a self-inactivating lentiviral vector in which, from 7 GT patients, we retrieved and sequenced >60 million proviral/host genomic junctions from samples of peripheral blood and bone marrow harvested at different points with a follow up of 36 months after treatment. The overall analysis required <24 hours computational time on a single workstation using 16 CPUs, thus allowing all laboratories working on GT-based studies to process NGS datasets of IS and retrieve reliable IS using standard computational resources.

TOP

A099 - Bioinformatics analysis pipeline for canine whole genome sequencing data identifies a novel mutation in canine dental dysplasia

Meharji Arumilli, University of Helsinki, Finland
Marjo Hytönen, University of Helsinki, Finland
Salmela Elina, University of Helsinki, Finland
Lukinmaa P, University of Helsinki, Finland

Sarkiala-Kessel E, University of Helsinki, Finland
Nieminen P, University of Helsinki, Finland
Kere J, University of Helsinki, Finland
Hannes Lohi, University of Helsinki, Finland

Short Abstract: INTRODUCTION: Dogs have emerged as clinically and genetically relevant large animal models for human inherited disorders. Majority of the over 600 described genetic disorders in dogs are similar to human conditions and are inherited in a Mendelian way. Unique breed structure facilitates gene discovery. We aimed to study the genetics of a novel canine disease with vigorous tooth attrition using a whole genome sequencing approach.
METHODS: We used clinical, pathological and pedigree analyses to characterize a novel canine dental disease and to establish a study cohort for gene discovery, including six affected and 16 unaffected Border Collies. The whole genome resequencing was performed for all the 22 dogs and a bioinformatics analysis pipeline was established to filter out candidate mutations for further experimental validation. For the bioinformatics pipeline we integrated tools for quality control (FASTX), alignment (BWA), variant calling (GATK, SAMTOOLS), annotation and pathogenicity prediction using snpEff adapted for dogs.
RESULTS: We describe a recessive condition with severe tooth attrition and hypomineralized dentin. Bioinformatics analysis of the data under recessive model identified two promising variants that were screened in the affected pedigree using Sanger sequencing. Only one of the variants segregated with the disease and it was confirmed in a cohort of ~400 dogs from the same and other breeds.
CONCLUSIONS: The establishment of a bioinformatics pipeline for whole genome sequencing data provides a powerful platform for gene discovery in canine disorders. The identification of the mutation for the dental disease will help to understand the molecular etiology of the disease in comparison to human disorder.

TOP

A100 - CoryneRegNet v7.0 – Updated backend for keeping up with a growing amount data.

Lucas Ferreira, University of Southern Denmark (SDU), Denmark
Richard Röttger, University of Southern Denmark (SDU), Denmark
Jan Baumbach, University of Southern Denmark (SDU), Denmark

Short Abstract: CoryneRegNet is an ontology-based data warehouse for corynebacterial transcriptional regulatory networks. Using the regulatory information available for the well-known bacteria C. glutamicum, combined with a large amount of microarray data and with literature-derived knowledge, the databases is not only a comprehensive source for gene regulations of C. glutamicum but also enables the computational transfer of known regulations of model organisms to non-model organisms. CoryneRegNet has grown drastically since its first version, especially due to the vast amount of data produced in the post-genomic next-generation sequencing techniques.
This growth of available data made a couple of technical insufficiencies of CoryneRegNet more apparent over time; especially queries to the ontology based database are increasingly becoming a problem and slowing down the responsiveness of the system. CoryneRegNet 7.0 facilitates a complete overhaul of the database structure and the backend resources resulting in a tremendous speed-up of queries and network transfers. These improvements will pave the ground for including more organisms and with that giving a better and more accurate view of bacterial gene regulation in the future.

TOP

A101 - MicroRNA expression profiling of lung adenocarcinoma in never-smoker females

Namhee Yu, Ewha Womans University, Korea, Rep
Sanghyuk Lee, Ewha Womans University, Korea, Rep

Short Abstract: Many microRNAs have been reported to play critical roles in tumor development, progression and treatment responses, thus representing a promising class of cancer biomarkers. But our understanding on their functional roles in lung adenocarcinoma is still limited in non-smoker female population.
Here we report a deep sequencing study of mRNA and microRNA profiling for paired tumor and normal tissues from 25 never-smoker female patients of non-small-cell lung adenocarcinoma (NSCLC). We have determined 34 important microRNAs that showed reliable differential expression (FDR<0.001 in edgeR) with consistent direction (over 80% of patients in the same direction) and valid expression level (average logCPM>4 in edgeR). We further obtained target mRNAs whose expression were inversely correlated with 34 microRNAs, and the subsequent pathway enrichment analysis identified cell mobility, angiogenesis, and cell signaling as important functions.
To interrogate microRNAs associated with smoking, we have identified 34 smoker cases with microRNA sequencing data for paired tumor and normal tissues in the TCGA lung adenocarcinoma cohort. Comparing two data sets showed that most differentially expressed microRNAs (DEmiRs) were common. Notably, we found that let-7 family microRNAs were differentially expressed only in the smoker data set. Further pathway enrichment analysis on target genes of DEmiRs revealed many regulatory processes that were distinct between smoker and non-smoker populations.
In conclusion, our study has identified not only the smoker-specific DEmiRs but also differential regulatory processes in non-smoker population.

TOP

A102 - An Algorithm for the Efficient Inference of Cancer Progression Models

Giancarlo Caravagna, University Milano-Bicocca, Italy
Daniele Ramazzotti, University Milano-Bicocca, Italy
Giulio Caravagna, University Milano-Bicocca, Italy
Loes Olde Loohuis, UCLA, United States
Alex Graudenzi, University Milano-Bicocca, Italy
Ilya Korsunsky, NYU, United States
Marco Antoniotti, University Milano-Bicocca, Italy
Bud Mishra, NYU, United States

Short Abstract: Cancer is a disease of evolution whose process is characterized by accumulation of somatic alterations to the genome, which selectively make a cancer cell fitter to survive [1]. The understanding of progression models for cancer, i.e., the identification of sequences of mutations that leads to the emergence of the disease, is still unclear.
The problem of reconstructing such progression models is not new; several methods to extract progression models from cross-sectional samples have been developed such as [2, 3, 4].
In this work, we propose a novel algorithm called CAPRI (CAncer PRogression Inference) to reconstruct DAGs, modeling the sequences of mutations, which characterize cancer evolution.
To the best of our knowledge, the existing techniques are based either on correlation or on maximum likelihood. Differently, we perform the reconstruct by exploiting the notions of probabilistic causation in the spirit of Suppes' causality theory [5]. We note that in the context of biological systems and cancer progression, the notion of causality can be interpreted as the notion of selective advantage of the occurrence of a mutation.
In those settings, we prove the correctness of our algorithm and, on synthetic data, we show that our approach outperforms the state-of-the-art. Moreover, for real cancer datasets, we highlight biologically significant differences in the progressions inferred with respect to other competing techniques.

[1] Hanahan D., Weinberg R.A. (2011). Cell 144
[2] Vogelstein B. et al. (1988) New England Journal of Medicine 319
[3] Desper R. et al. (1999). Journal Computational Biology 6
[4] Beerenwinkel N., et al. (2007). Bernoulli
[5] Suppes P. (1970). North Holland

TOP

A103 - Characterizing novel splicing sites from RNA-seq data in large sample sets

Sergei Häyrynen, University of Tampere, Finland

Matti Nykter, University of Tampere, Finland

Short Abstract: Cancer-related splicing alterations have been reported to contribute to cancer progression and prognosis. Large publicly available datasets can be used to uncover less frequent but functionally important alterations. However, standard approaches that include transcriptome-wide quantification or transcriptome assembly are computationally expensive or confined to known events. Our approach focuses on finding only novel events with possible biological implications.

Developed pipeline catalogues potential splice sites from RNA-seq data by focusing on gapped reads. Sequential low cost filtering and scoring steps result in a limited subset of candidate events for further, computationally more costly analysis. Previously annotated and low quality splice cites are filtered out. Proximal genome sequences of remaining sites are scanned for splicing breakpoint motifs and annotated in relation to genes to distinguish novel splicing features. Boundaries of novel exons are estimated from pooled coverage. Events are quantified in relation to known splicing patterns. Performance of feature and boundary detection is validated by randomly dropping known features from annotation and assessing the quality of their detection.

Remaining subset of features is examined in combination with known isoforms to extract events with possible biological implications by inspecting changes in reading frames and protein domains. If available, mutation call data is integrated to report mutations associated with splicing breakpoint and branch point motifs. The proposed analysis allows us to catalogue novel splicing events across large datasets such as TCGA , subsequently allowing us to associate the identified events to clinical phenotypes.

TOP

A104 - Matrix factorization with features for drug activity modeling

Adam Arany, KU Leuven, Belgium
Jaak Simm, KU Leuven, Belgium
Yves Moreau, KU Leuven, Belgium

Short Abstract: The in-silico modeling of interactions between small molecular entities and their targets is an important tool in pharmaceutical research to solve problems like drug target identification, prediction of side-effects and deconvolution of phenotypic effects.

Previous studies have modeled each protein target separately, but that approach does not take into account the common patterns in the data. The problem can also be formulated as a factorization of an incompletely filled matrix where the goal is to predict the unknown values. The matrix factorization methods have been very successful in large scale recommender system tasks, like the Netflix challenge where the data set consisted of 200 million movie ratings from half a million users [1]. We generalized the matrix factorization framework to simultaneously factorize multiple relations (like protein and functional assay data) and also incorporate additional entity-level features like chemical fingerprints.

Our work can be considered as treating each protein as a task in a multi-task learning setting, and formulating a large learning problem of predicting compound activity based on their chemical structure [2].

We applied the method on IC50 values from ChEMBL database [3]. The size of the feature space of compounds is 300k, the number of compounds is 75k and the number of protein targets is about 300.

TOP

A105 - A Computational Framework For The Prioritization Of Disease-Gene Candidates

Fiona Browne, Ulster University, United Kingdom
Haiying Wang, Ulster University, United Kingdom
Huiru Zheng, Ulster University, United Kingdom

Short Abstract: The ongoing development of large-scale technologies such as high throughput sequencing technologies has resulted in an explosion of ‘omic’ data. These technologies and data have been pivotal in the identification of disease-gene candidates from patient cohorts. Notwithstanding these significant improvements, the challenge of identifying meaningful disease-associated genes
from a long list of candidate genes still remains.
To address this need, we have developed a novel computational framework to integrate heterogeneous data and prioritize disease-gene candidates for further experimental investigation. Our framework integrates diverse data including: expression, protein-protein interaction network, ontology-based similarity and network topology measures. Furthermore, we incorporate tissue-specific expression data to evaluate prioritized candidates as pathology caused by genetic defaults is usually highly tissue-specific.
Our computational approach was applied to prioritize Alzheimer Disease (AD) genes whereby Human AD gene expression data were obtained from the Gene Expression Omnibus. Using our approach, a list of 31 prioritized genes was generated whereby key AD susceptible genes: INPP5D and PSEN1 were identified. Biological process enrichment analysis revealed the prioritized genes are modulated in AD pathogenesis including: regulation of neurogenesis and generation of neurons. KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis identified significant hub involvement in the Neurotrophin signaling and Huntington Disease pathways. Furthermore, our evaluation demonstrated a relatively high predictive performance (AUC: 0.73) when classifying AD and normal gene expression profiles from individuals. This work provides a foundation for future investigation of diverse heterogeneous data integration for disease-gene prioritization.

TOP

A106 - Latent Tree Models for Cancer Phylogeny

Emily Hindalong, BC Cancer Agency, Canada
Hossein Farahani, BC Cancer Agency, Canada
Sohrab Shah, BC Cancer Agency, Canada

Short Abstract: Studying how cancer genomes evolve has the potential to further our understanding of the events and conditions that underlie oncogenesis. We have developed a method to learn evolutionary histories, or phylogenies, from point-mutation profiles of deeply sequenced single-cell cancer data. Traditional algorithms for learning phylogenetic trees put all the extant species on the leaves of the tree. In cancer single cell data, some of the sampled cells are identical copies of already divided cells. Because traditional phylogenetic algorithms cannot capture this phenomenon, we developed a new method based on general latent tree models, allowing us to learn phylogenies for distinct genotypes instead of individual cancer cells.

TOP

A107 - Community Detection in Disease Gene Networks Highlights Comorbidity between Metabolic Diseases and Alzheimer´s Disease

Thanh-Phuong Nguyen, Systems Biology Group, Life Sciences Research Unit, University of Luxembourg., Luxembourg
Thomas Sauter, Systems Biology Group, Life Sciences Research Unit, University of Luxembourg., Luxembourg

Short Abstract: Over the past decades, the molecular background of the phenotypic variability in metabolic diseases has been investigated and a spectrum of relations between clinical syndromes and molecular features has been identified. Although some genes have emerged as important players in the pathogenesis, the precise molecular machinery involved in metabolic diseases remains largely unknown. We present a systems medicine approach for exploring the complex phenotypic interdependencies and comorbidities of metabolic diseases and thus inferring underlying pathogenic mechanisms. Network mining is employed to analyze interaction networks of disease-related genes. We manually curate a wide range of 812 metabolic diseases, 64 malnutrition disorders, 44 over-nutrition disorders, and 137 liver diseases from the MeSH database. A total of 1,920 disease genes are investigated to construct the interaction network. We obtain a large network of 5,064 genes and 13,226 interactions from the HPRD database. The top 157 crucial genes are discovered by computing network centralities. We carry out the Glay clustering algorithm to identify four highly-connected communities from the subnetwork of 157 genes and their 998 interactions. The pathway-enrichment analysis shows that the communities are functionally related to cancer pathway, signaling pathway and especially neurodegenerative diseases-related pathway. Interestingly Alzheimer´s disease genes are found to be highly-enriched including CASP3, PSEN1, GSK3B, MAPT, CYCS, SNCA, CASP8, and CALM1. This result

supports the hypothesis of the comorbidity between metabolic diseases and chronic neurodegenerative diseases, especially Alzheimer´s disease. The findings can be further used for studying the interacting biological modules and pathways linking metabolic diseases and Alzheimer´s disease.

TOP

A108 - Predicting drug-drug interactions by using pharmacological similarity

Kyunghyun Park, Korea Advanced Institute of Science and Technology, Korea, Rep
Doheon Lee, Korea Advanced Institute of Science and Technology, Korea, Rep

Short Abstract: As patients taking multiple medications have serious adverse drug reactions by drug-drug interactions (DDIs), it is crucial for predicting them in drug development. To analyze and predict DDIs on a large scale, computational approaches have been developed. Most of previous similarity-based approaches assumed that similar drugs have similar DDIs. However, these approaches are hard to understand the mechanism of the predicted DDIs by the assumption.
In this study, we present a new method that can predict DDIs with possible mechanisms based on pharmacological similarity. The underlying assumption is that if drugs similar to drug A interact with a group of drugs with identical pharmacological properties significantly, then drug A is likely to interact with the drug group through the pharmacological properties.
Our method achieved high accuracy and also provided possible mechanisms of DDIs. Therefore, we expect that our method could be used to predict and prevent DDIs in drug development.

TOP

A109 - Exome-sequencing phenotype interpretation using VarElect, the NGS phenotyper

Gil Stelzer, Weizmann Institute of Science, Israel

Short Abstract: Next generation sequencing is becoming an effective instrument for identifying genes that cause Mendelian disorders. However, its routine clinical application is only beginning to emerge. Upon standard filtration methods that rely on variant population frequency, deleterious effect on the protein as well as evolutionary conservation a variant short-list often containing hundreds of candidates is generated. To overcome the hurdle of connecting one variant to the patient's phenotype, we constructed VarElect, a new Variant Election software tool that attains phenotype-dependent variant prioritization, leveraging the comprehensive information within GeneCards and MalaCards. Users submit phenotype/disease related keywords and a gene list. VarElect then produces a prioritized list of contextually annotated genes, according to publication information, disease association, gene function and various other data. In addition to connecting between genes and phenotypes, GeneCards associates between genes through shared pathways, protein interactions, paralogs and publications. By these means an indirect link may be produced between a gene candidate, resulting from NGS experiments but not directly connected to the phenotype of interest, and an implicating gene that is directly connected to the phenotype. In this manner genes in the short-list are scored according to their likelihood to be phenotype related, thus enabling to perform the last decision step in NGS analysis in a fast and objective manner.

TOP

A110 - Optimizing cancer genome sequencing and analysis

Malachi Griffith, Washington University, United States
Christopher Miller, Washington University, United States
Obi Griffith, Washington University, United States
Kilannin Krysiak, Washington University, United States
Zachary Skidmore, Washington University, United States
Avinash Ramu, Washington University, United States
Jason Walker, Washington University, United States
Ha Dang, Washington University, United States
Lee Trani, Washington University, United States
David Larson, Washington University,
Ryan Demeter, Washington University, United States
Michael Wendl, Washington University, United States
Rachel Austin, Washington University, United States
Vincent Magrini, Washington University, United States
Sean McGrath, Washington University, United States
Amy Ly, Washington University,
Li Ding, Washington University, United States
Tim Ley, Washington University, United States
Elaine Mardis, Washington University, United States
Richard Wilson, Washington University, United States

Short Abstract: Tumors are typically sequenced to depths of 75-100x (exome) or 30-50x (whole genome). We demonstrate that current sequencing paradigms based on this coverage are inadequate for tumors that are impure, aneuploid, and/or clonally heterogeneous. To reassess optimal sequencing strategies, we performed ultra-deep (up to ~312x) whole genome sequencing and exome capture (up to ~433x) of a primary acute myeloid leukemia, its subsequent relapse, and a matched normal skin sample. We tested 7 alignment algorithms and 7 single-nucleotide variant callers, and validated ~200,000 putative SNVs by sequencing them to mean depths of ~1,000x. Additional targeted sequencing provided over 10,000x coverage and ddPCR assays provided up to ~250,000x sampling of selected sites (of up to 2 ug of input DNA per assay). Using these data, we evaluated the effects of different library generation approaches, depth of sequencing, and analysis strategies on the ability to effectively characterize a complex tumor. This dataset, representing the most comprehensively sequenced tumor described to date, will serve as an invaluable community resource.

TOP

A111 - Identification of cancer-relevant alterations in GIST with exome and rna-seq data

Alejandra Cervera, University Of Helsinki, Finland
Harri Sihto, Translational Cancer Biology Program, University of Helsinki, Finland
Heikki Joensuu, Translational Cancer Biology Program, University of Helsinki and Department of Oncology, Helsinki University Central Hospital, Finland
Maarit Sarlomo-Rikala, Department of Pathology, Helsinki University Central Hospital, Finland
Pietr Rutkowski, Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Poland
Sampsa Hautaniemi, University of Helsinki, Finland
Rainer Lehtonen, University of Helsinki, Finland

Short Abstract: Gastrointestinal stromal tumors (GIST) that carry mutations in KIT or PDGFRA genes (80-95%) respond well to adjuvant therapy with imatinib or sunitinib, but advanced cases or tumors with specific genotypes usually recur after surgery. Here, we present
the analysis of 12 exome (11 with matching blood samples that were used to rule out germline variants) and 5 matching transcriptome deep sequencing tumor samples with the aim of finding new molecular markers in GIST. Our workflow integrates several filtering steps that significantly reduce the number of false positive variant and indel calls both in exome and rna-seq. Annovar is used for functional annotation of the variants and indels, which are then scored using CADD and MutSig, allowing us to rank them based on their deleteriousness. By combining exome with expression information we have been able to identify possibly silencing mutations, such as ones found in EPHA7, FBLN1 and GRB14 genes, were patients carrying the mutations show diminished expression in those genes. Other preliminary results show PRIM2, CSMD1 (tumor suppressor), MAP2K3, and MUC4 (associated to several cancers) to be heavily mutated in most of our samples. Furthermore, in the rna-seq patient samples we found a fusion gene, not been previously reported in GIST, that is also present in rna-seq data from one of two GIST cell lines tested.

TOP

A112 - An Phenotype specific blustering method in cancer data

KIM JUNGRIM, Yonsei University, Korea, Rep
PARK SANGHYUN, Yonsei University, Korea, Rep

Short Abstract: Gene clustering is a method for finding gene sets which are related to the same biological processes or molecular function. In order to find these gene sets, previous studies have clustered genes which showed similar mRNA expression or a specific expression pattern in a (sub) sample set. However, for two contrasting groups of samples, it is not easy to identify gene sets which show significant expression pattern in only one group using current gene clustering methods. Existing biclustering methods use only one group (disease) of samples. It is hard to identify disease specific biclusters which are differentially expressed in the disease although those methods can find biclusters which have specific expression pattern. Here, we proposed a novel method using a genetic algorithm in gene expression data, in order to find gene sets which can represent specific subtype of cancer. Proposed method finds gene sets which have statistically differential mRNA expression on two contrasting samples and fraction of cancer samples. The resulting gene modules share higher number of GO (Gene Ontology) terms related to a specific disease than gene modules identified by current algorithms. We also identify that when we integrate protein-protein interaction data with gene expression data of colorectal cancer samples, proposed method can find more functionally related gene sets.

TOP

A113 - Driver mutation profiles in UKLS CT-screen detected tumours

Russell Hyde, University of Liverpool, United Kingdom
Michael Davies, University of Liverpool, United Kingdom
Michael Marcus, University of Liverpool, United Kingdom
Ying Chen, University of Keele, United Kingdom
John Gosney, Royal Liverpool and Broadgreen University Hospital Trust, United Kingdom
John Field, University of Liverpool, United Kingdom

Short Abstract: Diagnosis of lung cancer patients at a late stage of tumour development is a major constraint on disease outcome. Computed tomography (CT)-screening of high risk individuals may address this limitation by identifying lung tumours at a presymptomatic stage when surgical or therapeutic intervention may be more effective. Exomic mutational profiles of 9 stage-IA lung tumours (3 squamous cell carcinomas and 6 adenocarcinomas; each patient being a current or former smoker) that were detected in the CT-screened arm of the UK Lung Screening (UKLS) trial were characterised. By restricting to targeted, non-repetitive, highly mappable coding regions of the genome, that had high coverage across all samples and low background error rate, we were able to compare DNA from formalin-fixed paraffin embedded (FFPE; sequenced to ~ 100x coverage) lung tumours with matched blood-derived DNA (sequenced to ~ 50x). Within these high confidence genomic regions, between 69 and 465 somatic mutations (SNVs and indels) were identified per sample. Though somewhat low, the mutation frequencies were not significantly different from that observed in comparable regions of TCGA lung tumour exomes (the current / former smokers therein), although mutation frequency within a list of cancer driver genes was significantly lower in the CT-detected tumours than in the TCGA dataset. The results show that FFPE-derived lung tumour material is an effective source for the genomic characterisation of primary lung tumours detected by CT-screening and subtle differences can be found in the location of mutations, but not the overall level when compared to TCGA data.

TOP

A114 - RNA-Seq analysis for clinical testing in an individualized oncology setting

Eric Klee, Mayo Clinic, United States
Gavin Oliver, Mayo clinic, United States
Diane Grill, Mayo Clinic, United States
Naresh Prodduturi, Mayo Clinic, United States
Jie na, Mayo Clinic, United States
Jeanette Eckel-Passow, Mayo Clinic, United States

Short Abstract: DNA sequencing is being implemented for the analysis of patients' genomic tumor profiles, to identify genomic aberrations indicative of a tumor's vulnerability to targeted treatments. DNA sequencing provides a molecular characterization of the tumor that appears to be improving theragnostic decision making in some patients. In an effort to improve the utility of molecular guided therapeutic decision-making, we have piloted the integration of RNA-Seq analysis in a clinical oncology setting, within Mayo's Individualized Medicine Clinic. We demonstrate that RNA-Seq is a versatile tool with significant potential to improve clinical decision-making through the verification of the effects of DNA mutations at the level of the transcriptome or as a standalone molecular test. Sample profiling includes tumor-only analysis to identify expressed fusions and profile allelic-bias in called DNA variants. We also examined the utility of differential expression profiling and discuss the analytical challenge of N=1 scenarios, where the question of a suitable reference for differential expression profiling exists. We explored the use of a normal reference cohort derived from publicly available expression data; including a matched tissue-type reference set, as well as, a disparate-tissue reference set, for the purpose of differential expression profiling. Results are presented in the context of therapeutic decision-making, including amplification/deletion characterization and identification of activated networks through transcription factor profiling. Our findings suggest in some situations, DNA analysis considered in isolation could mislead clinical decision-making and potentially detriment patient care. However, RNAseq in N=1 context is complicated by the lack of a well defined reference.

TOP

A115 - A systems view on mining common pathway deregulation profiles in autoimmunity, autoinflammation and inflammation

Lilit Nersisyan, Institute of Molecular Biology NAS RA, Armenia
Arsen Arakelyan, Institute of Molecular Biology NAS RA, Armenia
Anna Hakobyan, Institute of Molecular Biology NAS RA, Armenia

Short Abstract: Despite distinct phenotypes, systemic autoimmune diseases share genetic associations, treatment response and clinical manifestations with other characterized by involvement of chronic inflammation. In this study, we aimed at global assessment of similarity and specificity of downstream molecular events through evaluation of pathway activity changes in systemic and organ-specific autoimmune, autoinflammatory and inflammatory disorders.
We have performed a meta-analysis of 1692 samples from 65 disease-tissue combinations, including healthy controls obtained from the Gene Expression Omnibus repository.
The amount of pathway activity deregulation was calculated based on the gene expression and KEGG pathway topology using Pathway Signal Flow algorithm. Similarities in patterns of pathway activity changes among the studied conditions we analyzed pathway signal flows using a self-organizing map algorithm using the R package "oposSOM". Finally, a graph describing degree of similarity in pathway deregulation profiles was constructed and graph community search was performed to identify groups of similar diseases.
The results show that, four disease groups can be identified, from which the first group gathers various tissues from healthy conditions. The other groups were formed by diseases showing highly divergent clinical manifestations, but having similar pathway deregulation profiles. Furthermore, our data indicate that autoimmune diseases, such as systemic lupus erythematosus and rheumatoid arthritis had a relatively high similarity in pathway deregulation patterns in different tissues. Overall, we offer a methodology, which allowed analyzing pathway activity alterations in systemic autoimmune and other chronic inflammatory disorders, which may contribute to a better understanding of common and specific molecular mechanisms guiding their progression.

TOP

A116 - Diagnostic role of exome sequencing in immune deficiency disorders

Steven Brenner, University of California, Berkeley, United States

Short Abstract: To interpret genomic variant data, we have developed an analysis protocol whose distinctive features enabled solving numerous clinical cases. The first steps are mapping and variant calling. To yield high quality sets of variants, we use multiple callers and employ multisample calling. The pipeline integrates variant annotation, variant filtering, and gene prioritization to reduce the millions of called variants to a manageable shortlist of possible causative variants. We have 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) at birth. We discuss

cases with immune deficiencies including severe combined immunodeficiency syndrome (SCID), Nijmegen breakage syndrome (NBS), and ataxia telangiectasia (AT). 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.

TOP

A117 - The Evolution of Human Cell Types

Julian Gough, University Of Bristol, United Kingdom
Adam Sardar, University Of Bristol, United Kingdom
Matt Oates, University Of Bristol, United Kingdom
Hai Fang, University Of Bristol, United Kingdom
Hideya Kawaji, RIKEN, Japan
Owen Rackham, University Of Bristol, Singapore

Short Abstract: Humans are composed of hundreds of cell types. As the genomic DNA of each somatic cell is identical, cell type is determined by what is expressed and when. Until recently, little has been reported about the determinants of human cell identity, particularly from the joint perspective of gene evolution and expression. Here, we chart the evolutionary past of all documented human cell types via the collective histories of proteins, the principal product of gene expression. FANTOM5 data provide cell-type–specific digital expression of human protein-coding genes and the SUPERFAMILY resource is used to provide protein domain annotation. The evolutionary epoch in which each protein was created is inferred by comparison with domain annotation of all other completely sequenced genomes. Studying the distribution across epochs of genes expressed in each cell type reveals insights into human cellular evolution in terms of protein innovation. For each cell type, its history of protein innovation is charted based on the genes it expresses. Combining the histories of all cell types enables us to create a timeline of cell evolution. This timeline identifies the possibility that our common ancestor Coelomata (cavity-forming animals) provided the innovation required for the innate immune system, whereas cells which now form the brain of human have followed a trajectory of continually accumulating novel proteins since Opisthokonta (boundary of animals and fungi). We conclude that exaptation of existing domain architectures into new contexts is the dominant source of cell-type–specific domain architectures.

TOP

A118 - Whole Cancer Genome Annotation in the Catalogue of Somatic Mutations In Cancer

Chai Yin Kok, , United Kingdom
SA Forbes, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
D Beare, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
N Bindal, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
P Gunasekaran, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
K Leung, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
M Jia, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
H Boutselakis, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
R Shepherd, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
D Minjie, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
S Bamford, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
S Ward, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
C Cole, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
T De, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
JW Teague, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
MR Stratton, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom
PJ Campbell, Cancer Genome Project, Wellcome Trust Sanger Institute, UK, United Kingdom

Short Abstract: COSMIC, the Catalogue Of Somatic Mutations In Cancer (http://cancer.sanger.ac.uk) is designed to store and display somatic mutation information relating to human cancers. The information is curated from the primary literature, international consortium data portals and the laboratories at the Cancer Genome Project, Sanger Institute, UK. In the v72 release, whole-genome annotations were curated across 1,103,964 tumour samples, detailing 3,158,657 mutations in 28,305 genes, together with 61,232 genomic rearrangements. This curation currently comprises 19,672 whole genome samples from papers and 60 cancer studies from the two main cancer portals, the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). All these data are mapped to the GRCh38 reference sequence, although an archive site is available to view GRCh37 coordinates.

With increasing quantities of whole-genome cancer data being curated into COSMIC, users can examine mutation distribution and characteristics across a range of tumour types. Scientists can use COSMIC to explore genome wide cancer mutation data, and identify new candidate cancer genes and mutational patterns beyond the 572 already identified in the Cancer Gene Census (http://cancer.sanger.ac.uk/census). In addition, the COSMIC Genome Browser (http://cancer.sanger.ac.uk/cosmic/browse/genome) can be used to view COSMIC data in genomic context and is especially useful for exploring gene expression, methylation, copy number and non-coding variant data.

We are currently mapping tumour classifications to the NCI thesaurus and cancer genes to Gene Ontology and we will provide cross referencing to NCI_CODE and GO ID in the near future.

COSMIC data is freely available to download by academic and non-profit organisations (https://cancer.sanger.ac.uk/cosmic/login).

TOP

A119 - Prediction of relapse in early stage lung adenocarcinoma using calibrated loss minimization and proximal methods

Michel Barlaud, I3S - UMR7271 - UNS CNRS, France
Marie Deprez, I3S - UMR7271 - UNS CNRS, France
Agnes Paquet, IPMC, UMR 7275 CNRS / UNS, France
Nicolas Glaichenhaus, IPMC, CNRS, INSERM, UNS UMR7275, UMR_S 1080, France
Lionel Fillatre, I3S - UMR7271 - UNS CNRS, France
Bernard Mari, IPMC, UMR 7275 CNRS / UNS, France

Short Abstract: About 35% of patients diagnosed with early stage lung adenocarcinoma (ES-LUAD) face relapse after surgery. Although several prognostic biomarkers have been proposed, none have been validated as accurate predictors of relapse in clinical trials. Proximal methods are more efficient and more accurate for feature selection and classification than state of the art classification methods, such as support vector machine. The aim of our study was to apply proximal methods to improve existing molecular classification and identify better predictors of relapse in ES-LUAD.

We used a subset of TCGA LUAD dataset including matched somatic mutations for 25 oncogenes, RNAseq and relapse data for 203 patients. We applied gene length and variance stabilization corrections to RNAseq count data, and we computed the occurrence of somatic mutations as a binary value. Our method is based on minimization of calibrated loss with sparsity induction by proximal splitting to predict relapse and identify biomarkers based on both transcriptome and mutation data.

We obtained a maximum accuracy of 82% in relapse prediction using a signature composed of the expression of 100 genes and the mutational status of all 25 tested oncogenes. Interestingly, the presence of mutations in MAP2K1, CTNNB1, PTEN and EGFR was retained as predictor, but with low weights. The biological relevance of the higher weights gene expression predictors will be discussed.

In conclusion, the use of calibrated loss minimization and proximal methods integrating tumor gene expression and somatic mutation status may provide a new reliable assay to predict relapse in ES-LUAD patients.

TOP

A120 - Mutant U2AF1 alters splicing in hematopoietic cells in vitro and in vivo

Brian White, Washington University, United States
Cara Shirai, Washington University, United States
James Ley, Washington University, United States
Sanghyun Kim, Washington University, United States
Matthew Ndonwi, Washington University, United States
Brian Wadugu, Washington University, United States
Theresa Okeyo-Owuor, Washington University, United States
Timothy Graubert, Massachusetts General Hospital/Harvard Medical School, United States
Matthew Walter, Washington University, United States

Short Abstract: Myelodysplastic syndromes (MDS) are the most common adult myeloid malignancy. As many as 50% of MDS patients have heterozygous somatic mutations in spliceosome genes. Specific mutations in the U2AF1 splicing factor, which participates in recognition of 3' splice site, occur in ~11% of MDS patients. U2AF1 mutations are mutually exclusive with mutations in other spliceosome genes (principally, SF3B1 and SRSF2) and most often occur in the founding clone. These findings strongly suggest that spliceosome mutations contribute to disease, though their mechanism of doing so remains unclear. To investigate the possibility that perturbations in pre-mRNA splicing contribute to disease pathogenesis, we performed a meta-analysis of one novel murine and two published human transcriptional (RNA-seq) profiles that characterize isoforms expressed by mutant or wildtype U2AF1. The results highlight common U2AF1-induced splicing alterations, which are enriched in RNA processing genes, ribosomal genes, and genes recurrently mutated in MDS and acute myeloid leukemia. Further, the murine model of U2AF1 mutation exhibits hematopoietic phenotypes of MDS, including a reduction in white blood cell counts and progenitor cell population expansion. These findings support the hypothesis that mutant U2AF1 alters downstream gene isoform expression, thereby contributing to abnormal hematopoiesis in MDS patients.

TOP

A121 - A Signalling Regulatory Loop Harbored in PIK3CA Driver-mutated Breast Tumors Prolong Patient Survival

Shauna McGee, McGill University, Canada
Naif Zaman, McGill University, Canada
Chabane Tibiche, National Research Council of Canada, Canada
Mark Trifiro, McGill University, Canada
Edwin Wang, National Research Council of Canada, Canada

Short Abstract: Somatic DNA mutations are the causal drivers of tumor progression yet predictions based on mutation profiles have proved challenging. Gain-of-function mutations in the PIK3CA gene is one of few prevalent in many cancer types; up to 30-40% of luminal subtype breast cancers harbor this mutation yet literature is divided of its effect on tumor metastasis and patient survival. Increasing availability of genome sequencing and gene expression data, as well as construction of human protein interactions maps, enables network analysis of biological systems 'as a whole' and not only of its parts. We constructed a breast cancer, luminal specific 'survival' network using mutation data from 358 breast tumours, focusing on positive gene regulatory loops; frequently recurring patterns involved in regulation and progression of tumorigenesis. Results: ~70% of the PIK3CA driver-mutated breast luminal tumours contained a SCH1 protein "feedforward" loop (FFL) that is predominately enriched. Surprisingly, these patients had significantly longer survival than those whose tumours contain PIK3CA mutation only, suggesting that this FFL serves as a protective mechanism by halting proliferation. We also compared proliferation rate between groups with PIK3CA mutant/SHC1-loop+ and PIK3CA mutant/SHC1-loop- using expression values of the established 11 'proliferation index signature genes', revealing a near 5 fold decrease in proliferation. Future directions: biological validation using genome editing to target one or multiple genes to assessing effects in specific cell lines. Clinical: aim to identify targets for drugs in regulatory loops involved in the progression of breast cancer, perturbing them, possibly slowing cancer progression.

TOP

A122 - PUB: A Pile-Up Based Backfill Pipeline for Identification of Sub-Clonal Mutations in Tumor DNA Sequencing Data

Yan Asmann, , United States
Vivekananda Sarangi, Mayo Clinic, United States
Jaime Davila, Mayo Clinic, United States
Chen Wang, Mayo Clinic, United States
Xue Wang, Mayo Clinic, United States
Jaysheel Bhavsar, Mayo Clinic, United States
Jean-Pierre Kocher, Mayo Clinic, United States
Grzegorz Nowakowski, Mayo Clinic, United States
Thomas Habermann, Mayo Clinic, United States
Andrew Feldman, Mayo Clinic, United States
Anne Novak, Mayo Clinic, United States
Susan Slager, Mayo Clinic, United States
James Cerhan, Mayo Clinic, United States

Short Abstract: Background: Tumors evolve by iterative steps of clonal expansion, genetic mutation, and sub-clonal selection due to growth advantages of the fittest sub-clones and/or external mutation induction and selection pressures such as radiation- and chemo-therapies. Studying the evolutionary dynamics of subclonal mutations is essential to understand how tumors progress and respond to therapy. However, the detection of subclonal mutations is still a significant challenge in bioinformatics. In addition, somatic tissues used in DNA sequencing are oftentimes low in tumor purity. Therefore the detection of clonal mutations can also be challenging due to the low concentration of reads supporting the mutation alleles.
Method: We implemented a comprehensive workflow for low-concentration and sub-clonal mutation detection which is based on the positional pile-up data (PUB). The variant positions with alternative bases were further evaluated and prioritized using a boosting method. The GLM was used as base learners with a subset of VQSR parameters. The somatic mutations were identified using modified Fisher's Exact Test.
Results: PUB is more sensitive and more specific for both clonal and sub-clonal variant detection, compared to the existing somatic callers. We tested the pipeline in tumor-normal paired exome sequencing and compared the characteristics of the "somatic" mutations called by PUB and by other existing somatic callers. We showed that existing methods missed large number of mutations with good variant qualities, and at the same time the existing methods called many false somatic variants where the alternative alleles clearly present in paired normal samples.

TOP

View Posters By Category

N) Sequence Analysis
O) Systems Biology and Networks
P) Other

Search Posters:

Last Name

search

TOP