1st European Conference on Translational Bioinformatics

Copenhagen

April 26-27, 2016

Organized by:

Novo Nordisk Foundation
Center for Protein Research

Co-organized by:

DET SUNDHEDSVIDENSKABELIGE FAKULTET
KØBENHAVNS UNIVERSITET

SYNAPSE
RESEARCH MANAGEMENT PARTNERS

MedBioinformatics

elixir
DENMARK

The work presented in the framework of the MedBioinformatics project has received funding from the European Union's Horizon 2020 Research and Innovation Programme 2014-2020 under Grant Agreement no 634143

Abstracts, talks

Data-driven Precision Medicine in the Global Context

Søren Brunak, University of Copenhagen, Denmark, **Ferran Sanz**, IMIM, UPF, Barcelona, Spain

Computational methods are increasingly playing a role in the clinic. Molecular level characterization of patients generates large amounts of heterogeneous data, data that needs to be integrated and analyzed in the context of corresponding phenotypic data, big biomedical data, from the healthcare sector. Disease progression patterns of patients with more than one disease have recently received strong attention within both molecular level systems biology as well as in epidemiology. Disease co-occurrences may be informative in relation to the underlying network biology of shared and multi-functional genes and pathways and in delivering knowledge on the interaction between the molecular level and external exposures stemming from diet, lifestyle and patient care. The opening remarks will mention these developments and describe briefly the MedBioinformatics project, a consortium coordinated by prof. Ferran Sanz, Barcelona. The MedBioinformatics project is a project within a group of projects under the general heading "Advancing bioinformatics to meet biomedical and clinical needs" which are funded within Horizon 2020 by the EU.

Somatic Mutation and Germline Variant Identification and Scoring from Cancer Patient Tumor Molecular Profiling and ct-DNA Monitoring by High-throughput Sequencing

Francisco De La Vega et. al, Stanford University School of Medicine, USA

Cancer tumor profiling by targeted resequencing of actionable cancer genes is rapidly becoming the standard enable the selection of targeted therapies and clinical trials for relapse cancer patients. In this clinical scenario, a tumor sample is obtained from a FFPE block and sequenced by targeted next-generation sequencing (NGS) to uncover actionable somatic mutations in relevant cancer genes. One of the challenges for the analysis of this data, is to distinguish between tumor somatic mutations, germline variants (which the cancer cell harbors), and sequencing errors (which occur at a rate of at least 0.5% for most NGS platforms). Most genetic variants in the tumor tissue would be germline, and distinguishing them form somatic mutations could be best accomplished by comparing to data obtained from normal tissue, which is not often available in the current standard of care. In addition, while the primary aim of the test is to inform targeted therapy selection from observed somatic aberrations, germline variants can inform the decision making and may be of relevance for the patient relatives if they confer disease susceptibility. Therefore, the analysis strategy cannot simply filter germline variants but should aim to properly identify these. Finally, monitoring of therapy and disease progression has been recently proposed by sequencing cell-free

tumor DNA from plasma samples. This more challenging assay, that needs to detect a few haploid copies of cancer cell DNA from a few milliliters of plasma, can be informed by the previous patient findings in their primary or metastatic tumor profiling. Here we present a principled approach to identify both single-nucleotide and small insertion/deletion somatic mutations and germline variants from NGS data of tumor tissue that leverages the allele fraction patterns in tumors an prior information from external databases through the use of a Bayesian Network. This approach allows us to score each putative mutation or variant with respect to their probability of belonging to each class or being a sequencing error. These scores can be used to define empirical filtering schemes for clinical interpretation. As more samples are analyzed, we can leverage their information as priors to improve the performance of our method. In addition, our method allows the joint calling of related samples form the same patient, in particular the case where a cf-DNA sample is sequenced from a patient where prior primary or metastatic tumor profiling is available, improving the limits of detection and scoring of somatic mutations in monitoring. We validate our method by analyzing data obtained with the TOMA OS-Seq targeted sequencing RUO assay for 98 cancer genes from model system samples of mixtures of well known genomes, patient cases where normal, tumor and cf-DNA are available, and a retrospective analysis of tumor patient data that underwent clinical tumor profiling for therapy selection.

Cancer Drivers and Their Therapeutic Opportunities

Nuria Lopez-Bigas, ICREA and UPF, Barcelona, Spain

Distinguishing the mutations directly involved in cancer, driver mutations, from the myriad of somatic mutations in a tumor genome is one of the major challenges of cancer research. This challenge is accentuated and currently unsolved for mutations in non-coding regions. Given the evolutionary principles of cancer, one effective way to identify genomic elements involved in cancer is by tracing the signals left by the positive selection of driver mutations across tumours. We have identified 459 cancer genes with driver mutations by analyzing close to 7000 tumor exomes from 28 different cancer types, and we have search for their targeted therapeutic opportunities. Currently we are analyzing hundreds of tumor whole-genomes to identify non-coding elements, including promoters, enhances, 5' and 3' untranslated regions, microRNAs and lncRNAs, with cancer driver mutations.

PanCancer to Patient-specific Pathways

Josh Stuart, UC Santa Cruz, USA

The particular alterations and mutations that arise in an individual's tumor may be shared or they may be distinct. Common molecular events may reflect the cell-of-origin, the oncogenic process, and the disrupted genetic pathways that contribute to tumorigenesis. Over the past several years, TCGA and other projects have amassed databases of tumor samples cataloging diverse events using various high-throughput platforms. These data have enabled a systematic classification of the different manifestations of cancer. While most tumors from similar tissues share common molecular signatures, some share cross-tissue similarities. I will present some surprises revealed by pan-cancer analyses and how unanticipated connections might be used to suggest treatments in pediatric cancers where few options remain. Our ultimate goal is to create a patient-specific model that captures not only the common aspects of a tumor that it shares more broadly with other patients with a particular subtype, but also its unique qualities. Our hypothesis is that the identification of n-of-1 networks, in which we adapt a pathway model to reflect both the common and unique aspects of disease, will help prioritize treatment options. I'll show an example in which we predict networks for men with metastatic prostate cancer using n-of-1 networks.

Genomic Classification and Personalized Prognostics for Acute Myeloid Leukaemia

Moritz Gerstung, European Bioinformatics Institute, Hinxton, UK

Acute myeloid leukaemia (AML) is an aggressive blood cancer with median survival of approximately 3yrs. We are presenting data from a screen of 111 cancer genes and conventional cytogenetics in 1540 patients, which allow for a comprehensive characterisation of the genomic landscape of AML and its association with clinical outcomes. A probabilistic clustering approach shows that there are at least 11 genomic subtypes, each characterised by particular constellations of genomic lesions. High-dimensional survival regression shows that approximately 2/3 of the explained differences in overall survival are related to aggregated genomics, which are build up by many small contributions from individual mutations. Novel multistage approaches for modelling 6 concurrent outcomes of treatment show that patient fate can be twice as accurately and more granularly predicted compared to current strata. These approaches allow for modelling the impact of haematopoietic stem cell transplants either in first complete remission or after relapse on a per patient basis. This provides a quantitative basis of clinical decisions and indicates that about 10% of transplant might be saved maintaining the same population level survival. Power calculations show that 10,000 samples are needed for clinical decision support algorithms with errors <1%.

APERIM: Advanced Bioinformatics Platform for Personalized Cancer

Zlatko Trajanoski, Medical University of Innsbruck, Austria

Cancer treatment platforms that involve the use of the adaptive immune system have demonstrated profound tumourregressions including complete cure. Importantly, technological advances in next-generation sequencing (NGS) allow for the first time the development of personalised cancer immunotherapies that target patient specific mutations.However, clinical application is currently hampered by specific bottlenecks in bioinformatics, which we aim to address in this proposal. The overall objective of our trans-disciplinary network of leading experts in bioinformatics and cancer immunology is to develop an Advanced bioinformatics platform for PERrsonalised cancer IMmunotherapy (APERIM).

Specifically we are developing:
1) Database for the integration of NGS data, images of whole tissue slides of tumour sections, and clinical data. To enhance the usability and the data sharing we will use semantic web technologies, and will provide standardised interfaces to a set of analytical tools.

2) Tools for automated quantification of tumour-infiltrating lymphocytes using whole tissue slide images and NGS data for patient stratification.

3) Analytical pipeline for NGS-guided individualised cancer vaccines including crucial NGS data analysis and epitope selection components for the selection of vaccination targets.

4) A method for deriving T-cell receptor (TCR) sequences from NGS data and predicting TCR specificity.

JAK-STAT Correlates How Protective Inflammatory Diseases Are to Alzheimer ’s disease

Alejo Nevado, Oxford University, UK

A well-documented epidemiological relationship exists between Alzheimer’s Disease (AD) and inflammatory diseases (McGeer et al 1996, Am Acad Neurol; Wallin et ak 2012, J Alzh Dis v31) and anti-inflammatory drugs (Lu et al 2015, Ann Rheum Dis). Certainly some biological and/or behavioural mechanism is producing this epidemiological relationship, but its identification has remained elusive. In our study we first show that combining epidemiological with genomic evidence points towards JAK-STAT as a possible mechanism of the AD-inflammation epidemiological link. Secondly, we analyse this hypothesis in two gene expression datasets, which confirm JAK-STAT anomalies exist in AD patients.

From Circadian Rhythms to Precision Medicine

Pierre Baldi, UC Irvine, USA

Circadian rhythms date back to the origins of life, are found in virtually every species and every cell, and play fundamental roles in functions ranging from metabolism to cognition. These rhythms play also important roles in health and disease states and should be taken into account in precision medicine, for instance to determine the optimal time at which a drug should be taken.

Modern high-throughput technologies allow the measurement of concentrations of transcripts, metabolites, and other species along the circadian cycle under a variety of conditions, thus creating novel computational challenges and opportunities for improving our fundamental understanding of circadian biology and its applications to precision medicine.

We will present several experimental results that have led to the development of new computational tools in circadian biology, including a general framework for understanding the pervasiveness and plasticity of circadian rhythms at the molecular level. We will also present deep learning methods to detect periodicity in time series and impute time from a set of high-throughput measurements, two necessary prerequisites for the application of circadian biology to precision medicine.

Paradigm Shifts of Precision Medicine in Oncology: Colorectal Cancer as a Model

Rodrigo Dienstmann, Vall d'Hebron Institute if Oncology, Barcelona, Spain & Computational Oncology Group, Sage Bionetworks Seattle, USA

In the early days of tumor genomic profiling, clinicians classified the disease using the single aberration perspective in order to make therapeutic decisions: one marker = one drug (KRAS exon 2 wild-type = anti-EGFR antibody). Initial results were somehow disappointing, as the majority of the patients did not benefit from a matched drug. With increased understanding of the complexity of the tumor genome, dynamics of target inhibition, clonal evolution under treatment pressure (spatial and temporal heterogeneity) and advances in drug development, we now deal with the multi marker = multi drug paradigm (all RAS and BRAF wild-type = anti-EGFR antibody in combination with MEK inhibitors). Results are promising and for the first time, clinicians are taking into consideration the genomic context to select the most appropriate (combination of) targeted drugs in order to delay emergence of resistant clones. In the near future, recognizing the interaction with tumor microenvironment, we will reach a multi-omics = (adaptive) immune drug paradigm, whereby a systems biology integrative analytical pipeline will determine successful clinical translation of novel biomarkers. I will discuss advances in matched targeted therapies in Colorectal Cancer as a conceptual model for Precision Medicine in Oncology, advancing from a clonal to a stromal-immune perspective.

The 100.000 Genomes Project

Tim Hubbard, Kings College & Genomics England

In December 2012 the UK Prime Minister announced the 100,000 genomes project to introduce whole genome sequencing for treatment into the UK National Health Service (NHS) on a large scale. Since then more than 6,000 whole genomes have been sequenced through pilots organised by Genomics England, the body set up to deliver the project. In addition major components to deliver the main project have been put in place: 13 NHS Genome Medicine Centres have been setup across England involving ~90 hospitals which will recruit patients and collect samples for sequencing and associated clinical data. Illumina was announced as the partner to deliver the whole genome sequences. Several companies have been contracted to provide initial genome interpretation services. Finally Genomics England has invited applications from UK researchers and NHS Clinicians to join its new Clinical Interpretation Partnership to analyse the data generated from the project. I will introduce the project and discuss the bioinformatics challenges of handling clinical grade whole genome sequence at scale to deliver both timely and usable summary reports to clinicians and a secure environment for research.

Targeting Cancers Using Individual Systems Medicine

Krister Wennerberg, FIMM, Finland

Our rapidly increasing understanding of cancer genomics holds great promise for driving precision cancer medicine. However, there are still big gaps between the genetic and molecular information we can generate today and what can be translated to the clinic. The Individualized Systems Medicine program established between researchers at the Institute for Molecular Medicine Finland (FIMM) and our clinical collaborators aims to address this translational gap by combining comprehensive functional chemosensitivity profiling and deep molecular and genetic profiling of cancer patient cells directly with clinical information and translation. Central to the program is the Drug Sensitivity and Resistance Testing where we profile the responses of primary leukemic cells to a comprehensive clinically oriented oncology collection of 525 clinical and investigational compounds. The drug sensitivity information is used to identify signal and network dependencies as well as effective drug combinations, and is further compared to molecular profiling information to establish hypotheses on individual cancer-selective targeting drug combinations and their predictive biomarkers. I will present i) informatic challenges we have encountered and some of our solutions we have established as well as ii) how we use the information to identify personalized therapies in leukemia and other cancers.

Exploring Disease Through the Lens of Data-driven Genomics

Tune Pers, Harvard School of Public Health & State Serum Institute, Denmark

Genomics has become a powerful approach to understand human disease. However, understanding how genetic polymorphisms impact health and disease ideally requires agnostic approaches, which are not biased towards previous conceptions about the disease. I will discuss how we can integrate genetic data with large-scale expression data to identify disease-specific tissues, cell types and biological pathways.

The Role of ELIXIR in Precision Medicine: Perspectives for the European Infrastructure for Biological Data

Niklas Blomberg, ELIXIR Hub, Hinxton, UK

The challenges in storing, integrating and analysing the data from modern biological experiments needs a coordinated effort that involves both national and international resources. ELIXIR, the European life-science infrastructure for biological information, is a European research infrastructure that bring together national life-science data centres, services, and core bioinformatics resources from the 20 member states into a single, coordinated infrastructure.

Open access to bioinformatics resources provides a valuable path to discovery. ELIXIR is identifying core data resources that are essential to the larger international community and is developing a robust framework to secure their long-term sustainability and accessibility. Some of these datasets are highly specialised and by coordinating local, national and international resources – hosted at over 120 institutes - the ELIXIR infrastructure will meet the data- related needs of Europe's 500,000 life-scientists.

ELIXIR is currently constructing a distributed e-infrastructure of bioinformatics services – a data nodes network - built around established European centres of excellence. This talk will discuss our approaches to handling, accessing and archiving large and also highly diverse data-sets in the human translational medicine space and how ELIXIR, in partnership with national cohorts and efforts such as the Global Alliance for Genomics and Health. The talk will discuss experiences in data integration and the need for establishing data-management plans within projects that address the issues of meta-data annotation and long term archiving.

Integrative Methods for Post-GWAS Functional Interpretation at Ensembl

Daniel Zerbino, European Bioinformatics Institute, Hinxton UK

The current abundance of genotype data and GWAS results has highlighted the need for reliable post-GWAS techniques to step confidently from summary association data to actionable drug targets. A majority of candidate causal SNPs fall outside of coding regions, and understanding their role in disease mechanisms is still more art than science. To overcome this obstacle, EMBL-EBI and the Ensembl project in particular are collecting a diversity of reference datasets covering genomes, variants, epigenomes and cis-regulatory interactions. Because of the size and diversity of these datasets, it is no longer practical to download them in bulk, hence we are deploying new technologies to make this data available through an open and integrated framework that allows users to perform complex analyses right on our servers. We demonstrate the power of this architecture with novel integrative multi-omic pipelines for genome annotation and post-GWAS functional interpretation.

Germline Mutation Hotspots at Functional Regulatory Sites

Martin Taylor, MRC Human Genetics Unit, University of Edinburgh, UK

Genetic mutations provide the raw material for evolution, they are responsible for heritable disease and driving the development of cancer. We have shown that the binding of chromatin and regualtory proteins to DNA can interfere with replication and lead to region with locally elevated mutation rates. Mechanistically this process appears to involve the trapping of DNA polymerase alpha synthesised DNA in the fully replicated genome; a process we have explored with a novel method, EmRiboSeq, that tracks replicative polymerase activity in vivo.

Extending this work we are measuring the patterns of chromatin accessibility and protein binding specifically in the mammalian germline and related it to the distribution of polymorphism and mutation, to reveal the terrain of replication associated mutations in mice and humans. This provides a means of adjusting neutral substitution rate estimates for fine-scale mutation rate fluctuation when identifying regions of selective constraint. We also identify likely hotspots of paternal lineage mutations within functional regulatory sites.

Integrative Omics to Study Mitochondrial Disorders

Holger Prokish, Helmholtz-Muenchen, Germany

Impairment of the mitochondrial energy metabolism presents with a wide range of clinical phenotypes. Causative defects have been identified in about 300 genes and a presumably large number of additional disease genes still await identification. Whole exome sequencing (WES) has proven to be valuable to genetically diagnose these patients and to suggest rational treatment options (e.g in cofactor metabolism). Therefore, WES is underway to be implemented at an early stage in the diagnostic algorithm of suspected infancy-onset mitochondrial disease.

Despite these successes, many disease causing mutations in patients with mitochondrial disorders still have to be identified. We applied WES in more than 600 unrelated individuals with suspected mitochondrial disorder. Still, in about half of the patients we were unable to identify the disease causing mutation. A significant part of the causing variants that are not identified by WES might be regulatory. In principle, whole genome sequencing (WGS) approaches allow the discovery of all variants not seen by WES. However, with the increased number of rare variants by orders of magnitudes variants prioritization and interpretation becomes the challenge.

Here I suggest performing integrative omics studies to directly identify regulatory defects, for patients with mitochondrial disorders. To this end, we established a collection of more than 200 fibroblast cell lines from patients, with WES or WGS, respiratory chain complex investigation, transcriptome sequencing, and quantitative proteome profiling. Combined analysis of DNA variation, allele specific expression and protein steady state levels in a cell line guides identification of DNA variants involved in the etiology of the disease.

Virus Discovery and Epidemic Tracing from High Throughput Metagenomic Sequencing (VIROGENESIS)

Kristof Theys, KU Leuven, Belgium

To date, only a proportion of the millions of short-length sequence fragments generated from metagenomes by Next-Generation-Sequencing (NGS) platforms is used for virus discovery, due to the lack of sensitive methods and tools that can accurately classify and assemble known and unknown viruses. Even for well-known pathogenic viruses, appropriate methods that can handle and characterize large and incomplete sequence datasets are lacking. The H2020 VIROGENESIS project will develop new mathematical, statistical and computational methods to address major bioinformatics bottlenecks in the analyses of new, diverse and complex virome data resulting from high-throughput NGS technologies. The project specifically aims to increase the resolution of current metagenomic classifiers, to improve the performance of phylogenetic and phylodynamic inference methods for NGS analyses and to design dynamic visualization software than can present

the wealth of information resulting from these bioinformatics software applications. VIROGENESIS will accelerate the translation of NGS analyses for viral pathogen discovery and detection, clinical diagnostics as well as near real-time epidemiological tracing disease control.

Integrative Bioinformatics Supporting Biomedical Research

Ferran Sanz, Research Programme on Biomedical Informatics (GRIB) & Hospital del Mar Medical Research Institute (IMIM), Universitat Pompeu Fabra Barcelona

The integrative analysis of the Biomedical Big Data (BBD) offers new opportunities for understanding the complex basis of diseases and, consequently, for designing better treatments for them. This BBD is constituted by information resulting from biological and pharmaceutical research (‘omics information, knowledge contained in the biomedical literature, etc.), data generated in the clinical practice (electronic health care records, medical imaging, etc.), as well as the health-related information that is published in social media (Web 2.0). Since most of this information is stored in not-structured formats, the computational techniques for automatic knowledge retrieval (e.g. text-mining) are paramount. This presentation includes several examples of integrative analyses of the BBD, such as the biological substantiation pipeline developed in the framework of the EU-ADR Alliance, a collaborative framework framework for drug safety studies (http://synapse-managers.com/projects/eu-adr-alliance/), as well as several analyses on disease commorbities carried out using the DisGeNET (http://www.disgenet.org) and PsyGeNET (http://www.psygenet.org) resources on gene-disease associations.

Computerome: Secure Private Cloud Computing for Person-sensitive Data

Peter Løngreen, Technical University of Denmark

The talk will address the design of the Danish National Life science computer supercomputer. The system is optimised for the workloads resulting from the heterogneous data deluge within life science. In connexion the talk will discuss how to deliver supercomputing capabilities through private/public cloud. The talk will discuss how to build private secure clouds and how to manage sensitive data by utilising Bare Metal Provisioning and Virtualisation techniques in a private cloud setting.

Disease Trajectories for Precision Medicine

Søren Brunak, University of Copenhagen, Denmark

It is increasingly acknowledged that biomarker information is often not identifiable in a bottom-up manner, and that clinical data and fine-grained phenotypes, e.g. from electronic patient records, are needed in order to establish many useful relationships. A fundamental question in establishing genotype-phenotype relationships is the basic definition of phenotypic categories. Patient record data remain a rather unexplored, but potentially rich data source for discovering correlations between diseases, drugs and genetic information in individual patients. Given the availability in Denmark of longitudinal data covering long periods of time we have the possibility of suggesting new phenotype definitions based on temporal analysis of clinical data in a more life-course oriented fashion. The talk will describe how the use of an unbiased, national registry covering 6.2 million people from Denmark can be used to construct disease bbcan "condense" millions of trajectories into a smaller set which reflect the most frequent and most populated ones. This set of trajectories can be interpreted as re-defined phenotypes representing a temporal diseaseome as opposed to a static one computed from non-directional comorbidities only. Such data makes it also possible to link comorbidities to the treatment history of the patients. A fundamental issue is to resolve whether specific adverse drug reaction stem from variation in the individual genome of a patient, from drug/environment cocktail effects, or both. It is essential to perform temporal analysis of the records for identification of ADRs directly from the free text narratives describing patient disease trajectories over time. ADR profiles of approved drugs can then be constructed using drug-ADR networks, or alternatively patients can be stratified from their ADR profiles and compared. This type of work can potentially gain importance in projects involving population-wide genome sequencing in the future.