ECCB 2014 Accepted Posters with Abstracts

G: Bioinformatics of health and disease

G01: Emile Rugamika Chimusa, Jacquiline Wangui Mugo and Nicola Mulder. Leveraging ancestry along the genome of admixed individuals to resolve missing heritability in disease scoring statistics

Abstract: Human genetics has been haunted by the mystery of “missing heritability” of common traits. Although studies have discovered several variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability. Resolving missing heritability, the difference between phenotypic variance explained by associated SNPs and estimates of narrow-sense heritability (h2), will inform strategies for disease mapping and prediction of complex traits. Among biased estimates of h2 due to epistatic interactions and rare variants not captured by genotyping arrays have been cited to be the most can be the most explanations for missing heritability. Here, we present an approach for estimating heritability of traits based on sharing local ancestry segments between pairs of unrelated individuals in an admixed population. From simulation data and real data, we demonstrated that our approach outperformed current approaches for estimating heritability of traits and holds values in admixture mapping for deconvoluting genes underlying ethnic differences in complex diseases risk.

G02: Sylvain Mareschal, Pierre-Julien Viailly, Philippe Bertrand, Fabienne Desmots-Loyer, Elodie Bohers, Catherine Maingonnat, Karen Leroy, Thierry Fest and Fabrice Jardin. Next-Generation Sequencing applied to tailor targeted therapies in lymphoma: the RELYSE project

Abstract: Non-Hodgkin Lymphomas (NHL) are lymphoid cell malignancies accounting for about 4% of all cancers, with an incidence rate of 12 cases per 100,000 and per year in Europe. While recently developed immuno-chemotherapies like Rituximab have significantly enhanced their prognosis, a large part of these patients still relapses or is refractory to treatments. Recent advances in the field of high-throughput sequencing have provided concrete solutions to support these patients, as whole-exome sequencing of large NHL cohorts has highlighted several recurrent mutations, and benchtop sequencers are now available to quickly screen these mutations at diagnosis to tailor personalized therapies.
The RELYSE project is an embodiment of this strategy, initiated by the french LYSA group in 2013. It consists of four major steps:
- The selection of a panel of 34 genes of interest in lymphomas, to be sequenced at diagnosis. The large collection of whole-exome studies already published in NHL was mainly used for this task, and was completed by a few more exomes of atypical (4 “leg-type” diffuse large B cell lymphomas) and refractory lymphomas (14 normal-tumoral pairs of diffuse large B cell lymphomas relapsed in less than a year).
- The parallel adaptation of this panel of genes to the main benchtop sequencer technologies: the Personal Genome Machine (Life Technologies) and the MiSeq (Illumina). For each sequencing technology, the available sequencing chemistries (AmpliSeq, Haloplex, SureSelect ...) were compared in terms of coverage, depth and ease-of-use for inclusion in the project.
- The sequencing of this gene panel in a retrospective series of 500 lymphoma biopsies, to assess clinical relevance and determine somatic mutation frequencies.
- The prospective analysis of about 200 refractory or relapsing lymphomas over 3 years, for

inclusion into current LYSA clinical trials. Dedicated software was developed to produce a synthetic report for physicians on the variations found and the available clinical trials to propose to the patient.

While still running, this project is an enriching experience in the progressive translation of Next-Generation Sequencing from the research field to clinical routine initiated a few years ago. Aside from contributing a useful stratification of patients for clinical trials, it will provide valuable experience on the implementation of such a strategy for larger scale future projects.

G03: Vivien Deshaies, Alban Lermine and Elodie Girard. Galaxydx – a web-server dedicated to ngs diagnosis data analyses

Abstract: Early cancer diagnostic is a challenge that can dramatically improve cancer treatment efficiency. High throughput sequencing technology is the more promising solution to reach this goal, but the analysis of their output is not straightforward and most of the time, need to launch software only available via command line interface.

Galaxy is a web platform that aim to: (1) make command line softwares accessible in an easy to use web interface, (2) construct personal workflows, (3) make analyses reproducible among time, (4) share know-how (workflow sharing) as well as data and annotations.

We built Galaxydx, an implementation of Galaxy containing a suite of softwares used for the analyses of diagnosis sequencing data (PGM torrent suite, BWA, GATK, VarScan, Annovar, … etc). Galaxydx allows Clinicians as well as Biologists to be autonomous to perform a complete set of analyses such as: (1) mapping, (2) variant calling, (3) variant filtering, (4) variant annotation, (5) rearrangements calling and (6) visualization through diagnosis dedicated Genome browser (Alamut).

We also work on data integrity and confidentiality by modifying the Galaxy writing methodology.

Analyses in Galaxydx are organized by project and user, output files are owned by the user who generates them. It allows us to systematically check system rights on data before any process (Does the current user can read input data? Does the current user can write in this project?)

G04: Pravinkumar Patchaiappan. In silico molecular docking studies of biological active compounds from aegle marmelos against p53

Abstract: Described as “the guardian of the genome”, p53 plays a critical role in tumour suppression by triggering apoptosis or cellular senescence in response to oncogenic stress or DNA damage. Consequently, mutational inactivation of p53 is the most frequent event found in about 50% of human cancers. This feature forms the basis of p53 being one of the most appealing targets for anti- cancer research, drug discovery and therapy. In the traditional Indian system of medicine, the Ayurveda, Aegle marmelos, commonly known as Bael, is an important medicinal plant. It is endowed with certain photochemical constituents that possess antineoplastic, chemoprotective and chemopreventive effects, properties efficacious in the treatment and prevention of cancer. Utilizing the technique of molecular docking, this study aims to predict interaction energy between p53 and 2 phytochemical constituents of Agele marmelos, namely Aegelinoside A and Marmin. The insilico molecular docking studies showed that while no interaction was seen between Aegelinoside A and p53, Marmin exhibited good interaction with p53. This potentially can pave way for future developments in the attempt to modulate function of p53 with Marmin, with an ultimate goal of restoring or controlling p53 functions in cancer patients.

G05: Marcos Avila, Daniel Torrente, Ludis Morales, Francisco Capani, Janneth Gonzalez and George E. Barreto. Structural insights from GRP78- NF-κB binding interactions: A computational approach to understand a possible neuroprotective pathway in brain injuries

Abstract: GRP78 participates in multiple functions in the cell during normal and pathological conditions, controlling calcium homeostasis, protein folding and Unfolded Protein Response. GRP78 is located in the endoplasmic reticulum, but it can change its location under stress, hypoxic and apoptotic conditions. NF-κB represents the keystone of the inflammatory process and regulates the transcription of several genes related with apoptosis, differentiation, and cell growth. The possible relationship between GRP78-NF-κB could support and explain several mechanisms that may regulate a variety of cell functions, especially following brain injuries. Although several reports show interactions between NF-κB and Heat Shock Proteins family members, there is a lack of information on how GRP78 may be interacting with NF-κB, and possibly regulating its downstream activation. Therefore, we assessed the computational predictions of the GRP78 (Chain A) and NF-κB complex (IkB alpha and p65) protein-protein interactions. The interaction interface of the docking model showed that the amino acids ASN 47, GLU 215, GLY 403 of GRP78 and THR 54, ASN 182 and HIS 184 of NF-κB are key residues involved in the docking. The electrostatic field between GRP78-NF-κB interfaces and Molecular Dynamic simulations support the possible interaction between the proteins. In conclusion, this work shed some light in the possible GRP78-NF-κB complex indicating key residues in this crosstalk, which may be used as an input for better drug design strategy targeting NF-κB downstream signaling as a new therapeutic approach following brain injuries.

G06: Robin Haw. Reactome Knowledgebase - Linking biological pathways, networks and disease.

Abstract: Modern health initiatives and drug discovery are focused increasingly on targeting diseases that arise from perturbations in complex cellular events. Consequently, there has been a tremendous effort in biological research to elucidate the molecular mechanisms that underpin normal cellular processes. A reaction-network pathway knowledgebase is the tool of choice for assembling and visualizing the “parts list” of proteins and functional RNAs, as a foundation for understanding cellular processes, function and disease. The Reactome Knowledgebase (www.reactome.org) is a publicly accessible, open access bioinformatics resource that stores full descriptions of human biological reactions, pathways and processes. Curated pathway knowledgebases, like Reactome, are uniquely powerful and flexible tools for extracting biologically and clinically useful information from the flood of genomic data. Specific features of Reactome support the visualization of interactions of many gene products in a complex biological process, and the application of bioinformatics tools to find causal patterns in expression data sets. To maximize Reactome’s coverage of the genome, we have supplemented curated data with a conservative set of predicted functional interactions (FI), roughly doubling our coverage of the translated genome. We have developed a Cytoscape app called “ReactomeFIViz”, which utilizes this FI network to assist biologists to perform pathway and network analysis to search for gene signatures from within gene expression data sets or identify significant genes within a list. Pathway and network-based tools for building and validating interaction networks derived from multiple data sets will give researchers substantial power to screen intrinsically noisy experimental data in order to uncover biologically relevant information.

G07: Niek de Klein, Sophie Rodius, Peter Nazarov, Arnaud Muller, François Bernardin, Céline Jeanty, Simone Niclou, Laurent Vallar and Francisco Azuaje. Connecting multiple gene expression signatures with candidate drugs for boosting heart regeneration potential.

Abstract: Cardiac injury in humans is mended by replacing cardiomyocytes with fibrillar collagen. This limits the amount of heart muscle available for pumping blood, causing greater strain on the heart. After damage, cardiomyocytes in humans proliferate at a very slow rate. While the initial steps of heart repair in the zebrafish are similar to those observed in humans, the zebrafish, in contrast to humans, is capable of heart regeneration. However, there is evidence for the possibility of boosting this property in humans. Here we aimed to find drug candidates that can, in the longer term, be repositioned to induce faster cardiomyocyte proliferation and fibrosis replacement in humans.

We identified genes that are involved in heart repair by performing differential expression analysis in healthy and injured heart samples from zebrafish. At different time points after cardiac injury, different groups of genes were activated. The top up- and down-regulated genes were selected at different post-injury time points as "expression signatures" of the different stages of heart repair. We searched the Connectivity Map (cmap) with these signatures. The cmap is a database of expression profiles measured in different human cell lines after treatment with different drugs, together with matched (untreated) controls. The cmap ranks the drugs based on the similarity between a query signature and a reference expression profile in the database. Typically, this resource is used for searching candidate drugs with single (query) signatures. In our case, multiple signatures are important at different stages of the heart regeneration process. To address this need, we developed a new method to combine the results from multiple matched signatures. We ranked the cmap results based on their “enrichment” score. We then used the rank product test to combine the ranking results from the different signatures into a single ranking. The highest ranking drugs are those likely to induce expression profiles that are similar to many of the query signatures. To focus researchers’ attention on promising candidate drugs, only the most biologically informative signatures are used for the combined ranking. Statistical enrichment of Gene Ontology terms of the human genes included in the signatures assisted in the selection of top promising signatures by selecting signatures with enriched functions related to the heart or regeneration. The top drugs are used for further investigation, including future independent experimental validation.

In conclusion, we have developed a new method for searching the cmap with multiple related signatures. This method is included in an analysis pipeline that uses zebrafish as an in vivo model of heart regeneration. This method will help us to find candidate drugs that may be repositioned for potential clinical use in humans. Our integrated search and ranking strategy may be extended to other cmap-based applications.

G08: Amelie Desvars, Linda Vidman, Chinmay Dwibedi, Maria Furberg, Pär Larsson, Anders Sjöstedt, Anders Johansson and Patrik Rydén. Modeling the spatiotemporal distribution of tularemia in Sweden

Abstract: Background: Tularemia is a zoonotic disease, caused by the highly infectious bacteria Francisella tularensis. F. tularensis can infect humans through different routes but mosquitoes are considered as the main vectors in Sweden. Incidence of tularemia in Sweden is one of the highest in the world and seasonality of outbreaks is highly marked (late summer-early autumn). It has been argued that spatial distribution of cases correlates with ecological niches and presence of rivers and lakes.

Objectives: The overall aims of the project are to: i) Answer questions about the geographical distribution of the disease which includes identification of endemic regions, identification of

correlated ecological niches and study of the geographical distribution of different genotypes; ii) Identify abiotic variables that significantly contribute to tularemia outbreaks within the endemic regions and to develop predictive models for these outbreaks.
Methods: Data on location, date of disease contraction, gender, age, and in some cases genotypes, were reported for tularemia cases in Sweden between 1984 and 2012. Demographic data were retrieved from Statistics Sweden and daily data from 336 meteorological and hydrological stations were obtained from the Swedish Meteorological and Hydrological Institute. To identify endemic regions we used the spatial scan statistic implemented in the SaTScan software. For some endemic regions several isolates were genotyped. The Spatial Distribution Independency (SDI) test was developed to assess if the geographical distribution of cases varied between genotypes. Similarly Mantel´s test was used to assess if the isolates genetic distances correlated with their geographical distribution. The inverse distance weighted method was used to interpolate meteorological and hydrological data. Mosquito data was only available for one location and a limited time period, but these data were used to model the relative mosquito abundance for all endemic regions. A number of explanatory variables were considered and negative binomial regression together with backward selection was used to model the annual number of tularemia cases within the endemic regions
Results: The spatial scan method identified seven endemic regions: Boden, Dalarna, Hammarö, Ljusdal, Ockelbo, Piteå, and Örebro. The SDI test showed no significant difference in the spatial distribution the genotypes in the Örebro region while the Mantel´s test showed that genetically close genotypes are geographically closer. The annual number of tularemia cases was successfully modeled using five explanatory variables where the relative mosquito abundance was among the most influential.
Conclusion: Our study identified endemic tularemia regions in Sweden. In a public health perspective, identifying factors that locally influence occurrence of outbreaks and spatial distribution of cases and genotypes are essential steps to reduce health and economical impacts of this disease.

G09: Nora Speicher and Nico Pfeifer. Integrative cancer subtype discovery using multiple kernel learning

Abstract: Cancer is a leading cause of death worldwide. To be able to treat cancer patients specifically, researchers aim to identify cancer subtypes that show distinct behavior concerning certain characteristics, e.g. survival time. Since tumorigenesis is a complex process often involving changes on different molecular levels of the cell, subtypes that are merely based on one data type, e.g. gene expression, do not fully reflect the subtleties of the tumor. Therefore, research nowadays also turns towards subtype discovery based on multidimensional data which can be provided by large-scale cancer genome projects, e.g. The Cancer Genome Atlas (TCGA).
We conducted a study on integrative subtype identification for glioblastoma multiforme (GBM). For data integration, we use the multiple kernel learning for dimensionality reduction framework proposed by Lin et al. [1]. Initially, we construct a kernel matrix using the radial basis kernel function for each data of a specific type. These kernel matrices are then linearly combined and the corresponding data points are projected into a lower dimensional subspace. Since simultaneous optimization of the projection matrix and the kernel weights is difficult, coordinate ascend optimization is conducted. Subsequent clustering of the projected data points determines the subtypes. Different parameter choices, e.g. initialization of the kernel weights or number of clusters, lead to different results of which we choose the best according to the average silhouette value of the clustering.
We applied this methodology to a TCGA GBM data set (preprocessed by Shen et al.[2])

consisting of 55 patients with measurements of gene expression, DNA methylation, and copy number variation. The clustering with the optimum silhouette value consisted of three clusters, one of them showing significantly increased survival time (p-value=0.025). The analysis of differentially methylated genes identified genes known from literature to have an influence on GBM, such as FNDC3B, DGKI, and FSD1 which affect the response to a specific treatment. Compared to iCluster[2], an integrative clustering algorithm based on a joint latent variable model, we achieved similar significant groups while having much lower computation time.

We showed that this framework is able to integrate information from different data types resulting in reasonable GBM subtypes. Furthermore, one can also use different kernels to capture further non-linear characteristics of the data. Besides, this methodology can easily be extended towards semisupervised learning and can therefore help understanding differences between cancer subtypes and assisting in treatment decisions.

[1] Multiple Kernel Learning for Dimensionality Reduction. YY Lin, TL Liu. IEEE Transactions on Pattern Analysis and Machine Intelligence (2011)
[2] Integrative Subtype Discovery in Glioblastoma Using iCluster. R Shen, Q Mo, N Schultz, VE Seshan, AB Olshen, J Huse, M Ladanyi, C Sander. PLoS ONE (2012)

G10: Andrew Nightingale, Tunca Dogan, Diego Poggioli, Guoying Qi, Jie Luo and Maria-Jesus Martin. The Role of UniProt's Protein Sequence Databases in Biomedical Research

Abstract: UniProt is the global hub for protein annotation in biological research. UniProtKB provides detailed manually curated annotation on proteins; plus an extensive set of cross-references through UniProt's close collaborations with bioscience dedicated data resources to further enrich its protein annotation with functional details, structural features, pathway interactions and alteration by genetic variability. In addition, UniProt provides human disease information that has extensive cross-references to disease relevant databases such as: Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM) definitions of the disease and other resources that are then mapped to UniProtKB entries. In order to further enhance the functional annotations relevance to biomedical research; UniProt has recently developed a pipeline for importing protein altering variants from globally recognised genetic variant repositories with the aim to extend the manually curated set of natural protein altering variants provided by UniProt. By combining all these resources UniProt aims to be an important resource for biomedical research and drug target identification.

Utilising the cross-references and the extensive protein structure and functional annotations, imported protein altering variants and the disease dictionary within the knowledge base; UniProt has been exploring how structural, functional and chemical ligand annotations can be used to identify relationships between a protein and disease causing variants.

UniProt has developed methodologies that link protein altering variants found within the binding site of proteins, with the structural and functional annotations for protein sequences in order to determine the potential these variants could have upon the binding of natural ligands or small molecule drugs. UniProt has been able to identify specific COSMIC variants within the binding sites of proteins associated with breast cancer that when examined at atomic resolution modify the mode of interaction the variant residue has with the proteins natural ligand or approved therapeutic drug. In these cases we find that the variant residue is either loosing specific interactions with the natural ligand or gaining interactions with the therapeutic drug.

These results illustrate the potential UniProtKB has as a useful resource for biomedical research and therefore from these findings UniProt is planning, in collaboration with its partner databases, to develop new services to aid biomedical research and drug discovery pipelines.

G11: Cristina Menni, Steven Kiddle, Massimo Mangino, Ana Vinuela, Maria Psatha, Claire Steves, Martina Sattlecker, Alfonso Buil, Stephen Newhouse, Sally Nelson, Stephen Williams, Nicola Voyle, Hilkka Soininen, Iwona Kloszewska, Patrizia Meccoci, Magda Tsolaki, Bruno Vellas, Simon Lovestone, Tim Spector, Richard Dobson and Ana Valdes. Circulating proteomic signatures of chronological age

Abstract: Objectives: To elucidate the proteomic features of aging in plasma and how these relate to early developmental pathways related to age-related diseases.
Methods: Using proteomics profiling on 1,129 blood proteins, we searched for age related proteins in 202 females from the TwinsUK cohort and replicated our findings in 677 independent individuals from the AddNeuroMed, Alzheimer's Research UK and Dementia Case Registry cohorts. We validated the results using RNAseq data from whole blood in TwinsUK. Finally, we investigated the role of specific proteins on a likely determinant of healthy aging (birthweight) and on each individual's 10-years Framingham cardiovascular score.
Results: 11 proteins were found to be associated with chronological age after adjusting for family relatedness and multiple testing, and were replicated at protein level in an independent population and further investigated at gene expression level in 384 females from the TwinsUK cohort. The most strongly associated protein was chordin-like protein 1 (CHRDL1, meta-analysis Beta(SE)=0.013(0.001), P= 3.66 x10-46) which also significantly correlated with weight at birth (0.06(0.02), P=0.005), a well know developmental marker of health status in mid-life and old age, and with the individual Framingham 10-years cardiovascular risk score in TwinsUK (0.71(0.18), P= 9.9x10-5). Another protein strongly associated with chronological age, was pleotrophin (PTN) (0.012(0.005), P= 3.88x10-41), a secreted growth factor with a plethora of functions in multiple tissues, and known to be a marker for cardiovascular risk and osteoporosis.
Conclusion: Our study highlights the importance of proteomics to identify some key molecular mechanisms involved in human health and aging. However, longitudinal studies are needed to understand the role of some of these molecules in the aging process.

G12: Johnathan Watkins, Kayleigh Ougham, Markus Mayrhofer, Anders Isaksson, Andrew Tutt and Anita Grigoriadis. Exploring large-scale somatic variation in cancer for biomarker discovery

Abstract: Motivation: Many tumour types, including triple-negative breast cancer (TNBC) and ovarian cancer are characterised by high levels of genomic instability. The scars left in the genome by these mutational processes may be measured to uncover mechanisms of genomic instability and develop biomarkers of response to drugs that exploit the defects driving genomic instability.
Methods and results: We used Affymetrix SNP6.0 microarray data on several different cohorts to generate allele-specific segmented copy number (CN) profiles. We first used the profiles of 8 colorectal cancer cell lines to test the accuracy of 6 different SNP microarray-based methods for predicting whole chromosome CN using publicly-available spectral karyotying (SKY) data as our reference. The CN of segments overlapping the centromere generated the least error on average. We next incorporated centromeric segment-informed chromosome-wise CN into a workflow to uncover complex patterns of aberration, tandem duplications (TDs), acquired uniparental disomy (ASUPD), structural CN aberrations (SCNAs), and whole genome doublings (NWGDs). We operationally defined complex regions of aberration that are likely to be the result of stepwise alterations to the genome, by testing whether the spatial clustering of aberration breakpoints deviated from what was

expected by chance. TDs were identified as regions of CN gain of < 2 Mb in size relative to the segments either side, and validated with publicly-available sequencing data for the 9 breast cancer cell lines. CN segments for which the minor allele CN differed from adjacent segments while retaining the same total CN were defined as ASUPDs. SCNAs were defined as segments whose CN deviated from the chromosome-wise CN, while NWGDs for a sample were calculated by looking at the proportion of major allele gain in the genome.
To explore the biology and possible translational utility of these measures, we applied our workflow to 940 TCGA breast and 552 ovarian tumours in order to characterise the prevalence, location and length of these different genomic measures on a tumour type-specific basis. Both SCNA and ASUPD were significantly more prevalent in TNBCs and in serous ovarian cancers compared to other breast and ovarian subtypes, respectively. SCNA alone was significantly associated with BRCA1/2 mutations in breast and ovarian tumours. Moreover, genomes from ovarian cancer patients who demonstrated good response to platinum therapy generally exhibited higher levels of SCNA than those who failed to respond.
Conclusion: We have established a workflow to capture a suite of large-scale genomic aberrations, in a manner that acknowledges the interdependency among chromosomal instability phenotypes. This workflow provides a framework for us to further investigate the origins, processes and outcomes associated with each genomic event measure. Some of these measures are likely to be reporting defects that bear relevance to drug biomarker identification.

G13: Tom Petty, S Cordey, I Padioleau, M Docquier, L Turin, O Preynat-Seauve, E Zdobnov and L Kaiser. ezVIR: a user-friendly bioinformatics tool for human virus diagnostics from high-throughput sequencing of clinical specimens

Abstract: High-throughput sequencing (HTS) provides the means to analyze clinical specimens at unprecedented molecular detail. While this technology has been successfully applied to virus discovery and other related areas of research, the sensitivity of HTS has yet to be exploited for use in a clinical setting for routine diagnostics. In this work a bioinformatics pipeline, ezVIR, was designed to process HTS data from any of the standard platforms and to evaluate the entire spectrum of known human viruses, providing diagnostic results that are easy to interpret and customizable. In addition to standard reports, ezVIR can also generate optional information for strain typing, detailed genome coverage histograms for comparing any of the detected viruses, and can perform cross-contamination analysis for specimens prepared in series. The pipeline was validated using HTS data from 20 clinical specimens representative of those most often collected and analyzed in daily practice. The specimens (5 cerebrospinal fluids, 7 bronchoalveolar lavages, 5 plasma, 2 serums and 1 nasopharyngeal aspirate) were originally found to be positive for a diverse range of DNA or RNA viruses by routine molecular diagnostics. The ezVIR pipeline correctly identified 14/14 specimens containing viruses with genomes < 40 000 bp, and 4/6 specimens positive for large genome viruses (e.g. members of herpesviridae). In this study, results indicate that the overall detection success rate, coupled to the ease of interpreting the analysis reports, makes it worth considering using HTS for clinical diagnostics.

G14: Pierre-Julien Viaillly, Arnaud Lefebvre and Hélène Dauchel. FunEVA: a user-friendly web application for Functional Effect Variation Analysis of human coding variations and their prioritization

Abstract: Whole and targeted-exome sequencing with next-generation sequencing (NGS) technologies have demonstrated to be reliable methods for mutation discovery and screening in medical genetics. These approaches provided new insights into molecular basis of

Mendelian disorders, complex diseases and driver mutations in cancer. The current bottleneck is no more acquiring sequences but the challenging NGS data management workflow, including efficient bioinformatics algorithms, reliable analytic pipelines and effective tools to help the variant interpretation. For the last step, command-line and rare web functional prediction tools propose to annotate and prioritize candidate genes [1]. Unfortunately, in the case of the web tools, the interfaces are not sufficiently thinking for biologist investigators, especially concerning the prioritization step. New free tools, taking advantage of effective prediction algorithms and up-to-date resources, and bridging the gap between the data deluge and friendly means of interpretation, are needed. Here, we present FunEVA, a user-friendly application designed to help functional interpretation in batch of coding human genetic variations and their prioritization. FunEVA combines as much as possible advantages for a non-programming biologist user. (i) Aggregation of multiple analyses of the potential functional impacts on corresponding proteins, based on reliable and updated international databases and tools resources (amino-acid properties, evolutionary conservation, potential pathogenicity predictions, functional signatures, structural properties, protein interaction, gene ontology terms and pathways, mutation repositories links). (ii) Simple input in tabular format (no heavy storage or upload of data) and batch submission of variants (up to five thousands), quick results (150 SNV/min). (iii) Friendly and exportable outputs in standard formats. For the overview of the full dataset, FunEVA displays aggregated results in a customizable, filterable and sortable table as well as in interactive visual representations, allowing an easy way for the prioritization. FunEVA also supplies, for every variation, a detailed individual report, cross-linked to numerous external resources and summarized thanks to an intuitive radar chart. We illustrate the functionalities and relevance outputs of FunEVA with two datasets of 1705 neutral SNV from Varibench [2] and 369 SNV from 14 exomes of Alzheimer individuals [3] pre-filtered thanks to the tool EVA [4]. In conclusion, within the context of large-scale studies in research and clinical settings, FunEVA provides a comprehensive functional interpretation of coding variants and an easy way to select a restricted number of candidate mutations before performing further investigations.

1.Pabinger et al., Brief Bioinform. 2014
2.Nair et al., Hum Mutat. 2013
3.Pottier et al., Mol Psychiatry. 2012
4.Coutant et al., BMC Bioinformatics 2012

G15: Arnaud Lefebvre, Alexandra Martins, Karim Labrèche, Vivien Deshaies, Alan Lahure, Pascaline Gaildrat and Hélène Dauchel. HExoSplice: a new software based on overlapping hexamer scores for prediction and stratification of exonic variants altering splicing regulation of human genes

Abstract: Sequencing exonic regions in disease-associated genes became a common practice for mutation screening in medical genetics. Besides their potential impact at the protein level, exonic sequence variations can induce aberrant splicing through disruption of Exonic Splicing Regulator elements (ESRs), leading to potential pathological effects. ESRs are described as 6-8 nucleotide motifs that regulate constitutive and alternative pre-mRNA splicing by recruiting trans-acting factors. Existing bioinformatics tools have limited performance to predict potential ESR signals and to assess potential alterations of ESRs by exonic variants. Moreover, their web interface present several restrictions in term of data submission and outputs.

Here, we present HExoSplice, a user-friendly web application designed for predicting and stratifying from a set of single nucleotide substitutions and with a quantitative predictive measure, candidate exonic mutations affecting splicing regulation through ESRs alteration. Concerning the scoring method, HExoSplice takes advantage of two previously published

studies [1, 2]. Firstly, it integrates individual scores of all possible RNA hexamers (4096) ranking their potential functions as ESRs [1]. Secondly, HExoSplice implements the computing of a new global quantitative predictive value (Total ESRseq score change) for any given exonic single nucleotide substitution to alter potential splicing regulatory signals [1, 2]. Briefly, it calculates the net effect of the six hexamers overlapping the nucleotide substitution, both in the variant and wide type (WT) contexts, and then it computes the differential value of Total ESRseq score change (variant versus WT). On the web interface, as input, HExoSplice takes one WT exonic sequence and simultaneously a set of variants within this sequence. After immediate batch computing, the output figures generated by HExoSplice allow a quick inspection of the full dataset. The results are displayed as predictive maps of potential ESRs and variation-induced alterations. Moreover, HExoSplice displays comprehensive score tables which allow ranking Total ESRseq score change values for stratification of exonic variant as potential splicing regulatory mutations.

In conclusion, HExoSplice implements a new scoring method for the prediction of the splicing regulatory mutations through an effective and user-friendly web interface. HExoSplice could help to quickly prioritize a restricted number of variations before performing further investigations. This approach could be particularly relevant for the evaluation of the so called Variations of Unknown Significance (VUS). More generally, HExoSplice could contribute to the annotation and filtering strategies of genetics variations identified by whole or targeted exome sequencing.

1.Ke et al., Genome Res. 2011, PMID:21659425

2.Di Giacomo et al., Hum Mutat. 2013, PMID:23983145

G16: Yong Li, Ekkehart Lausch, Anika Salfelder, Karl Otfried Schwab, Natascha van der Werf-Grohmann, Tanja Velten, Dieter Lütjohann, Pablo Villavicencio Lorini, Uta Matysiak-Scholze, Bernhard Zabel and Anna Köttgen. Whole exome sequencing identifies variants causing different monogenic diseases in one nuclear family

Abstract: Purpose: Whole exome sequencing has greatly facilitated the identification of mutations causing single-gene disorders, and has the potential to implicate genes previously not known to cause monogenic diseases. The purpose of this study was to search for causal mutations in two children from a consanguineous marriage. Both children are affected by hypomagnesemia and congenital hypothyroidism; one sibling also displayed hyperlipidemia. Methods: Whole exome sequencing was carried out using the Agilent SureSelect Exome Enrichment kit and the 5500 SOLiD system. Lifescope software was used for alignment and variant calling. Variants were tested for reproducibility using GATK. Quality metrics such as coverage, number of variants, percentage of known variants and the Ti/Tv ratio were evaluated. Results: Across individuals, mean coverage of the target region ranged from 40x to 63x, and target coverage at >20x ranged from 83% to 91%. After variant filtering (<1% population frequency; stop, splice, frameshift or missense variants; autosomal recessive mode of inheritance; and region of homozygosity obtained homozygosity mapping), 19 candidate variants were identified. One of these is the c.2667+1G>A mutation in TRPM6, which encodes a magnesium transporter expressed in intestine and kidney. This splice variant leads to skipping of exon 19 and is a known cause of autosomal recessive hypomagnesemia with secondary hypocalcemia (MIM #602014). In the child with hyperlipidemia, we identified a homozygous c.C1336T mutation in ABCG5, a sterol transporter. This known mutation introduces a premature stop codon and causes autosomal recessive sitosterolemia (MIM #210250). The other family members are heterozygous for this mutation. Biochemical analyses confirmed the diagnosis of sitosterolemia with elevated serum plant sterols. Mutations in both TRPM6 and ABCG5 were confirmed by Sanger sequencing. We are currently following up an intronic variant in TG gene as candidates for the thyroid phenotype.

Both children are treated with thyroxine since shortly after birth and develop normally. Conclusion: Different mutations causing single-gene disorders cluster in offspring from consanguineous marriages, and can give rise to complex clinical pictures. Treatments for all three single-gene disorders are available; the lipid-lowering treatment can be adapted for the specific diagnosis.

G17: Anthony Mathelier, Calvin Lefebvre, Jiarui Ding, David J. Arenillas, Wyeth W. Wasserman and Sohrab P. Shah. Cis-Regulatory Somatic Mutations and Gene-Expression Alteration in B-cell Lymphomas

Abstract: With the increasing availability of whole genome sequencing data, one has now access to a wealth of somatic mutation events specific to patients carrying a cancer. Classically, researchers focus on the events lying within protein-coding regions in order to extract the ones disruptive to protein structure and/or function. One can use whole-genome sequencing information to extract mutations lying within cis-regulatory regions with potential impact on gene regulation. Finding the cis-regulatory mutations driving carcinogenesis is an ongoing challenge.

We analyzed 84 matched tumour-normal whole genomes from germinal center B-cell like diffuse large B-cell lymphoma patients to elucidate potential cis-regulatory mutations involved in B-cell lymphomas. We specifically focused on mutations overlapping transcription factor binding sites (TFBSs), considered as cis-regulatory elements, for their potential impact on gene expression regulation. From a set of ~600 compiled ChIP-seq experiments, TFBSs have been predicted within peak regions using the most up-to-date JASPAR TF binding profiles collection. Looking at the substitution to insertion-deletion (indel) ratios within cis-regulatory elements and protein-coding exons, we show that the regulatory and coding spaces are under similar functional constraints. Hence, we highlight here the importance of analyzing the cis-regulatory space of the genome and not only the protein-coding space. By comparing somatic signatures of single nucleotide variation (SNV) mutation trinucleotide context

within TFBSs to the ones within protein-coding exons, we highlight the prevalence of C>T mutations in exons. We observe that B-cell lymphoma mutations overlapping TFBSs significantly cluster at promoter regions of genes involved in apoptosis or carcinogenesis in general.

By coupling mutated genes (in protein coding exons and/or TFBSs) with RNA-sequencing data, we used a modified version of the xSeq tool (under

development) to predict candidate cancer driver genes in samples where the mutated genes show altered expression in combination with altered expression of their known interacting genes in biological networks. Focusing on SNVs and indels overlapping TFBSs and looking for their potential impact on gene expression, we extract and prioritize candidate driver genes with disrupted expression associated to cis-regulatory mutations. This analysis culminates with potential cis-regulatory mutations altering gene expression of candidate driver genes. Our analyses demonstrate the importance of in-depth characterization for mutations lying within cis-regulatory elements and provide potential candidate as cis-regulatory variations impacting gene regulation in B-cell lymphomas. They represent the preliminary steps toward large-scale analysis and prioritization of mutations altering gene regulation through the disruption of cis-regulatory elements in the genome.

G18: Johanna Mazur, Isabella Zwiener and Harald Binder. Combining gene expression measurements from different platforms with a stratified boosting approach

Abstract: To develop reliable gene expression risk prediction signatures in a survival setting, a high number of samples is needed. Combining several data sets simultaneously is one feasible way to overcome this limitation. However, direct pooling of individual patient data is not possible for gene expression studies that are performed on different platforms, like RNA-Seq and microarrays.

We propose here a stratified boosting approach for regularized estimation of Cox regression models to still be able to combine gene expression measurements from different platforms. For every study, i.e. every stratum, a componentwise likelihood-based boosting algorithm is performed where the variable that is updated in every step is the one where the score statistic is the largest across studies.

Our stratified boosting approach is evaluated on simulated data for the two following settings:

1. the prediction performance is compared to the prediction performance of the pooled analysis
2. the performance to identify important genes is compared to the pooled analysis' performance to identify important genes and a setting where only gene lists but not the data itself is available for the gene expression studies

Additionally, we evaluate our approach to RNA-Seq and gene expression microarray data from kidney clear cell carcinoma patients.

The results show that our newly proposed stratified boosting approach gives comparable results to the pooled analysis, where the latter is feasible. In addition, it makes it possible to combine gene expression studies from different platforms and gives a good starting point to combine studies from different molecular platforms, like gene expression and methylation.

G19: Medi Kori and Kazım Yalcın Arga. Uncovering the Interconnectivity Between Infertility-Associated Woman Diseases

Abstract: Background: Infertility became a major problem worldwide. Nowadays, approximately 50 million couples worldwide are affected from infertility. The goal of this study is to integrate transcriptomics datasets with biological networks to (i) identify candidate genes, proteins and metabolites that have the potential for being biomarker for infertility-associated woman diseases, and (ii) map the interconnectivities between genetic mechanisms of these diseases.

Method: Nineteen transcriptomics datasets for infertility-associated woman diseases (polycystic ovary syndrome, endometriosis, ovarian, cervical cancer and uterine leiomyomas) were analysed statistically (via RMA and LIMMA methods) to identify differentially expressed genes (DEGs) for each disease. Proteins encoded by DEGs were determined and data was integrated with reconstructed protein-protein interaction network and human Recon2 metabolic model for further analyses (via BioMet toolbox) to identify reporter genes, proteins and metabolites. Enrichment analyses were performed (via DAVID bioinformatics tool) to map the interconnectivities between diseases and biological pathways.

Results:Integrated analyses indicated that (i) samples of same disease obtained from different donors were representing variable gene expression profiles (i.e., 1-25% of the genome was differentially expressed); (ii) the inspected diseases were representing high interconnectivities at protein-protein interaction level; and (iii) several genes and proteins were representing potential for being biomarker for these diseases.

Conclusion:Analysis of whole genome expression profiles provides a good opportunity for systems level investigation to characterize the interconnectivity between infertility-associated woman diseases and provides hypotheses for further in vivo studies to characterize novel biomarker proteins.

G20: Pedro Brazão-Faria, Alexandra M. Carvalho, Susana Vinga and Nuno L. Barbosa-Morais. Analyses of alternative splicing landscapes in clear cell renal cell carcinomas reveal putative novel prognosis factors

Abstract: The recent development of next-generation sequencing (NGS) largely improved our means to study transcriptomes. By RNA sequencing (RNA-seq, the use of NGS to sequence cDNAs reversely transcribed from RNAs), one can not only quantify gene expression (GE) levels, with a higher resolution than microarrays, but also quantitatively reveal unknown transcripts and splicing isoforms. However, the use of RNA-Seq to find cancer transcriptomic signatures beyond GE has been very limited, partly due to a lack of accurate and efficient computational tools. For instance, to our knowledge, the wealth of The Cancer Genome Atlas (TCGA) dataset has not yet been employed to systematically study deregulation of alternative splicing (AS, regulatory process by which multiple, distinct RNAs are originated from the same primary transcript by different splicing patterns) in cancer, despite the lying of disease-causing single nucleotide polymorphisms within putative splicing regulatory sequences and abundant evidence for the role of splicing regulation in epithelial-mesenchymal transition and cellular programs altered in oncogenesis.

We conducted GE, AS and associated survival analyses on RNA-seq data for 62 clear cell renal cell carcinomas and their matched normal kidney samples from the TCGA project. To identify cancer- and stage-specific AS patterns, we ran MISO (Mixture of Isoforms) on the TopHat-generated genomic alignments of reads and then employed non-parametric methods (Wilcoxon rank-sum and Kruskal-Wallis analysis of variance, respectively) to compare between samples. In addition, Log-rank tests were used to carry out survival analysis in order to identify AS prognostic factors. Multiple testing corrections were also performed in the selection of the putatively relevant events.

We observe that AS patterns primarily separate normal from tumour samples, with some exons exhibiting a normal/tumour “switch” pattern in their inclusion levels. This is the case, for example, for genes such as CD44 and FGFR2, previously reported to undergo AS alterations in cancer. Several AS events appear to be associated with tumour stage and survival, being therefore identified as potential novel disease biomarkers and prognostic factors.

These results suggest a great potential of AS signatures derived from tumour transcriptomes in providing etiological leads for cancer progression and as a clinical tool. A deeper understanding of the contribution of splicing alterations to oncogenesis could lead to improved cancer prognosis and contribute to the development of RNA-based anticancer therapeutics, namely splicing-modulating small molecule compounds.

This work was partially supported by FCT, Portugal under contracts PEst-OE/EME/LA0022/2011, CancerSys EXPL/EMS-SIS/1954/2013, IF/00653/2012, co-funded by the European Social Fund (ESF) through the Operational Program Human Potential (POPH) and Marie Curie International Outgoing Fellowship FP7-PEOPLE-2010-IOF EvoAltSplice.

G21: Anita Schuerch, Debby Schipper, Maarten A. Bijl, Jim Dau, Kimberlee B. Beckmen, Claudia M.E. Schapendonk, V. Stalin Raj, Albert D. M. E. Osterhaus, Bart L. Haagmans, Morten Tryland and Saskia L. Smits. Metagenomic survey for viruses through iterative assembly of taxonomic units

Abstract: Recent outbreaks of previously unknown (‘emerging’) viruses, such as the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) highlight the need for vigilant surveillance to detect new threats to public health. As many (re-)emerging viral infections originate from animal sources, it is important to obtain insight into viral pathogens present in

the animal reservoir from a public health perspective.

To allow a reproducible analysis of the metagenomic data from tissue or body fluids, we developed an automatic analysis pipeline. The pipeline enables demultiplexing, stringent quality trimming and an exhaustive iterative assembly of the reads to define taxonomic units in the metagenome, followed by homology search and taxonomic assignment.

This workflow was applied to a viral survey of the Western Arctic herd of barren ground caribou (Rangifer tarandus granti) in Alaska, USA. It identified the presence of several mammalian viruses, including different papillomaviruses, a novel parvovirus, polyomavirus, and a virus that potentially represents a member of a novel genus in the family Coronaviridae.

We benchmarked the assembly approach in terms of contig length, chimera generation and number of assembled reads. Rigorous assembly does not only reduce the input to the subsequent homology search but also increases the likelihood of identification of a homolog and taxonomic assignment. Moreover it allows for near-complete assembly of viral genomes with sufficient coverage.

The development of viral metagenomics platforms and bioinformatics pipelines enables a comprehensive inventory of viruses present in animals and humans. In silico identification of viral genomes from metagenomes allows rapid characterization of viruses and timely measures to be taken in response to an infectious disease outbreak.

G22: Jelmar Quist, Johnathan A. Watkins, Pierfrancesco Marra, Andrew Nj Tutt and Anita Grigoriadis. Uncovering driver mechanisms of cancer using a gene module based approach

Abstract: The analysis of human cancers by various types of arrays revealed that many tumours contain a large number of genes with altered copy numbers or expression levels. Many of these genes will not provide a selective advantage to the tumour and are therefore referred to as 'passenger genes'. Genes that do provide a selective advantage, and thus promote cancer progression, are referred to as 'driver genes'. The task of identifying such driver genes in large datasets has proved challenging, but could lead to the discovery of new therapeutic targets.

For this purpose, a computational workflow was developed to perform an integrative analysis of copy number and gene expression data. We adapted the previously published COpy Number and EXpression In Cancer (CONEXIC) algorithm to handle absolute copy number data, which was derived from Allele-Specific Copy number Analysis of Tumors (ASCAT) and adjusts for tumour ploidy and normal cell contamination. We applied the workflow independently to a discovery (n=88) and a validation (n=112) triple-negative breast cancer (TNBC) cohort. Copy number data was available from the Affymetrix SNP 6.0 array and gene expression data from the Affymetrix GeneChip Human Exon 1.0 ST array and Illumina HumanHT-12 v3 BeadChip, respectively.

The workflow identified 44 gene modules in the discovery cohort, each consisting of genes with a similar gene expression pattern. Of these gene modules, 38 were driven by 29 potential driver genes whose expression levels were copy number driven. In the validation cohort, 44 gene modules were identified of which 36 were found driven by 47 potential driver genes. Between the discovery and validation cohort, 110 gene module combinations were found to contain more genes in common than expected by chance. From these, 2 gene module combinations were found to be driven by the same potential driver gene, i.e. EXO1. This exonuclease is implicated in a multitude of DNA metabolic pathways and helps to maintain genomic integrity. Interrogating EXO1-driven gene modules revealed functional commonalities further supporting the identification of EXO1 as a potential driver gene in this type of cancer. Thus, our computational workflow adapted to the tumour-specific characteristics of TNBCs will help to extend our knowledge on the copy number driven transcriptional changes and their potential functional mechanisms affecting TNBC biology.

G23: Mathilde Daures, Anne Degavre, Cécile Julier and Anne Philippi. GAMES: Genetic Analysis and Mining of Exome Sequencing

Abstract: In recent years, there has been a generalization of Next Generation Sequencing technics, which are now widely used in human genetics. Currently, Whole Exome Sequencing (WES) has been generally favored over Whole Genome Sequencing, both for cost and technical reasons, and because they have mainly targeted Mendelian diseases, which are known to be generally caused by protein impairing variants[1]. Despite the restriction to mostly coding variants in WES, it may still be challenging to identify the disease causing mutation among a mean of 22000 variants (400 novel ones) per individual[1].

To date, several bioinformatics tools dedicated to WES analysis have been developed, which however have some limitations: organism specificity[2], web server form[3], basic filtering[4], disease specificity[5], or non-availability or cost (private companies). Here, we present a new bioinformatics pipeline to filter and exhaustively annotate exome variants in order to select and shortlist candidate variants according to specific criteria.

GAMES is a 2-stage python tool. In the first stage, it filters the variants according to their quality, the consistency with the chosen genetic disease model and frequency constraint (e.g. absence) in control databases (public and internal). It also integrates available genetic data (linkage, homozygosity regions…) and records the impact of all variants, including the predicted severity of the protein coding ones based on Polyphen and SIFT. In the second stage, it annotates the 1st stage selected variants with an exhaustive set of relevant information, such as the summary function description of affected genes (from Refseq), their pathways (from Gene Ontology), their related bibliography (from PubMed) and their specific tissue expression (from UCSC). This is achieved by directly calling these databases using python modules (Biopython and SqlAchemy). We also cross-reference relevant in-house information on selected genes, diseases and phenotypes.

Using GAMES on 40 patients, we were able to reduce the number of variants, depending on the given constraints (e.g. recessive transmission, rare variants (frequency < 1%)). Combining with the complementary information of stage 2, as described above, we short-listed a mean of 5 genes (1 – 10) by patient as first candidate.

In conclusion, GAMES is a useful tool to filter and annotate exome variants in order to prioritize and select them for subsequent experimental validation. GAMES is very flexible and adaptable to various genetic scenarios from Mendelian diseases to complex traits. At term we will extend further the gene annotation options, generate a prioritization score and create a graphical interface.

1. Bamshad, M. J. Nat. Rev. Genet. 12, 745–55 (2011).
2. Al-Shahib, A. BMC Bioinformatics 14, 326 (2013).
3. Vuong, H. Bioinformatics 1–2 (2013).
4. Pope, B. J. BMC Bioinformatics 14, 65 (2013).
5. Swaminathan, G. J. Hum. Mol. Genet. 21, R37–44 (2012).

G24: Therese Kellgren and Patrik Rydén. Experimental designs for finding disease-causing mutations in rare diseases

Abstract: Background: Rare diseases are commonly explained by a mutation in a single gene. Many of these genetic mutations can be passed on from one generation to the next, explaining why certain rare diseases run in families. A common objective in healthcare and medicine research is to locate these mutations to find a treatment for the disease. A frequently used method to find these hereditary mutations is exome sequencing, but the technique is relative expensive which limits the number of sequenced individuals. We consider the situation when

there is a family described by a pedigree where some healthy and diseased individuals are available for exome sequencing. Here we investigate statistical methods to identify the risk allele in rare diseases with a focus on experimental design.

Objectives: The aim is to develop a statistical approach for identifying disease-causing mutations in rare diseases which minimizes the number of false positives. The approach includes both a tool to derive an experimental design and a pipeline for analyzing the exome sequencing data.

Methods: A pipeline for downstream analysis after performing exome sequencing was created. The pipeline can be considered as a function g(XS(n),yS(n),p) generating a list L of potential disease-causing mutations. Here XS(n) denotes the data from the n sequenced individuals (S indicates which individuals were selected), y indicates which individuals were diseased and p indicates whether the disease was recessive or dominant. Suppose that the pipeline always identifies the risk-allele, then the objective with the experimental design is to find the selection S(n) that minimizes the expected length of the list L. We use Monte-Carlo simulations to determine the optimal selection S(n).

Results and future work: The pipeline was used in an exome sequencing study of a rare eye-disease, Bothnia Corneal Dystrophy. One risk allele was identified and confirmed by independent functional studies. This result gave us an indication on the most profitable settings. To investigate whether this study had an optimal experimental design we plan to simulate genomes and examine how different experimental designs effects the number of expected false positives.

G25: Elizabeth Baker, Hamel Patel, Mizanur Khondoker, Stephen Newhouse and Richard Dobson. The use of Polygenic Risk Scores for predicting rate of cognitive decline in Alzheimer’s disease

Abstract: Background: There is huge variability in how patients with Alzheimer’ Disease (AD) progress in terms of their rate of cognitive decline. It is reasonable to assume there is a genetic component to rate of progression, yet whilst there has been considerable progress in determining the genetic underpinnings of AD itself, with a recent study by Lambert el al identifying 11 new loci in the largest study to date, few studies have investigated the genetics of decline beyond the Apolipoprotein E gene. With rate of cognitive decline being a progressive phenotype with sigmoidal change over time, it may well be that multiple genes of low effect are acting together to infer susceptibility to AD progression. In this study we investigate the association between an AD polygenic risk score (PRS) and a patients rate of decline, to investigate whether patients having a high burden of risk alleles tend to be rapid decliners.

Method: 874 subjects from the EU IMI AddNeuroMed (ANM) have been genotyped and imputed based on the 1000 Genomes datasets. Using SNPs from two large GWAS studies of AD diagnosis (Harold et al 2009 and Lambert et al 2013), a PRS has been calculated for each AD individual within ANM. Rate of cognitive decline estimates have also been generated for these subjects and as such the association with PRS will be investigated.

Results: This study will quantify the association between the genetic risk burden harbored by a patient and their rate of cognitive decline in disease.

Discussion: Our results will add to understanding of the genetic component of cognitive decline in AD subjects and improve our understanding of the contribution of genetic factors to the variability of AD progression. The polygenic risk profile may aid prediction of future cognitive decline in AD patients.

References

Harold et al Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer’s disease October 2006 41(10) 1088-1156

Lambert et al Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease December 2013 45 (12) 1450-1458

G26: Owen Lancaster, Tim Beck, Raymond Dalgleish and Anthony Brookes. Cafe Variome: enabling the federated discovery of genetic variant and phenotype data

Abstract: Cafe Variome (http://www.cafevariome.org/) enables the fully open and comprehensive discovery of variant, phenotype and patient data that cannot be openly shared by conventional means for legal, ethical or competitive reasons. Cafe Variome makes the existence rather than the substance of the data openly accessible, by providing a "shop window" onto what and where data exist. When used between networks of diagnostic laboratories or disease consortia that know/trust each other and share an interest in certain causative genes or diseases, Cafe Variome provides the mechanism for the discovery of rare sequence variants or patients with rare disease phenotypes.

The system allows users to openly search the full content of the data, via a sophisticated central "Google-like" query interface, and thereby determine whether or not a record of interest exists in an information resource. Users of the system can subsequently access the hit data (according to pre-set permissions) in line with one of the following conditions:

1. Open Access: the user may access variant and patient records directly and freely
2. Linked Access: the user can view summary data and is provided with a link to the source database to access the full record
3. Restricted Access: the user may access variant and patient records if they belong to a pre-approved group or must request access from the data owner to the full record

Phenotype descriptions in Cafe Variome can be standardised by annotating each variant with any number of terms from relevant NCBO BioPortal ontologies. The flexibility with regards to the number of permissible annotations allows a variant to be associated with a single disease term, or a complex combination of phenotype descriptions. The phenotype annotations can be interrogated via a term autosuggest feature from the central query interface, or via searchable term trees that are generated for each of the ontologies used.

Cafe Variome offers a complete data sharing software solution (either a hosted or an "in-a-box" solution) all controlled by an intuitive administrator dashboard, which gives owners complete control over their data and installation. Multiple installations can be connected together to form federated networks to allow controlled queries across nodes within the network.

G27: Clara Benoit-Pilven, Amandine Rey, Léon-Charles Tranchevent, Marie-Pierre Lambert, Hussein Mortada, Emilie Chautard, Laura Corbo, Béatrice Eymin and Didier Auboeuf. Alternative splicing and resistance to cancer targeted therapies

Abstract: Targeted therapies are commonly used to treat cancer but they often fail due to resistance to the treatment of some tumours. Resistance can happen via different mechanisms, and alternative splicing appears to be one of them. Indeed, alternative splicing is the process of creating distinct proteins from a single gene, and recent reports indicate that therapeutic targets often produce isoforms that do not respond to the targeted therapy.

We propose to better define the role of alternative splicing in cancer drug resistance with a systems biology approach and experimental validations on breast and lung cancer cell lines. The main objectives are to develop a computational method that will help users to analyse the role of alternative splicing of therapeutic targets in resistance and to develop a proof-of-concept in cancer cell lines by predicting which cell lines exhibit de novo resistance to a given treatment, owing to alternative splicing.

The first outcome is a web interface freely available for users to analyse the therapeutic target

variants up to the protein level to assess the effect of splicing on protein domains and therefore on function. Another outcome is an experimentally validated method to predict which cell lines are resistant to which treatment.

This represents a first step towards the development of more advanced methods that will take into account all possible alterations observed in human cancers in order to refine the population stratification for clinical trials and possibly in the future modify the way cancer is treated with multiple targeted therapies.

G28: Anna Feldmann and Nico Pfeifer. Predicting and Understanding HIV-1 Susceptibility to Broadly Neutralizing Antibodies

Abstract: The advent of the highly active antiretroviral therapy (HAART) in 1996 revolutionized HIV-1 treatment. However, clinicians still have to face drug-resistant viral strains, while the amount of both available drugs and drug targets remains limited. Recently, combination therapy with broadly neutralizing antibodies (bNAbs) was introduced as a prospective HIV-1 treatment that is capable to reduce viral load under detectable levels for up to 60 days in humanized mice. An additional study in SHIV-infected rhesus monkeys showed that treatment with a combination of bNAbs can be successful in primates. However, similarly to HAART, the emergence of neutralization resistant strains is a major problem to select an efficient combination therapy of bNAbs for an individual patient.

In this work, we developed prediction models for seven different bNAbs using support vector machines. The prediction models are able to determine the neutralization susceptibility of unseen viral strains to the specific bNAb based on the viral envelope sequence (Env). Different string kernels as well as the polynomial kernel and the Gaussian RBF kernel were tested in a nested cross-validation. Among all considered kernels, the oligo string kernel performed best. High prediction performance could be traced back to learnt discriminant features that are supported by literature.

The classifiers for the V3 loop directed antibodies (PGT121 and 10-1074/10-996) achieved the highest performance (mean classifier AUC = 0.81 ± 0.03) identifying the N-glycosylation site at position 332 as the most discriminant residue for neutralization susceptibility and the N-glycosylation site at position 334 as an indicator for neutralization resistance. For the V1/V2 loop site directed bNAbs (PG9 and PG16), the N-glycan site at position 160 was found by the classifiers to be most discriminant for neutralization susceptibility (mean classifier AUC = 0.69 ± 0.02). The classifiers for the CD4 binding site-directed bNAbs (VRC01 and VRC-PG04) recognized different known binding sites to the CD4 molecule on the gp120 Env subunit to be associated with neutralization susceptibility of viral strains and had a mean AUC of 0.70 ± 0.01.

In addition, a new procedure was developed that is able to visualize the classification results of non-linear kernels to increase model interpretability and acceptance by potential users. The robustness of the learnt models implies that similar models can also be learnt for additional bNAbs given the availability of neutralization panel data.

G29: Noémie Robil, Benoit Grellier, Fabien Petel, Ronald Rooke and Jacques Haiech. A new gene expression-based tool for selecting putative membrane cancer antigens named KANT

Abstract: Monoclonal antibodies are promising agents for use in cancer treatment. Passive immunotherapy could gain in efficacy in several cancer types by increasing the number of responding patients, decreasing disease recurrence or diminishing off-target-related toxicity. This is why we need to find putative cancer antigens for new therapies and biomarkers to identify subtypes of cancer. We describe here a new method for identifying potential cancer antigens that would be accessible to antibodies in the extracellular milieu and encoded by

genes specifically overexpressed in cancers.

This identification can be broken down into four key phases: (a) The annotation of transmembrane proteins; (b) the analysis of gene overexpression in specific cancers; (c) the prioritization of the most relevant proteins and (d) biological validation. We focused on the first three steps with the development of a method applicable to all types of cancers.

We first tested various algorithms of transmembrane domain prediction using a benchmark dataset: the best accuracy (97%) was obtained using a modified version of MEMSAT-SVM algorithm. We applied this modified algorithm to the Uniprot dataset, and predicted 5000 proteins (25%) to include a transmembrane domain. We then developed a new expression-based score to identify, among genes coding for transmembrane proteins, those showing specific overexpression in cancer cells, and low or null expression in non-tumor cells. To validate this strategy, we analyzed a large publicly available breast cancer dataset and identified 13 candidate cancer antigens. Of these candidates, seven are already known to be involved in breast cancer, and one, MUC1, has been validated as a therapeutic target and is under clinical investigation in a phase 2 trial. We also developed a method to prioritize these candidates.

Results of the first step and the second step are available as an R package.

G30: Abhishek Dixit and Richard Dobson. BBGRE: brain and body genetic resource exchange

Abstract: Studies of copy number variation (genomic imbalance) are providing insight into both complex and Mendelian genetic disorders. Array comparative genomic hybridization (array CGH), a tool for detecting copy number variants at a resolution previously unattainable in clinical diagnostics, is increasingly used as a first-line test at clinical genetics laboratories. Many copy number variants are of unknown significance; correlation and comparison with other patients will therefore be essential for interpretation. We present a resource for clinicians and researchers to identify specific copy number variants and associated phenotypes in patients from a single catchment area, tested using array CGH at the SE Thames Regional Genetics Centre, London. User-friendly searching is available, with links to external resources, providing a powerful tool for the elucidation of gene function. We hope to promote research by facilitating interactions between researchers and patients. The BBGRE (Brain and Body Genetic Resource Exchange) resource can be accessed at the following website: http://bbgre.org DATABASE URL: http://bbgre.org.

G31: Yupeng Cun and Holger Froehlich. Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics

Abstract: Predictive, stable and interpretable gene signatures are generally seen as an important step towards a better personalized medicine. During the last decade various methods have been proposed for that purpose. However, one important obstacle for making gene signatures a standard tool in clinics is the typical low reproducibility of signatures combined with the difficulty to achieve a clear biological interpretation. For that purpose in the last years there has been a growing interest in approaches that try to integrate information from molecular interaction networks. We here propose a technique that integrates network information as well as different kinds of experimental data (here exemplified by mRNA and miRNA expression) into one classifier. This is done by smoothing t-statistics of individual genes or miRNAs over the structure of a combined protein-protein interaction (PPI) and miRNA-target gene network. A permutation test is conducted to select features in a highly consistent manner, and subsequently a Support Vector Machine (SVM) classifier is trained. Compared to several other competing methods our algorithm reveals an overall better

prediction performance for early versus late disease relapse and a higher signature stability. Moreover, obtained gene lists can be clearly associated to biological knowledge, such as known disease genes and KEGG pathways. We demonstrate that our data integration strategy can improve classification performance compared to using a single data source only. Our method, called stSVM, is available in R-package netClass on CRAN.

G32: Kerstin Haase, Silke Raffegerst, Dolores Schendel and Dmitrij Frishman. Expitope: Web server for Epitope Expression

Abstract: Adoptive T cell therapies based on introduction of new T cell receptors (TCRs) into patient recipient T cells is a promising new treatment for various kinds of cancers. A major challenge, however, is the choice of target antigens. If an engineered TCR can cross-react with self-antigens in healthy tissue, the side-effects can be devastating. We present the first ever web server for assessing epitope sharing when designing new potential lead targets. The user can search the database for expression of any potential specific target epitope and for epitopes with partial identity, with an allowed number of mismatches, when approximate matches are of interest. All proteins deposited in RefSeq [Pruitt, et al., 2014] that contain the given peptide or variants thereof are reported. Furthermore, transcripts that are associated with the identified proteins and isoforms are then checked for their expression values in various human tissues. We calculated the number of fragments per kilobase of exon per million fragments mapped (FPKM) for 15 normal tissue types from the Illumina Human Body Map Project (NCBI GEO accession GSE30611) and three ENCODE cell lines [The ENCODE Project Consortium, 2012]. Additionally, gene expression values for multiple brain tissues, provided by the Burge group [Wang, et al., 2008] are incorporated.

All expression values are reported to the user ranked by the predicted score of the corresponding epitope, which reflects the probability for it to be produced by proteosomal cleavage [Keşmir, et al., 2002; Nielsen, et al., 2005] and its affinity to the TAP transporter [Peters, et al., 2003] and user-defined MHC class I alleles [Lundegaard, et al., 2008; Lundegaard, et al., 2008; Nielsen, et al., 2003].

The web server was used to investigate two previous TCR gene therapies in which unanticipated cross- recognition of healthy tissues led to patient deaths. When the specific epitopes of the TCR were submitted to our web server, the cross-reactive epitopes could be identified. In the first example, Morgan and coworkers reported cross-recognition of a MAGE-A3 TCR with a MAGE-A12 epitope that was later found to be expressed in the brain; the treatment was fatal for some patients [Morgan, et al., 2013]. In the second case, Linette and coworkers used a different MAGE-A3-specific TCR that was found to show cross-recognition of an epitope present in titin, a protein expressed in the heart, although the titin-associated epitope had four mismatches compared with the original MAGE-A3 epitope [Linette, et al., 2013]. Both cases of cross-recognition were identified by our web server when allowing for a sufficient number of mismatches with the specific epitope. Thus, our Expitope web server can help to exclude potential cross-reactivity in the early stage of TCR selection for use in design of adoptive T cell immunotherapy.

Availability: http://webclu.bio.wzw.tum.de/expitope

G33: Claudia Rincon, Isabel Brito and Philippe Hupé. Evaluation of different algorithms to stratify cancer tumors based on gene interaction networks and somatic mutations data

Abstract: Recently, [1] pointed out the important role played by genomic instability in cancer cells as some genetic changes can “orchestrate hallmarks of cancer capabilities”. Somatic mutations profiles, enriched in the tumor cell population, become a promising new source of data for tumor stratification.

Because among these mutations are causal drivers of tumor genesis and progression, the classification of patients based on their mutations profile is of major significance. Usual mathematical clustering solutions are misleading since this kind of data shows two particularities: individual profiles show very few mutations and the set of patients have not many mutations in common. [1] emphasizes that “cancer is a disease not of individual mutations, nor of genes, but of combinations of genes acting in molecular networks corresponding to hallmark processes”. Therefore, [2] proposed to integrate somatic mutation data with the knowledge of the molecular network architecture of human cells.

In this work, we compare different classification algorithms using as reference the Network-Based Stratification of Tumor Mutations (NBS) method [2]. Two main steps are implemented: First, patients’ mutations are propagated by a random walk that allows signal to spread following the influence of each gene over its network neighborhood, defined by a gene interaction network. Second, unsupervised classification techniques: Non-negative Matrix Factorization –NMF– (propagation + NMF corresponding to NBS algorithm), k–means, Self-organizing Maps –SOM– and Highly Connected Components –HCC– are applied in order to get tumors stratification. These methods are denominated respectively, NBS, k–means like–NBS, SOM like–NBS and HCC like–NBS. We tested all these algorithms on four datasets of somatic mutations profiles downloaded from the TCGA portal: Breast, Ovarian, CESC and UCEC cancer datasets, and use three gene interaction networks: STRING, HumanNet and PathwayCommons. We compare the performance of these classification methods by the evaluation of their stability, their intrinsic cluster quality and their accuracy to predict phenotypic types.

SOM like–NBS is the most stable method. The intrinsic quality of all the methods is dataset dependent. NBS ability to predict phenotypic classes is restricted to UCEC and CESC tumors whereas SOM like–NBS is the best method to predict phenotypic classes for Breast and Ovarian tumors. The impact of the gene interaction networks does not appear to be relevant. Perspectives to improve tumor classification algorithms, particularly concerning the use of gene interaction networks, are discussed.

References

[1] Hanahan D., Weinberg R., (2011), Hallmarks of cancer: the next generation, Cell. 44:646-74.

[2] Hofree M., Shen J., Carter H., Gross A. Ideker T., (2013), Network-based stratification of tumor mutations, Nature Methods. 10:1108-15.

G34: Adrin Jalali and Nico Pfeifer. Interpretable per Case Weighted Ensemble Method for Cancer Associations

Abstract: Over the past decades, biology has transformed into a high throughput research field both in terms of the number of different measurement techniques as well as the amount of variables measured by each technique (e.g., from Sanger sequencing to deep sequencing) and is more and more targeted to individual cells [1]. This has led to an unprecedented growth of biological information. Consequently, techniques that can help researchers find the important insights of the data are becoming more and more important. Molecular measurements from cancer patients such as gene expression and DNA methylation are usually very noisy. Furthermore, cancer types can be very heterogeneous. Therefore, one of the main assumptions for machine learning, that the underlying unknown distribution is the same for all samples in training and test data, might not be completely fulfilled.

In this work, we introduce a method that is aware of this potential bias and utilizes an estimate of the differences during the generation of the final prediction method. For this, we introduce a set of sparse classifiers based on L1-SVMs [2], under the constraint of disjoint features used by classifiers. Furthermore, for each feature chosen by one of the classifiers, we introduce a

regression model based on Gaussian process regression that uses additional features. For a given test sample we can then use these regression models to estimate for each classifier how well its features are predictable by the corresponding Gaussian process regression model. This information is then used for a confidence-based weighting of the classifiers for the test sample. Schapire and Singer showed that incorporating confidences of classifiers can improve the performance of an ensemble method [3]. However, in their setting confidences of classifiers are estimated using the training data and are thus fixed for all test samples, whereas in our setting we estimate confidences of individual classifiers per given test sample.

In our evaluation, the new method achieved state-of-the-art performance on many different cancer data sets with measured DNA methylation or gene expression. Moreover, we developed a method to visualize our learned classifiers to find interesting associations with the target label. Applied to a leukemia data set we found several ribosomal proteins associated with leukemia that might be interesting targets for follow-up studies and support the hypothesis that the ribosomes are a new frontier in gene regulation.

[1] E. Shapiro, T. Biezuner, and S. Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet., 14(9):618–30, Sept. 2013.

[2] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In ICML, volume 98, pages 82–90, 1998.

[3] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. Machine learning, 37(3):297–336, 1999.

G35: Alejandra Medina-Rivera, Lina Antounians, Jessica Dennis, France Gagnon and Michael Wilson. Defining the cJun regulatory network in vascular endothelial cells from multiple species.

Abstract: Vascular endothelial cells conform the inner linings of blood vessels, in general they regulate hemostasis, and their dysfunction has been associated with cardiovascular diseases. Throughout our body, depending on the local conditions, endothelial cells show gene expression heterogeneity. Transcription factors (TF) and epigenetic modifications control the dynamic regulatory landscape by mediating interaction between transcriptional machinery and gene expression. The extent to which the gene regulatory events that control vascular endothelial cells are conserved in human and other model organisms is unknown. Detecting conserved protein-DNA interactions is potentially a powerful way to identify critical regulatory regions that control tissue-specific and process-specific functions. cJun, an AP-1 monomer, is a TF involved in many cellular processes, including cell proliferation, apoptosis, and stress response. In order to identify conserved and species-specific cJun binding in endothelial cells, we used chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) in aortic endothelial cells of human (HAEC) and rat (RAEC). Preliminary results show that approximately 5% of cJun protein-DNA interactions are conserved between human and rat. Of these 1,339 sites, 74.3% are also bound by cJun in human umbilical vein endothelial cells (HUVEC), indicating that conserved cJun binding is relevant to both arteries and veins (p-value=0.036). In order to identify conserved cJun binding events which were specific to endothelial cells we compared HAEC cJun binding patterns to publicly available ChIP-seq data in K562, HeLa, and HepG2 cell lines. We found that conserved cJun endothelial cell-specific binding enriched for vascular and blood vessel development functions. Our ongoing work is focused on identifying rare and common variants in the human genome, which have the potential to affect these potentially functional cJun-binding regions. Having a comprehensive multi species, multi-tissue regulatory map will enable us to understand how cJun regulates gene expression in endothelial cells.

G36: Mahmoud Elhefnawi, Asmaa Ezzat and Mohd Noor Isa. Metagenomic and Metatranscriptomic analyses of the hepatocellular carcinoma-associated microbial communities and the potential role of microbial communities in liver cancer

Abstract: Background & Aims: 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.

G37: Maxim Ivanov, Kamil Khafizov and Sergey Kovalenko. OrphaPRED – functional effect prediction tool for mutations associated with rare diseases

Abstract: As next generation sequencing projects lead to the generation of massive amounts of data, new computational tools for predicting possible phenotypic effects of mutations are being developed. In spite of the significant number of methods and tools that have been developed for this task, development of the field is more extensive rather than intensive – each new program uses the well-known algorithms and combines them in a different way. Thus, the effectiveness of such tools is expected to be rather limited, which was discussed in many reviews. At the same time, the adjacent bioinformatics fields are developing much faster but the newly emerging methods still cannot find their applications in programs/tools aimed at studying phenotypic effects of the mutations. Here, we present a new program OrphaPRED that provides a new way to predict the functional effects of protein sequence variations. Besides the well-known methods that use the conservation of mutation sites, sequence annotations and trivial geometric parameters of 3D protein structure, our tool also estimates the difference in free energy between the wild type and mutant proteins. Binary classification of mutations is based on a score-function derived from the calculated descriptors using empirical rules. Our method was trained on a set of 1,521 mutations. On a selected test-subset (1,019 deleterious and 498 neutral mutations) the predictive ability of our method was shown to be higher than the corresponding abilities of widely used tools: accuracy is 80.6%, Matthews correlation coefficient - 0.61, area under the ROC curve - 0.88. In addition to the predictive ability of binary classification, using TP53 and alpha-Galactosidase (Fabry disease) as examples we also observed a correlation between the developed score-function and biological activity of mutant protein (the linear correlation coefficient is 0.74). Thereby, our program can significantly help to interpret the unknown-value variations and to narrow down the search space of possibly deleterious variants.

G38: Laura Buzdugan and Peter Bühlmann. High-dimensional, predictive GWAS

Abstract: Motivation: Traditionally, genome wide association studies relied on the common disease – common SNP hypothesis, assuming that a small number of SNPs explain most of the phenotypic variability. Given this assumption, the approach is to regress each SNP against the response, and compare their p-values to a genome-wide significance threshold. However, new evidence suggests that a more likely scenario is the one in which individual SNPs have small effect sizes. Identifying such SNPs has proven problematic with the standard methods.
Results:We propose a method that addresses this issue, by creating a unified model. This model includes all genotyped SNPs, thus estimating the joint rather than the marginal effects. Our statistical method has 2 stages: a screening stage and a testing stage. In the screening stage an L1 penalized linear model is used to select a set of candidate SNPs. In the testing stage we compute p-values not only for individual SNPs, but also for groups of SNPs. The groups can have varying sizes and be based on different criteria. These groups are constructed by hierarchically clustering the SNPs. The testing is done on this hierarchy, in a top-down manner. The procedure finds the smallest significant cluster of SNPs, while controlling the FWER and correcting for multiple testing. We used our method on a GWAS with socio-economic phenotypes and obtained promising results.

G39: Francesco Iorio, Hayley Francies, Jayeta Saxena, Graham Bignell, Cyril Benes, Ultan McDermott, Simon Cook, Mathew Garnett and Julio Saez-Rodriguez. Systematic Prediction of Transcription Factors modulating Drug Response in Human Cancer Cell Lines

Abstract: We designed DoRothEA (Discriminant Regulon Enrichment Analysis), a computational method that, starting from a set of genome-wide gene expression profiles, infers transcription factors whose post-translational regulatory activity is significantly associated with an underlying factor on sub-populations of samples.

By systematically applying this method to the collection of 1,000 basal gene expression profiles of the cell lines in the Genomics of Drug Sensitivity in Cancer (GDSC) screening [1] (version July2014), in combination with drug-response data from the same study, we inferred associations between the activity of ~200 human transcription factors (including complexes and aberrantly fused proteins) and the level of sensitivity/resistance to 140 drugs.

In DoRothEA the level of activity of a given transcription factor (TF) in a given cell line is considered proportional to the basal expression of its target genes (i.e. its regulon) in that cell line. To allow the computation of TF activities, a collection of 931 regulons, corresponding to a total number of ~400 unique human TFs has been assembled from different public repositories and a statistical method, making use of a phenotypic independent sample-wise version of the Gene Set Enrichment Analysis (GSEA) [2] tool, has been conceived on purpose. Finally the inferred TF activity scores have been systematically correlated with drug response through a systematic multivariate analysis of variance (MANOVA).

Stratification of the samples based on the constitutive expression of the genes targeted by a transcription factor associated to a drug, often improves the predictive ability of the established genomic markers for that drug. More importantly, novel transcriptional markers are identified where no genomic markers are available at all. Additionally, the availability of a transcription factor paired with a transcriptional signature of drug response, makes the expression of the genes in the signature potentially ‘actionable’ (by knocking-down the corresponding transcription factor or by inducing its expression) to observe whether this modulates drug resistance/sensitivity (experimental validation is in progress).

Results have been collected into a user-friendly web-resource (accessible at http://wwwdev.ebi.ac.uk/saezrodriguez/dorothea, access credentials available on request), providing interactive plots and statistical tests performed on-the-fly on user defined subpopulations of cell lines based on transcription factor activities, oncogenic lesions, histology, primary sites and other features.

References:
[1] Garnett,M.J. et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483, 570–575.
[2] Subramanian,A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102, 15545.

G40: Wolfgang Raffelsberger, Hélène Polveche, Amélie Weiss, Mickael Renaud, Anne Maglott-Roth, Johann Foloppe, Benoit Grellier, Etienne Weiss, Laurent Brino, Olivier Poch, Philippe Ancian and Philippe Erbs. Improving oncolytic viruses : Integrated data treatment pipeline

Abstract: Oncolytic viruses represent recent efforts of developing novel approaches in cancer therapy. Indeed, oncolytic viruses open very promising perspectives to efficiently target and kill cancer cells through virus mediated lysis. Recently, numerous clinical trials have been performed and helped bringing oncolytic virus based/assisted therapy closer to patient treatment. The Oncovaccine-consortium aims to genetically improve the Vaccinia virus to increase overall oncolytic efficiency and encompass current limitations through tumor resistance. Based on a combination of a genome-wide RNAi screen, transcription profiling of a wide panel of cancer cell lines, gene-network modeling and further validation efforts at protein level, we are currently working on this objective. Here, we present the bioinformatic efforts within the consortium to treat raw data and generate integrated analysis. RNAi data analysis is based on R and Bioconductor with tools previously developed at the LBGI, primary RNAseq data treatment on MWARD (Modular Workflow for Analysis RNAseq Data, conveniently integrating fastQC, Gsnap and rna-STAR, samtools and HTSeq-count), data integration of different types of omics data and statistical testing is performed in R, gene network analysis uses the String data-base, Cytoscape and tools previously developed at the LBGI. In summary, the integrated data analysis has the following aims : i) Identification of individual genes as targets for siRNA mediated gene-repression in novel Vaccinia virus strains with improved oncolytic activity, and ii) gaining novel system-wide understanding of molecular and cellular pathways participating in cancer cell specific oncolytic virus activity.

G41: Sumaiya Nazeen and Bonnie Berger. Integrative Analysis of Multiple Gene Expression Studies Reveals Genes and Pathways involved in Alteration of Steroidogenesis in Polycystic Ovary Syndrome (PCOS)

Abstract: Polycystic Ovary Syndrome (PCOS) is a complex endocrine abnormality observed in 6-10% of premenopausal women with a heritability of ~70%. Although the role of genetic factors in PCOS is strongly supported by its frequent occurrence in families, the exact etiology remains a mystery because of the heterogeneity of genetic and phenotypic features associated with it. Researchers have thus far focused on elucidating potential genes and pathways involved in PCOS by using differential expression analysis of single studies only, many of which were inconclusive due to having small sample sizes. In this study, we adopt a novel integrative approach to analyze gene expression data from multiple case-control studies to extract deeper biological insight into PCOS. Our integrative analysis reveals 47 candidate genes that have significantly altered expression levels across multiple studies, 45 of which are

novel. Canonical pathway enrichment analysis reveals as the top four hits signal transduction pathways controlling steroidogenesis, suggesting their involvement in PCOS.
Microarray datasets from human case-control studies of PCOS were downloaded from NCBI's GEO. Per inclusion criteria, only datasets from studies using affymetrix arrays were selected which gave us six gene expression studies– GSE34526, GSE10946, GSE6798, GSE5090, GSE48301, and GSE5850, published between 2006 and 2013. We combined the raw data from these studies into an integrated expression set removing the batch effect among different studies using the gene quantile normalization method. We calculated fold changes and student's t-test p-values for the genes and selected 47 genes (all those with p-value < 0.001) as candidates among which 33 were up-regulated and 14 were down-regulated. These genes are known to be significantly involved in cell death and survival, lymphoid tissue structure and development, tissue morphology, cell morphology and organ development. We also performed canonical pathway enrichment analysis using IPA (Ingenuity® Systems, www.ingenuity.com). The top four pathways that we found are: Protein Kinase-A (PKA) signaling, G-protein coupled receptor signaling, cAMP-mediated signaling, and fMLP signaling in neutrophils respectively. These are known to be involved in the steroidogenic machinery; in fact, the top three are gonadotropin-stimulated signaling pathways that regulate ovarian steroidogenesis. Our findings for the first time show conclusively the connection between gonadotrophin-stimulated signaling pathways and alteration of steroidogenesis in PCOS. However, determining the causal relationship between PCOS and these pathways needs further investigation. Looking at other sources of information (e.g., protein-protein interaction networks, gene ontologies etc.) may help us gain deeper insights into our initial findings.

G42: Tim Beck, Robert Hastings and Anthony Brookes. GWAS phenotype data: standardised in the GWAS Central resource and harmonised in the GWAS PhenoMap browser

Abstract: The terminology used in the scientific literature to describe genome-wide association study (GWAS) phenotypes requires standardisation if these data are to be effectively discoverable in databases. Once standardised using a particular ontology, these data must be harmonised if they are to be effectively compared with data that are compliant with other ontologies in other databases. The GWAS Central resource (http://www.gwascentral.org) provides standardised descriptions of GWAS phenotypes to allow comparisons between individual studies, while the GWAS PhenoMap browser (http://phenomap.gwascentral.org/) allows comparison of harmonised phenotypic descriptions between various GWAS-related databases.
GWAS Central is a comprehensive resource for the comparison and interrogation of summary-level GWAS data. Each study is manually evaluated for its range of phenotype content and appropriately chosen Medical Subject Headings (MeSH) terms are assigned to ensure that the phenotype descriptions are standardised across all studies. Further utility is provided by assigning Human Phenotype Ontology (HPO) terms to studies, either manually when a HPO term is more suitable than a MeSH term, or via the automatic mapping of MeSH and HPO equivalent terms.
Phenotype classification approaches differ between the GWAS-related databases involved in the GWAS PhenoMap initiative (participant databases listed on the website). Four different ontologies are used across the databases: MeSH, HPO, International Classification of Diseases and Disease Ontology Lite. We have harmonised the various phenotype descriptions by taking the totality of terms used from each ontology and mapping them to equivalent terms in the other ontologies through a series of manual and automated steps. The databases share a common identifier per study: the study publication PubMed ID. This permits the GWAS PhenoMap browser to be used to explore each of the four ontologies used across the

databases; the association of ontology terms (and child terms) with a GWAS publication; and the various database entries for that study.

The diversity of phenotypes reported in published GWAS makes it difficult for the GWAS databasing community to standardise data in this domain by implementing a single ontology across all related databases. Instead, the GWAS PhenoMap browser harmonises diverse phenotype annotations, from resources such as GWAS Central, to allow cross-database comparisons to be made based on phenotype.

G43: Khalid Abnaof, Joao Dinis and Holger Fröhlich. Using Consensus Clustering to Explore Biological Effect Similarities of Drug Treatments based on Integrated Biological Knowledge from Multiple Sources – An Example Study on HIV and Cancer

Abstract: Investigating the joint space of drugs and their putative targets, together with empirical drug bio-activities, might help to design new drugs or repurpose existing ones.

In this study we asked the question, whether approved and tested drugs for different diseases (here: HIV and cancer) would fall into separable clusters according to the biological response they induce by binding to known targets. It should be noted that one and the same drug can potentially bind to different protein targets with certain affinity.

To answer our question we collected for each known drug target (213 in cancer, 22 in HIV) a multitude of biological information, such as involvement into biological processes (Gene Ontology [1]), contained protein domain annotation (InterPro [2]) and position in specific pathways (KEGG [3], PathwayCommons [4]). Moreover, drug-target binding affinities (Ki values) were retrieved from ChEMBL [5], DrugBank [6] and OpenPhacts [7].

Based on this information we developed an integrated similarity measure for drugs comparing their expected biological effects. Our biological effect similarity (BES) combines in a first step drug-target similarities computed from individual information sources (e.g. shortest path distances for network information) in a probabilistic manner [8]. In a second step drug-target similarities are extended to drug similarities. That means for each drug pair we compute the similarities of all their targets, weighted by the similarity of binding affinities (normalized Ki values) to these targets. All these information are then combined into one BES measure per drug pair.

The BES measure allowed us to apply generalized k-means consensus clustering [9] in order to reproducible identify groups of drugs inducing similar biological effects. This way we found 33 non-singleton clusters in HIV and in cancer. These clusters were well separated according to a silhouette analysis. Furthermore, analysis of targets of compounds that were clustered together, revealed in the vast majority of cases a clear enrichment of distinct Gene Ontology terms, KEGG pathways, protein domains and sequence motifs. None of our identified clusters could be found in a traditional ligand-based clustering using fingerprints and the Tanimoto similarity coefficient. This was also true when a joint clustering of HIV and cancer drugs with respect to the BES measure was conducted. In that case we could specifically find clusters containing drugs from both diseases, hence indicating biologically close effects despite of a very different medical application area.

Taken together our work suggests the usefulness of a joint view on the compound-target space and the proposed BES measure in particular. Our results on cancer and HIV demonstrate that the latter may help to uncover drugs with biologically similar effects and in consequence also repurpose existing drugs for novel application areas.

G44: Jennifer E. Mollon, Steven J. Kiddle, Claire Steves, Kerrin Small, Martina Sattlecker, Katie Lunnon, Petra Proitsi, John Powell, Angela Hodges, Steven Williams, Tim Spector, Iwona Kloszewska, Patrizia Mecocci, Hilkka Soininen, Magda Tsolaki, Bruno Vellas, Simon

Lovestone, Richard J. B. Dobson and Stephen Newhouse. Identification and Replication of Cis and Trans Effect Protein Quantitative Trait Loci in Ageing Adults

Abstract: In this study we use an aptamer-based protein assay from SomaLogic and combine this with genotyped and imputed SNPs to identify protein quantitative trait loci (pQTLs). We map our hits to eQTLs in a subset of our samples as well as published eQTLS in other tissues. Samples from 106 individuals with Alzheimer's Diseases (AD), 90 with mild cognitive impairment (MCI) and 101 healthy elderly controls were selected from the AddNeuroMed (ANM) cohort. Genotyping and imputation of these samples resulted in 6 345 198 SNPs from 297 individuals, and 1001 proteins targeted by 1016 SOMAmers were assayed in plasma from the same individuals. We used regression to measure single-SNP effects, adjusting for age, gender, diseases status and 5 genetic principle component axes. A strict Bonferroni threshold was used, allowing for multiple SNP signals as well as the 1016 protein phenotypes. Replication was carried out in 102 unrelated individuals from the Healthy Aging in Twins Study (HATS).

We identified 100 novel pQTLs in 88 different proteins, with 22 cis- (within 300kb of the protein-coding gene) and 78 trans-effects. We found the pQTL SNPs to be highly enriched for hits in the NHGRI GWAS catalogue, and cis-hits were highly enriched for non-synonomous SNPs. Eighty-six SNP/protein combinations were available in HATS, with 19 cis and 12 trans associations replicating at a nominal level of p<0.05. Our top novel, replicated cis-hit was missense mutation rs3197999; the minor allele was associated with decreased levels of MST1 (p=3.83x10-74). This SNP has been implicated in inflammatory bowel disease and primary sclerosing cholangitis. The top novel, replicated trans-hit was rs12614 (missense mutation in CFB), with the minor allele associated with increased levels of MMP8 (p=2.88x10-70). We replicated 25/41 published pQTLs from other groups, and 85/94 from our previous pQTL. Only 2 pQTLs mapped to eQTLs in plasma from the same individuals while 3 pQTLs are eQTLS in blood in gTEX, suggesting pQTL proteins may be produced in other tissues. Several of our cis-pQTL SNPs are cis-eQTLS for other, nearby genes, for example rs62012908 and TPSAB1 (p=1.99x10-11). In gTEX this SNP/gene combination is an eQTL in blood, and rs62012908 is also an eQTL for TPSD1 in six tissues, and for RP11 in four tissues.

We have presented strong evidence of novel genetic control of protein expression. The study of such intermediary phenotypes may help unravel mechanisms through which genetic effects manifest in disease.

G45: David Källberg, Mattias Landfors, Yuri Belyaev and Patrik Rydén. Identifying subgroups of cancer – the blind two-sample test

Abstract: Background: It is important to identify new subgroups of cancer, since varying the therapy between subgroups can be beneficial for recovery. We consider the problem where data (e.g. gene expression) from n patients are observed and the aim is to test the suspicion that the disease can be divided into two subgroups, and to predict which subgroup the patients belongs to. If we only observe one variable, this is a reformulation of a familiar statistical problem: let x1,…,xn be observations on individuals with unknown class membership c1,…,cn (c is either A or B), and suppose that the objective is to test if the means differ between group A and B. We refer to this problem as the blind two-sample test. In the case the labels c are known this problem may be solved using the Welch two-sample t-test or the Wilcoxon rank-sum test.

Objectives: The aim is to develop and evaluate statistical methods for testing if a disease is homogeneous or can be divided into two subgroups in the case a single variable is observed. Furthermore, in the case there are evidence that the disease is heterogeneous, develop and evaluate statistical methods that allow the user to predict which patients belong to the same

subgroup.

Methods: We consider four approaches for the blind two-sample test, all of which also can be used to predict the cancer subtype of the patients. Two methods are based on estimating the parameters for the 2-component normal mixture model and using the absolute difference between the mean estimates as test variable; the EM-algorithm and the method of moments. Another test variable is the distance between the cluster centers obtained from the 2-means clustering algorithm. Finally, we consider a test statistic based on the differences between the mean of the (n-k) largest observations and the mean of the k smallest observations, k=2,…,(n-1). The methods are evaluated using microarray gene expression data and simulated data, and their performance for various effect sizes, number of patients, and distributions are studied. We compare the tests in terms of power, and the methods' ability to predict the class membership using the adjusted rand index.

Results: All methods can be used to test if a disease is homogeneous or can be divided into two subgroups. As expected the power is lower than for the Welch two-sample t-test and the Wilcoxon rank-sum test (for which the labels are known), but if the sample size or the effect size is large the methods have a sufficient power. Furthermore, all methods have the ability to make reasonable predictions of the patients' class memberships. The result reveals interesting differences between the suggested methods, both with respect to their performance (i.e. the power and the adjusted rand) and their running time.

Conclusions: We suggest methods that can be used to identify new subtypes of cancer and predict which subgroup the patients belong to, in particular if the effect size or the sample size is large.

G47: Laurence Pearl, Amanda Schierz, Simon Ward, Bissan Al-Lazikani and Frances Pearl. Drugging the DNA Damage Response

Abstract: The near universal development of genomic instability that accompanies tumourigenesis has brought the pathways that mediate the DNA Damage Response (DDR) to the fore as targets for development of new approaches to the treatment of a wide range of cancers. However, unlike the protein-kinase signalling pathways that have been the focus of much anti-cancer drug development in the past decade, the proteins that make up the DDR pathways are very diverse in structure and function and there is a need for major effort in target definition and identification before these targets can be fully exploited.

To help support these drug discovery activities we have assembled a comprehensive dataset of proteins involved in the DNA Damage Response. These have been systematically analysed using chemogenomic approaches to define both their genetic vulnerability to mutation and to identify their suitability for functional modulation by small molecule drugs – 'druggability'. Through data-driven evaluation of chemical activities data, gene expression data, mutational data, functional annotation, 3D structure, and analysis of interaction network we aim to identify 'druggable' points of intervention within the DNA Damage Response pathways, on which drug discovery can be most effectively focussed.

G48: Fiona Browne, Haiying Wang and Huiru Zheng. Network Driven Analysis for Biomarker Discovery in Alzheimer's Disease

Abstract: Alzheimer's disease (AD) is a genetically complex and heterogeneous disease and the most common form of age-related cognitive impairment. In order to develop specific drug therapies to target this disease, an understanding of AD mechanisms is required. This research analyses the human protein interaction network along with gene expression AD data to determine if significant hub proteins can predict biomarkers for AD.

A network of human protein-protein interactions (PPI) was derived from both manually curated and high throughput experiments. Hub proteins (proteins with a large number of interacting partners) were identified by measuring degree distribution in the network and selecting the top 15%. AD gene expression was obtained from Dunckley et al. (2006). This study examined the neurofibrillary tangles (NFT) a histopathology feature of AD and consists of gene expression profiles of NFT-bearing entorhinal cortex neurons from 10 mid-stage AD patients (Disease) compared with 10 histopathologically normal neurons (Control). The dataset was normalised using MASS 5.0 in R. To determine if significant hubs can differentiate between the Disease and Control groups, Pearson's correlation (PCC) of co-expression of the hubs and their interactors was calculated. Gene co-expression values were mapped to the PPI network nodes via NCBI gene IDs. PCC values for hubs and interactors were calculated for both the Disease and Control groups. The average hub difference (AverageHubDiff) in PCC values between each group was estimated. We assigned a statistical significance, P, value to each AverageHubDiff based on a random permutation test. To account for multiple testing, P values are adjusted using the Bonferroni correction.

A total of 1112 hubs were identified from the human PPI network. Hubs with corresponding $P<0.05$ were considered significant which resulted in 154 significant hubs. The significant hubs were compared to 22 known and recently discovered AD susceptibility genes from Lamberet et al. (2013). Using this approach, the significant hub INPP5D was correctly identified as an AD susceptible gene. Gene Ontology Biological Process Enrichment analysis revealed significant hubs are modulated in AD pathogenesis including regulation of neurogenesis and generation of neurons. KEGG pathway analysis identified significant hub involvement in the Neurotrophin signalling and Huntington Disease pathways. The results highlight the potential of using significant hubs to predict AD biomarkers. Currently, we are integrating additional data such as tissue information and applying classification techniques to further validate our analysis.

1. Dunckley,T., Beach,T.G., et al. (2006) Gene expression correlates of neurofibrillary tangles in Alzheimer's disease. Neurobiol.Aging, 27, 1359-1371.

2. Lambert,J., Ibrahim-Verbaas,C.A., et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat.Genet.

G49: Trevor Clancy and Eivind Hovig. Mining immune cell activity from tumor transcriptomes

Abstract: Despite decades of successful discoveries, the overwhelming molecular complexity of cells has been a significant barrier to gain a fundamental understanding of how cells work as a whole, how they are perturbed in cancer, and how they interact with signals from the immune system. To achieve a more complete understanding, will require the assistance of network modeling. By developing computational network models we can introduce cancer mutations to simulate pharmacological perturbations, or loss-of-function and gain-of-function experiments and study the dynamics of biochemical networks change under different immune signaling states. We are developing and applying rule-base (kappa) network modeling of signaling networks to study how microenvironment immune factors influence a tumor's behavior. In particular, that of immune mediated cellular senescence in tumor cells. We have built a rule-base network model of the biochemical network regulating immune mediated cell senescence in melanomas, which carry a BRAFV600 mutation in the MAPK signaling network. Using computational methods we have characterized the crosstalk with cytokine signaling networks and the crosstalk with other signaling networks. The development of such a computational model may help researchers understand how BRAFV600 cells in melanomas communicate with their environment by secreting various cytokines and growth factors, as it has become clear that this immune 'secretory phenotype' can have pro- as well as anti-tumorigenic effects.

By combining rule-based descriptions of the biochemical reactions associated to this phenotype with computer simulation we can open up new avenues for exploring such complex networks linked to melanoma progression.

G51: Mamunur Rashid, Alistair Rust, Jeroen Ridder and David Adams. Identification of Novel Non-Coding Driver Mutations in Cancer using Pattern Recognition

Abstract: The driver mutation model of cancer, in which only a small number of mutations play a role in the development and progression of a cancer, is the most widely accepted model in cancer genomics (Stratton et.al., Nature, 2009). Recent advances in Next Generation Sequencing technologies have allowed cancer genomes to be investigated at much higher resolution than ever before, giving rise to a number of large-scale cancer genome sequencing projects, such as the Cancer Genome Project and the International Cancer Genome Consortium. However, most research to date has focused on mutations in the protein coding regions of the genome. As a result, around 98% of cancer genome remains comparatively unexplored. A handful of recent studies (e.g. Khurana E. et.al, Science 2013; Vinagre J. et.al., Nature 2013) have begun to investigate non-coding mutations as potential driver mutations. Most of these studies have focused on prioritizing mutations and/or mutated genes by applying arbitrary filters based on incomplete biological assumptions. Moreover, most of these studies addressed generic human diseases rather than providing cancer-specific predictions (Graham R S Ritchie, Nature 2014).

Developing a robust technique to identify non-coding driver mutations is challenging because of i) inadequate number of biologically validated non-coding mutations, ii) a partially annotated non-coding human genome and iii) a diverse mutational landscape in cancer genomes. In this poster we present a novel annotation and classification software pipeline to detect potential non-coding driver mutations. This pipeline is specifically tailored to cancer as it includes a range of genetic, regulatory, population and cancer-specific features to annotate non-coding variants. We have compiled three different mutation datasets (a) COSMIC recurrent non-coding mutations, (b) non-coding driver mutations predicted by Khurana et.al. and (c) HGMD Disease Associated Mutations (Stenson PD, 2014) along with relevant control variant sets. We applied the feature annotation pipeline on these data sets and due to inadequate annotation of non-coding genome; the annotated feature matrix is often sparse in nature. By employing feature selection on these annotated feature matrices, we have identified two distinct sub-sets of predictive features. The first one is capable of discriminating generic disease associated mutations while the second is capable of discriminating cancer associated mutations from the relevant controls. We have adopted a classifier combination approach to and obtained a considerable improvement in discriminating non-coding cancer associated mutations.

G52: Russel Sutherland, Salvador Diaz-Cano, Jane Moorhead and Richard Dobson. Predicting tumour grade across multiple adenocarcinomas using exome sequence data.

Abstract: Background: There is a need for tumour grading tools that are applicable across multiple cancer types in order to make tumour grading a more objective process. Most tumour grading systems are only applicable to a single cancer type. The aim of this study was to develop a tumour grading prediction model using tumour/normal exome sequence data as a tumour grading decision support tool for Pathologists.

Methods: We used exome sequence tumour/normal variant data and clinical data from the Pan Cancer Analysis from 970 patients with Kidney Renal Cell Carcinoma, Ovarian Carcinoma and Endometrial Carcinoma. We created a sample (S) by protein coding gene (P) “binary mutation matrix” that indicated the presence of any protein coding mutation for a sample (s) at a protein (p). Using the “binary mutation matrix” together with age, gender and cancer stage clinical variables we used multivariate logistic regression and Akaike Information Criterion (AIC) based backwards model selection to create classification models to predict the high grade or low grade status of tumours across cancer types and within cancer types. It was important to include the gender clinical variable in order to control for any gender bias introduced by the Endometrial and Ovarian Carcinomas. Classification accuracy, sensitivity and specificity were measured in an independent test set along with the area under the receiver operator characteristic (ROC) curves.

Results: Across cancer types an AIC refined multivariate logistic regression model including 13 protein coding genes; TP53, MUC4, ANK2, CCDC171, CDH7, CTCF, CTNNB1, MYOF, NALCN, PARD3, PKD1L1, PTPRK and RAPGEF2 along with the cancer stage and age clinical variables was able to assign tumours to the correct grade status with an area under the curve (AUC) of 0.821. By comparison, the model using gender and cancer stage clinical variables refined from a model including clinical variables achieved an AUC of 0.756. The multivariate logistic regression model in Endometrial Carcinoma refined from protein coding genes and clinical variables age and stage included TP53, PTEN and cancer stage and performed with an AUC of 0.879 in comparison to a model including age and stage clinical variables that achieved an AUC of 0.753.

Conclusions: There are protein coding genes across cancer types that are predictive of tumour grade when adjusting for cancer stage. The genes ANK2, CDH7, and CTNNB1 are involved in tumour and stromal cell interactions important in tumour progression. MUC4 is involved in glandular differentiation, and has been implicated in cancer progression. Mutations in TP53 can be indicative of aggressive tumour types. The 13 protein features capture predictive information in addition to that captured by age gender and cancer stage clinical variables.

G54: Haeseung Lee and Wan Kyu Kim. An integrative analysis of multi-dimensional compound-disease signatures for in silico drug repositioning in human cancers

Abstract: With increasing cost of developing novel drugs, in silico drug repositioning (DR) methods have been actively sought. Here, we integrate multi-dimensional signatures for drug-disease associations and predict DR candidates for three different types of human cancer - glioblastoma, lung and breast cancer. We first collected known drugs (KD) from public drug databases and literature which are used to define the signature of disease. To map the compound data into disease signatures, we defined seven compound-disease associations for each compound(C). These association scores are divided into three groups according to the origin of compound signatures: i) structural similarity (S): structural similarities between C and KD, ii) Target based similarity (T): overlaps between the gene sets targeted by C and KD, overlaps between target genes of C and differentially expressed genes(DEGs) of KD, overlaps between target genes of C and DEGs under disease(D) conditions, iii) Expression based simialrity (E): overlaps between DEGs of C and target genes of KD, overlaps between DEGs of C and DEGs of KD, overlaps between DEGs of C and DEGs under D conditions. These association measures for each compound were integrated to evaluate a unified DR score. Using a benchmark set of known anticancer drugs, our method consistently showed high accuracy in the three types of cancer. For a systematic validation, we experimentally tested cytotoxic and migratory effects in glioblastoma cell lines for ~1000 FDA approved drugs. Our DR scores for glioblastoma showed a strong correlation with anti-tumor activity. Finally, Astemizole, Acemetacin, Terfenadine are chosen among the top DR candidates, whose anti-tumor activity was not previously reported for glioblastoma. In order to interpret their mode-of-actions, we reconstructed network models that revealed many direct and indirect targets leading to cytotoxicity and migration, as observed experimentally in the glioblastoma cell lines tested. It demonstrates the utility of our method to identify DR candidates can serve as a useful tool to provide a list of DR candidates and pathways relevant to their DR efficacy that can be tested in detailed functional studies.

G56: Bartosz Wojtas and B. Kaminska. TCGA-based analysis of gliomas uncovers a putative role of miRNA 155 in regulation of gene expression

Abstract: Introduction: Molecular mechanisms of progression from lower grade gliomas to anaplastic, highly malignant forms are poorly known. The Cancer Genome Atlas (TCGA) project provides an opportunity to study molecular aspects of transcription regulation among histological subtypes of gliomas in large patient cohorts.

Materials & Methods: In a present study both mRNA and miRNA expression datasets from histological subtypes of WHO II and III grade gliomas were acquired from the TCGA website. It includes next generation sequencing data for mRNA and miRNA for 54 astrocytoma, 53 oligodendroglioma and 37 oligoastrocytoma tumors. Many miRNAs were associated with differential gene expression in gliomas, threfore miRNA and mRNA expression was analyzed to find relationship between a level of specific miRNA and regulation of gene expression. Best inverse correlations for astrocytomas (rho < -0.5 and Bonferroni corrected p-value < 0.1) were identified and 4 algorithms (miRDB, microT4-CDS, Paccmit, Paccmit-CDS) were used for bioinformatics prediction of the possible target genes of miRNAs. Correlations confirmed by 2 out of 4 algorithms were reported. GSEA (gene set enrichment analysis) of Gene Ontology terms for biological function and CGP (chemical and genetic perturbations) enrichment was performed to investigate genes that were potentially regulated by miRNAs in gliomas.

Results: An integrated analysis of the observed miRNAs and mRNAs resulted in 100 highly correlated miRNA-mRNA pairs that fulfilled criteria of analysis. Strikingly 60% of them were putative regulations of miRNA-155, which is a very well known oncomir involved in glioblastoma development. GSEA analysis of putative targets of miRNA in astrocytoma revealed that amongst most enriched GO terms for biological functions ESTABLISHMENT AND OR MAINTENANCE OF CHROMATIN ARCHITECTURE and CHROMOSOME ORGANIZATION AND BIOGENESIS were most enriched functional groups. The most GSEA CGP enriched group consisted genes correlated with a proneural type of glioblastoma multiforme (WHO grade IV)

Conclusions: MiRNA-155 could be a key player in astrocytoma development. Main functional groups of miRNA targets are genes involved in chromosome organization and chromatin architecture. The expression pattern of genes that are putative targets of miRNA in astrocytoma is very similar to the expression patterns of the proneural type of glioblastoma multiforme.