VI Jornada de Bioinformàtica i Genòmica

Organitzada per:
Secció de Bioinformàtica i Biologia Computacional de la SCB
Secció de Genòmica i Proteòmica de la SCB
Associació Bioinformatics Barcelona – BIB

Patrocinada per:

PROGRAMA I RESUMS DE LES COMUNICACIONS

Auditori CaixaForum

Av. Francesc Ferrer i Guàrdia, 6-8.

Barcelona

20 de desembre de 2018

COMITÈ ORGANITZADOR:

Patrick Aloy (ICREA, IRB Barcelona)
Miquel Àngel Pujana (IDIBELL)
Ricard Gavaldà (UPC)
Mario Cáceres (ICREA, UAB)
Roderic Guigó (CRG-UPF)
Ana Ripoll (UAB, BIB)

SUPORT:

Mariàngels Gallego (SCB)
Maite Sánchez (SCB)
Begoña Duran (BIB)

PROGRAM

8:30 – 9:15 Registration

9:15 – 9:30 Wellcome and opening of the symposium
Sr Valentí Farràs (Fundació "La Caixa")
Dra Ana Ripoll (Bioinformatics Barcelona)
Dra Montserrat Corominas (Societat Catalana de Biologia)

SESSION I. Chair Marc Martí-Renom (ICREA, CRG-CNAG)

9:30 - 10:15 **Invited Lecture: Henk Stunnenberg** (RIMLS, Netherlands).
Epigenetic (de)regulation in health and disease.

10:15 - 10:30 **Pablo Baeza (CRG).** Combinatorial genetics reveals a scaling law for the effects of mutations on splicing.

10:30 - 10:45 **Pablo Latorre (IRB Barcelona).** Sensitive, high-throughput single-cell RNA-Seq reveals within-clonal transcript-correlations in yeast populations.

10:45 - 11:00 **Marta Puig (UAB).** Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR (ddPCR).

11:00 - 11:30 Coffee Break

SESSION II. Chair Baldo Oliva (UPF)

11:30 - 11:45 **Lídia Mateo (IRB Barcelona)**. Exploring the OncoGenomic Landscape of cancer.

11:45 - 12:00 **Bernhard Knapp (UIC)**. Predicting T-cell receptor binding using hierarchical natural Monte Carlo simulations.

12:00 - 12:15 **Janet Piñero (UPF).** Network, transcriptomic and genomic characterization of genes relevant for drug response.

12:15 - 13:00 **Invited Lecture: Júlio Sáez-Rodríguez** (BioQuant, Germany).
Dynamic logic models complement machine learning to improve cancer treatment.

13:00 - 14:30 Lunch and free poster viewing

SESSION III. Chair Xavier Daura (ICREA, UAB)

14:30 - 14:45 **Oriol Pich (IRB Barcelona).** Somatic and germline mutation rates in nucleosome-occupied DNA.
14:45 - 15:00 **Maria Pilar Francino (FISABIO).** Metabolic adaptation in the human gut microbiota during pregnancy and the first year of life.
15:00 - 15:15 **Sergio Picart-Armada (UPC)**. A tissue-specific network-based pathway test and application to GWAS data.
15:15 - 15:30 **Alejandro Caceres (ISGlobal)**. Extreme downregulation of chromosme Y and male disease.
15:30 - 15:45 **Marina Ruiz Romero (CRG)**. Time and tissue contribution to gene expression during tissue differentiation and development.
15:45 - 16:00 **Giovanni Iacono (CRG-CNAG)**. Single-cell transcriptomics unveils gene regulatory network plasticity.

16:00 - 16:30 Coffee Break

SESSION IV. Chair Patrick Aloy (ICREA, IRB Barcelona)

16:30 - 16:45 **Davide Cirillo (BSC)**. Training IBM Watson with MelanomaMine.
16:45 - 17:00 **Renée Beekman (IDIBAPS)**. Integration of genomic and epigenomic data, including the three-dimensional chromatin structure, refines regulatory mechanisms at chronic lymphocytic leukemia risk loci.
17:00 - 17:15 **Josu Aguirre Gómez (VHIR)**. The clinical costs of in silico tools: a novel approach to chose the best pathogenicity predictor for healthcare applications.

17:15 - 18:00 **Invited Lecture**: **Lars Juhl Jensen** (University of Copenhagen, Denmark). Population-wide data and text mining of electronic health records.

18:00 - 19:00 Poster viewing with authors and cocktail

19:00 - 19:15 *Genes* award to the best oral communication and poster and end of the symposium.

19:15 Free visit to the CaixaForum

Oral Presentations

COMBINATORIAL GENETICS REVEALS A SCALING LAW FOR THE EFFECTS OF MUTATIONS ON SPLICING

Pablo Baeza-Centurion[1,2*], Belén Miñana[2,3*], Jörn M. Schmiedel[1,2], Juan Valcárcel[2,3,4**], Ben Lehner[1,2,4,5**]

[1] Systems Biology Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, Barcelona 08003, Spain
[2] Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain
[3] Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, Barcelona 08003, Spain
[4] Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain
[5] Lead Contact
* These authors contributed equally to this work.
** These authors jointly supervised this work.

Presenting author: Pablo Baeza-Centurion (pablo.baeza@crg.eu)

Despite a wealth of molecular knowledge, quantitative laws for accurate prediction of biological phenomena remain rare. Alternative pre-mRNA splicing is an important regulated step in gene expression frequently perturbed in human disease. To understand the combined effects of mutations during evolution, we quantified the effects of all possible combinations of exonic mutations accumulated during the emergence of an alternatively spliced human exon. This revealed that mutation effects scale non-monotonically with the inclusion level of the exon in which the mutations are introduced, with each mutation having maximum effect at a predictable intermediate inclusion level. This scaling is observed genome-wide for *cis* and *trans* perturbations of splicing, including for natural and disease-associated variants. Mathematical modelling suggests that mutually-exclusive competition between alternative splice sites is sufficient to cause this non-linearity in the genotype-phenotype map. Combining the global scaling law with specific pairwise interactions between neighbouring mutations allows accurate prediction of the effects of complex genotype changes involving >10 mutations. Given the abundance of mutually exclusive molecular competitions, similar scaling of mutation effects is likely to be widespread in biology.

SENSITIVE, HIGH-THROUGHPUT SINGLE-CELL RNA-SEQ REVEALS WITHIN-CLONAL TRANSCRIPT-CORRELATIONS IN YEAST POPULATIONS

Mariona Nadal-Ribelles[1,2,3,4°], Saiful Islam[1,2°], Pablo Latorre[3,4°], Michelle Nguyen[1,2], Eulàlia de Nadal[3,4], Francesc Posas[3,4], Wu Wei[1,2,5], Lars M. Steinmetz[1,2,6].

° Equal contribution

1. Department of Genetics, Stanford University, School of Medicine, California, USA.
2. Stanford Genome Technology Center, Stanford University, California, USA.
3. Departament de Ciències Experimentals i de la Salut, Cell Signaling Research Group, Universitat Pompeu Fabra (UPF), Barcelona, Spain.
4. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain
5. CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China
6. European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany.

Presenting author: Pablo Latorre Doménech (pablo.latorre@irbbarcelona.org)

Single-cell RNA-seq (scRNA-seq) has revealed extensive cellular heterogeneity within many organisms but few methods have been developed for microbial clonal populations. The yeast genome displays an unusually dense transcript spacing with interleaved and overlapping transcription from both strands, resulting in a minuscule but complex pool of RNA protected by a resilient cell wall. Here, we developed a sensitive, scalable and inexpensive yeast single-cell RNA-seq (yscRNA-seq) method that digitally counts transcript start sites (TSS) in a strand- and isoform-specific manner with unique molecular identifiers (UMI). YscRNA-Seq detects expression of low-abundant, non-coding RNAs and at least half of the protein-coding genome in each cell. From just one single-cell transcriptome experiment of a bulk population, enough heterogeneity is uncovered between individual cells to identify biological associations without the need for perturbation experiments necessary to derive correlations from bulk data. Within cells of a single clonal population, we observe negative expression correlation of sense/antisense pairs while duplicated gene pairs and divergent transcripts co-express. By combining yscRNA-Seq with index sorting, which allows phenotypic characterization of cells, we uncover a linear cell size-dependent change in absolute RNA content. Although we detect an average of ~3.5 molecules per gene, a single cell tends to restrict expressed isoforms. Remarkably, there is a highly variable expression within metabolic genes, whose stochastic expression primes cells for fitness benefit towards the corresponding environmental challenge. These findings suggest functional transcript diversity as a mechanism for providing a selective advantage to individual cells within otherwise transcriptionally heterogeneous microbial populations.

DETERMINING THE IMPACT OF UNCHARACTERIZED INVERSIONS IN THE HUMAN GENOME BY DROPLET DIGITAL PCR (DDPCR)

Marta Puig[1], Jon Lerga-Jaso[1], Carla Giner-Delgado[1], Sarai Pacheco[1], David Izquierdo[1], Alejandra Delprat[1], Jack F. Regan[2], George Karlin-Neumann[2], Mario Cáceres[1,3]

[1] Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
[2] Digital Biology Center, Bio-Rad Laboratories, Pleasanton, CA, United States.
[3] ICREA, Barcelona, Spain.

Despite the growing interest in characterizing genomic structural variation, the presence of large repeats at the breakpoints is an important limitation that precludes the analysis of many variants. Here we describe a novel linkage-based application of droplet digital PCR (ddPCR) for the validation and genotyping of polymorphic inversions mediated by inverted repeats (IRs). We developed ddPCR assays for a total of 18 human inversions, including several predicted inversions that had yet to be validated, and others for which there are not direct high-throughput genotyping assays, like the well-characterized chromosome 17 inversion. Inversions ranged from 6.8 to 747 kb with IRs from 6.3 to 134 kb, and we genotyped them across 95 individuals from three different human populations. Our analysis allowed us to validate all the tested inversions and demonstrate that the technique is highly accurate and reproducible. Inversions show a wide variation in frequency and several have significant differences between continents. Moreover, all except two show clear signs of being recurrent, given the lack of association of the inversion alleles with SNPs or haplotypes. Finally, thanks to the generated genotyping data, we have been able to check the effect of these inversions on gene expression, validating gene expression differences reported previously for two inversions and finding new candidate associations that require further analysis. Therefore, in this work we provide a tool to screen these elusive variants quickly in a large number of samples for the first time, making possible to assess their potential functional effects and clinical implications and opening new avenues for future investigation.

EXPLORING THE ONCOGENOMIC LANDSCAPE OF CANCER

Lídia Mateo[1,*], Oriol Guitart-Pla[1], Miquel Duran-Frigola[1], Patrick Aloy[1,2]

[1] Institute for Research in Biomedicine (IRB Barcelons) and Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
[2] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

The widespread incorporation of next-generation sequencing into clinical oncology has yielded an unprecedented amount of molecular data from thousands of patients. A main current challenge is to find out reliable ways to extrapolate results from one group of patients to another and to bring rationale to individual cases in the light of what is known from the cohorts.

We present *OncoGenomic Landscapes*, a framework to analyze and display thousands of cancer genomic profiles in a 2D space. Our tool allows users to rapidly assess the heterogeneity of large cohorts, enabling the comparison to other groups of patients, and using driver genes as landmarks to aid in the interpretation of the landscapes. In our web-server, we also offer the possibility of mapping new samples and cohorts onto 22 predefined landscapes related to cancer cell line panels, organoids, patient-derived xenografts, and clinical tumor samples.

Contextualizing individual subjects in a more general landscape of human cancer is a valuable aid for basic researchers and clinical oncologists trying to identify treatment opportunities, maybe yet unapproved, for patients that ran out of standard therapeutic options. The web-server can be accessed at https://oglandscapes.irbbarcelona.org/.

PREDICTING T-CELL RECEPTOR BINDING USING HIERARCHICAL NATURAL MONTE CARLO SIMULATIONS

Bernhard Knapp
Bioinformatics and Immunoinformatics Research Group, UIC Barcelona, Spain

Presenter's email: bknapp@uic.es

The affinity of T-cell receptors (TCRs) to peptide/MHC (pMHC) complexes (e.g. a specific cancer antigen) can be improved by introducing mutations in TCR sequences. Testing a large number of mutant TCRs experimentally is expensive and time consuming. A computational high-throughput approach could test such large numbers of TCRs in short time and at little cost. This method would be of high value in many areas of immunology such as cancer, allergies and vaccines.
However, to date no such method exists. Current computational methods are mainly based on free energy calculations and computationally extremely expensive, which does not allow their application for screening a large number of different TCR mutants.
We have extended our previously developed MOSAICS protocol (Knapp et al. 2017 Bioinformatics; Sim et al. 2012 PNAS) for the prediction of TCR binding. MOSAICS implements a unique combination of hierarchical Monte Carlo, coarse-graining, stochastic chain closure, and (local) temperature annealing cycles. By this way MOSAICS allows to sample essential degrees of freedom while non-essential degrees are restricted. This allows the prediction of association, dissociation, and binding of TCRs with peptide/MHC using structural modelling. MOSAICS is orders of magnitudes faster than conventional molecular dynamics simulations and achieves high agreement with experimental data as we have shown before in studies of multiple other systems.
We think that this MOSAICS extension will be of high value for precision medicine and the development of patient specific TCRs.

NETWORK, TRANSCRIPTOMIC AND GENOMIC CHARACTERIZATION OF GENES RELEVANT FOR DRUG RESPONSE

Janet Piñero, Abel Gonzalez-Perez, Emre Guney, Joaquim Aguirre-Plans, Ferran Sanz, Baldo Oliva, and Laura I. Furlong

Understanding the mechanisms that underlie drug therapeutic action and toxicity is crucial for the prevention and management of drug adverse reactions, and paves the way for more efficient and rational drug design. The characterization of drug targets, proteins participating in drug metabolism, and proteins associated to side effects is a first step in this direction. Here, we hypothesize that proteins involved in the therapeutic effect of drugs and in their adverse reactions have distinctive transcriptomics, genomics and network features. To test this hypothesis, we explored the properties of these proteins within the context of global and organ-specific interactomes, using global, mesoscopic, and local network features. By leveraging genomic variation data from 60K subjects, we assessed the differences in the tolerance of these proteins to loss-of-function variants. Finally, we have analyzed their pattern of expression across healthy tissues. We found that drug targets that mediate side effects are more central in cellular networks, more intolerant to loss-of-function variation, and present a wider breadth of tissue expression than targets that do not mediate side effects. Actually, they behave similarly to the proteins that mediate drug adverse reactions. In contrast, drug metabolizing enzymes and transporters are more peripheral in the network, more tolerant to deleterious variants, and are more expressed in liver. Our findings highlight the importance of assessing genomic variability across drug related genes, and pinpoints to network, genomic and transcriptomic features as useful indicators to identify safer drug targets.

SOMATIC AND GERMLINE MUTATION PERIODICITY FOLLOW THE ORIENTATION OF THE DNA MINOR GROOVE AROUND NUCLEOSOMES

Oriol Pich[1], Ferran Muiños[1], Radhakrishnan Sabarinathan[1,5], Iker Reyes-Salazar[1], Abel Gonzalez-Perez[1,2,4], Nuria Lopez-Bigas[1,2,3,4]
[1] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain
[2] Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[4] Co-senior authors

Mutation rates along the genome are highly variable and influenced by several chromatin features. Here, we addressed how nucleosomes, the most pervasive chromatin structure in eukaryotes, affect the generation of mutations. We discovered that within nucleosomes, the somatic mutation rate across several tumor cohorts exhibits a strong 10 base pair (bp) periodicity. This periodic pattern tracks the alternation of the DNA minor groove facing toward and away from the histones. The strength and phase of the mutation rate periodicity are determined by the mutational processes active in tumors. We uncovered similar periodic patterns in the genetic variation among human and *Arabidopsis* populations, also detectable in their divergence from close species, indicating that the same principles underlie germline and somatic mutation rates. We propose that differential DNA damage and repair processes dependent on the minor groove orientation in nucleosome-bound DNA contribute to the 10-bp periodicity in AT/CG content in eukaryotic genomes.

METABOLIC ADAPTATION IN THE HUMAN GUT MICROBIOTA DURING PREGNANCY AND THE FIRST YEAR OF LIFE

María José Gosalbes[1,2], Joan Compte[1,3], Silvia Moriano-Gutierrez [1,4], Yvonne Vallès[1,5], Nuria Jiménez-Hernández[1,2], Xavier Pons[1], Alejandro Artacho[1], M. Pilar Francino1[,2]

1. Unitat Mixta d'Investigació en Genòmica i Salut, Fundació per al Foment de la Investigació Sanitària i Biomèdica de la Comunitat Valenciana (FISABIO-Salut Pública) / Institut de Biologia Integrativa de Sistemes (Universitat de València), València, Spain
2. CIBER en Epidemiología y Salud Pública, 28029 Madrid, Spain
3. Present address: Neurodegenerative Diseases Research Group, Vall d'Hebron Research Institute (VHIR)-Center for Networked Biomedical Research on Neurodegenerative Diseases (CIBERNED), 08035 Barcelona, Spain
4. Present address: Pacific Biosciences Research Center, School of Ocean and Earth Science and Technology, University of Hawaii at Mānoa, Honolulu, Hawaii, 96822, USA
5. Present address: Department of Biological and Chemical Sciences, The University of the West Indies, Cave Hill campus, Cave Hill, Barbados

Presenter's email: mpfrancino@gmail.com

The relationship between the gut microbiome and the human host is dynamic and we may expect adjustments in microbiome function if host physiology changes. Metatranscriptomic approaches should be key in unraveling how such adjustments occur. We employ metatranscriptomic sequencing analyses to study gene expression in the gut microbiota of infants through their first year of life, and of their mothers days before delivery and one year afterwards. In infants, hallmarks of aerobic metabolism disappear from the microbial metatranscriptome as development proceeds, while the expression of functions related to carbohydrate transport and metabolism increases and diversifies, approaching that observed in non-pregnant women. Butyrate synthesis enzymes are overexpressed at three months of age, even though most butyrate-producing organisms are still rare. This suggests that butyrate production may be ensured in the gut of young infants before the typical butyrate synthesizers of the adult gut become abundant. In late pregnancy, the microbiota readjusts the expression of carbohydrate-related functions in a manner consistent with a high availability of glucose, suggesting that it may be able to access the high levels of blood glucose characteristic of this period. Moreover, late pregnancy gut bacteria may reach stationary phase, which may affect their likelihood of translocating across the intestinal epithelium (Gosalbes et al., EBioMedicine Nov 2018).

A TISSUE-SPECIFIC NETWORK-BASED PATHWAY TEST AND APPLICATION TO GWAS DATA

Sergio Picart-Armada[1,2,3], Wesley K. Thompson[4,5], Alfonso Buil[4], Alexandre Perera-Lluna[1,2,3]

1. B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Spain
2. CIBER-BBN, Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine, Spain
3. Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Spain
4. Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Denmark
5. Department of Family Medicine and Public Health, University of California San Diego, United States

Biological pathway analysis is a common approach to relate molecular phenotypes to diseases by including contextual knowledge from annotated comprehensive databases. Finding over-represented biological pathways serves as a quality control and can point towards novel discoveries. The leverage of public tissue-specific interactomes enables including topological patterns to increase the number of findings and their specificity. Unfortunately, most pathway analysis tools do not account for tissue-specific knowledge.

We present a generalisation of over-representation that builds on tissue-specific networks. By using regression models, we test whether an input gene list is closer than expected to a target gene set (pathway) within a tissue-specific network. Such closeness is quantified through label propagation processes within the network being tested. Our findings are compared to classical over-representation and to other state-of-the-art network-based tests.

We have applied our approach on synthetic gene lists generated on a yeast interactome with 2,375 proteins and 11,693 high-confidence interactions. We prove that, on a grid of signal parameters for our theoretical model, our network-based approach is more powerful than classical over-representation and other network-based methods. Currently, we are benchmarking network-based tests on tissue-specific interactomes and GWAS summary statistics for several complex diseases, such as asthma and diabetes. We assess their capability to rediscover known relevant processes within their tissues, not found through classical over-representation.

EXTREME DOWNREGULATION OF CHROMOSME Y AND MALE DISEASE

Alejandro Caceres, Aina Jene, Tonu Esko, Luis Perez-Jurado, Juan R Gonzalez

Institution: ISGlobal

Mosaic loss of chromosome Y (LOY) in blood has been associated to all-cause mortality in men. However, it is unknown whether LOY in affected tissues precedes disease. The analysis of 6,898 expression arrays and multiple data from the genotype-tissue expression (GTEx) and the cancer genotype atlas (TCGA) projects revealed that the extreme down-regulation of chromosome Y (EDY) is a functional consequence of LOY and a likely route towards disease. We observed that EDY was associated to several LOY-related conditions and increased the risk of various cancers, where it correlated with differential changes in methylation across Y that were independent of LOY. Remarkably, we identified a subgroup of cancer patients with EDY and no LOY characterized by decreased survival and increased copy number duplications containing *EGFR* and *MDM2*. Our analyses support EDY as a novel risk factor of male disease, which can be triggered by different factors including LOY.

TIME AND TISSUE CONTRIBUTION TO GENE EXPRESSION DURING TISSUE DIFFERENTIATION AND DEVELOPMENT

Marina Ruiz-Romero[1,2], Cecilia Coimbra Klein[1,2,4], Sílvia Pérez-Lluch[1,2], Alessandra Breschi[1,2], Amaya Abad[1,2], Emilio Palumbo, Roderic Guigó[1,2,3]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[3]Institut Hospital del Mar d'Investigacions Mediques (IMIM), Barcelona 08003, Spain
[4] Departament de Genètica, Microbiologia i Estadística and Institute of Biomedicine (IBUB), Universitat de Barcelona. Barcelona, Catalonia, Spain,

During development most tissues undergo striking changes in order to develop into functional organs. All along this process, the identity of each tissue arise from the particular combination of master regulator transcription factors that specifically control the expression of relevant genes for growth, pattern formation and differentiation. In this scenario regulation of gene expression turns to be essential to determine cell fate and tissue specificity. Although many studies aim to decipher tissue signature through the analysis of their transcriptome profiles, most lack temporal information during development and, in consequence, many differentiation events are poorly understood.
To characterize specific tissue dynamic transcriptional profiles during differentiation, in this study, we track down the transcriptome of commited cells throughout differentiation of eye, leg and wing of *Drosophila melanogaster*. We identified three main sets of genes: time-specific (commonly expressed genes that change across time), tissue-specific genes and time-tissue-specific genes. Our analyses indicate that transcriptome profiles of different fly tissues at a specific stage of differentiation are more similar between each other than they are to their own lineage in subsequent stage, suggesting there is a tightly and coordinated regulation among tissues. Interestingly, in other fly tissues, for instance brain, this mechanism is also conserved. To describe deeply the regulatory program underlying differentiation, we study the dynamics of chromatin accessibility surrounding transcriptional changes, as well as the differential expression of associated Transcription factors (TFs). Overall, we designed a gene regulatory network to model the transcriptional program of eye, leg and wing differentiation.
In conclusion, we have seen that although differences among tissues increase over time, a common gene regulatory network is leading the differentiation process in all tissues.

SINGLE-CELL TRANSCRIPTOMICS UNVEILS GENE REGULATORY NETWORK PLASTICITY

Giovanni Iacono[1,*], Ramon Massoni-Badosa[1], Holger Heyn[1,2,3,*]

[1] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain
[3] Lead author

* Correspondence: holger.heyn@cnag.crg.eu (H.H.), giovanni.iacono@cnag.crg.eu (G.I.)

Single-cell RNA sequencing (scRNA-seq) plays a pivotal role in our understanding of cellular heterogeneity. Current analytical workflows are driven by categorizing principles that consider cells as individual entities and classify them into complex taxonomies. We have devised a conceptually different computational framework based on a holistic view, where single-cell datasets are used to infer global, large-scale regulatory networks. We developed correlation metrics that are specifically tailored to single-cell data, and then generated, validated and interpreted single-cell-derived regulatory networks from organs and perturbed systems, such as diabetes and Alzheimer's disease. Using advanced tools from graph theory, we computed an unbiased quantification of a gene's biological relevance, and accurately pinpointed key players in organ function and drivers of diseases. Our approach detected multiple latent regulatory changes that are invisible to single-cell workflows based on clustering or differential expression analysis, significantly broadening the biological insights that can be obtained with this leading technology.

TRAINING IBM WATSON WITH MELANOMAMINE

Cirillo D*, Cañada A, Corvi JO, Fernández González JM, Lopez-Martin JA, Capella-Gutierrez S, Krallinger M, Valencia A

Barcelona Supercomputing Center (BSC).

The outstanding breakthrough of Big Data is enhancing unprecedented opportunities to advance cancer research, especially in areas calling for improvement in early detection and prevention such as melanoma. The massive volume of biomedical information is largely composed of unstructured data, which includes text such as research articles and clinical reports. Navigating the current flood of unstructured information is a groundbreaking challenge that has been taken up by new technologies based on Natural Language Processing (NLP) collectively referred to as Cognitive computing systems. IBM Watson is one of the most acknowledged platforms for Cognitive computing.

In order to surface insights from massive volumes of unstructured data, Watson forms inferences by assessing the context pertaining to the specific information of interest. In this regards, of primary importance to Watson operations is the so-called *knowledge corpus*, providing the system with both immediate and broad domain-specific information to be teased apart using NLP techniques.

In this work, we employ MelanomaMine (http://melanomamine.bioinfo.cnio.es/), a text mining application designed to process melanoma-related biomedical literature, in order to generate a melanoma-specific knowledge corpus to be processed by Watson. MelanomaMine uses information extraction and machine learning approaches to score and classify textual data based on cancer relevance detected by Support Vector Machines (SVMs) techniques. Moreover, it enables a general free text retrieval and several semantic search options bound to the co-occurrence of a particular bio-entity (genes, proteins, mutations and chemicals/drugs).

In this presentation, I will discuss the steps and difficulties of the training process, the results showing how Watson surveys the content of melanoma knowledge corpus, and the future avenues for the application of Cognitive computing systems to biomedical problems.

This work has been founded by BBVA Foundation and the BSC Research Collaboration Agreement with IBM.

INTEGRATION OF GENOMIC AND EPIGENOMIC DATA, INCLUDING THE THREE-DIMENSIONAL CHROMATIN STRUCTURE, REFINES REGULATORY MECHANISMS AT CHRONIC LYMPHOCYTIC LEUKEMIA RISK LOCI

Beekman Renée[1,2*], Speedy Helen E.[3*], Chapaprieta Vicente[4], Orlando Giulia[3], Law Philip J.[3], Martín-García David[1,2], Gutiérrez-Abril Jesús[5], Catovsky Daniel[3], Beà Sílvia[1,2], Puente Xose S.[2,5], Allan James M.[6], López-Otín Carlos[2,5], Campo Elias[1,2,4,7], Houlston Richard S.[3**], Martín-Subero José I.[1,2,4**]

1. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), 08036 Barcelona Spain
2. Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain
3. Division of Genetics and Epidemiology, Institute of Cancer Research, London SW7 3RP, UK
4. Departament de Fonaments Clinics, Facultat de Medicina, Universitat de Barcelona, 08036 Barcelona, Spain
5. Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), Universidad de Oviedo, 33006 Oviedo, Spain
6. Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne NE2 4HH, UK
7. Hematopathology Section, Hospital Clinic of Barcelona, Barcelona, Spain
Presenter e-mail: beekman@clinic.ub.es

*These authors contributed equally to the work
**These authors contributed equally to the work

Genome-wide association studies have provided evidence for inherited predisposition to chronic lymphocytic leukemia (CLL), identifying 42 (non-HLA) genomic regions influencing CLL risk. However, efforts defining the mechanisms mediating the risk at these, largely non-coding, loci have been constrained by a lack of integrated genome-wide data in large CLL series.

Here, we aimed to refine the regulatory mechanisms underlying CLL predisposition by (i) analysing high-resolution chromatin state maps and (ii) integrating genetic, epigenetic and transcriptomic information in primary CLL cases, (iii) by *in silico* transcription factor (TF) binding analysis and (iv) by studying the three-dimensional (3D) chromatin structure using promoter capture Hi-C in normal B cells and CLL.

Eighty-one percent of the risk loci were enriched for regulatory elements in CLL, suggesting a specific regulatory role for these loci in CLL pathogenesis. This regulatory role was further emphasized by the fact that 30 risk loci were associated with genome activity (H3K27ac), chromatin accessibility (ATAC-seq) and/or gene expression quantitative trait loci (QTLs) in up to 452 primary CLLs. Yet, we characterised the mechanism of action at these loci in closer detail by identifying 105 single nucleotide polymorphisms as potential causal risk variants, which showed decreased binding affinity for B-cell related TFs and increased affinity for FOX, NFAT and TCF/LEF TF family members, among others. Finally, we observed significant interactions between the risk loci and 15 eQTL gene loci (e.g. TLE3, IPCEF and BMF) at the 3D chromatin level in CLL and normal B cells, highly suggestive for a direct regulatory link between the risk loci and expression of their target genes in relation to CLL pathogenesis.

In conclusion, by defining potential functional risk variants at CLL risk loci, their effect on TF binding and their association with downstream gene expression events we offer improved insights into the functional basis of CLL predisposition.

THE CLINICAL COST OF IN SILICO TOOLS: A NOVEL APPROACH TO CHOSE THE BEST PATHOGENICITY PREDICTOR FOR HEALTHCARE APPLICATIONS

Josu Aguirre[1], Natàlia Padilla[1], Casandra Riera[1], Xavier De la Cruz[1,2]

1. Vall d'Hebron Research Institute (VHIR), Passeig de la Vall d'Hebron 119-129, 08035 Barcelona, Catalonia, Spain.
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

Presenter e-mail: josu.aguirre@vhir.org

The applicability of NGS experiments in the clinical setting is presently limited by our inability to establish the pathogenic nature of genetic variants. Many tools have been developed for this purpose, but none of them has a perfect success rate. In this context, identifying the best tool for healthcare applications is a challenging problem. Current performance parameters (MCC, accuracy, ROC/AUC, etc.) only provide a limited answer, because they do not reflect the costs of specific applications. Here we present a novel approach to this problem, where the performance of pathogenicity predictors is estimated using a clinical cost-related parameter.

The simplicity of the formalism allows an easy and fast comparison between many different predictors simultaneously and the identification of the best clinical scenario for each of them. We illustrate this analysis using several standard predictors (SIFT, PolyPhen-2 Hvar, PolyPhen-2 Hdiv, MutationTaster2, CADD, PON-P2) and describe how they distribute along the cost range.

Apart from its simplicity, our approach has the virtue of providing an exact solution to the problem of comparing the performance of bioinformatics predictors when they have different coverages, a relevant and yet unsolved problem. Application of our formalism to this problem shows an interesting trend where some top-ranking methods suddenly drop in performance when considered in the clinical setting.

Overall, our results give an original view of the pathogenicity prediction field, a view in which the idea of a unique best predictor vanishes and is replaced by that of an optimal partition of the cost space among predictors.

Posters

ENCIRCLING THE REGIONS OF THE PHARMACOGENOMIC LANDSCAPE THAT DETERMINE DRUG RESPONSE

Adrià Fernández-Torras 1 , Miquel Duran-Frigola 1 and Patrick Aloy 1,2
1. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine
(IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

The integration of large-scale drug sensitivity screens and genome-wide experiments is changing the field of pharmacogenomics, revealing molecular determinants of drug response without the need for a priori, hypothesis-driven assumptions about drug action.

In particular, transcriptomic signatures of drug sensitivity may guide drug repositioning, prioritize drug combinations and suggest new therapeutic biomarkers. However, the inherent complexity of transcriptomic signatures, with thousands of genes differentially expressed, makes them hard to interpret, giving poor mechanistic insights and hampering translation to the clinics. Here we show how network biology can help simplify transcriptomic drug signatures, yielding functionally-coherent, less noisy drug modules.

We successfully analyzed 170 drugs tested in 637 cancer cell lines, proving a broad applicability of our approach and evincing an intimate relationship between modules' gene expression levels and drugs' mechanisms of action. Further, we have characterized multiple aspects of our transcriptomic modules. As a result, the drugs included in this study are now annotated well beyond the reductionist (target-centered) view.

EUROPEAN GENOME-PHENOME ARCHIVE (EGA) - GRANULAR SOLUTIONS FOR THE NEXT 10 YEARS

Name/affiliation/email of presenting author:
Audald Lloret-Villas (audald.lloret@crg.eu) - Centre for Genomic Regulation - European Genome-phenome Archive - ELIXIR-ES
Names/affiliations/email of any co-authors:
Jordi Rambla de Argila (jordi.rambla@crg.eu) - Centre for Genomic Regulation - European Genome-phenome Archive - ELIXIR-ES

Short description

As The European Genome-phenome Archive (EGA) (https://ega-archive.org) moves into it's 10th year it continues to play a pivotal role for public bio-molecular data archiving, sharing, standardisation and reproducibility. The EGA is currently listed as one of the ELIXIR core database services (https://www.elixir-europe.org/services/database).

As the genomics community awareness of data sharing and reproducibility increases, complex services and granular solutions are needed from the EGA. We will herein present several advanced features designed for a wide range of users; these new tools and technologies include the EGA Beacon (developed within the GA4GH framework), EGA APIs for metadata submission, retrieval and data access, as well as the data visualisation projects.

We will finally cover all the new advances achieved for human data federation. The EGA is currently coordinating the efforts, within the ELIXIR framework, for agreeing and developing necessary solutions towards national/local data governance (Local EGA) with a centralised metadata repository, which ensures proper discoverability.

Abstract

The EGA archives and distributes mainly genomic and phenotypic human data in an encrypted and secure manner; files are shared upon the authorisation of the relevant Data Access Committee (DAC).

Given the heterogeneous profile of EGA users, an important variety and granularity of services and tools are required. On one hand, small data submissions and individual researchers appreciate user-friendly interfaces and data management approaches. Whilst larger complex projects and international consortia require a more technical approach.

During the presentation we will provide the audience with a brief overview of how all of these varying requirements by EGA users are addressed.

The EGA Beacon is able to highlight existing variants within a dataset and the possibility to retrieve information from samples or experiments registered at the EGA will be described. In addition, the necessary technologies for data visualisation, such as data access API and htsget will be seen in the context of RD-Viz, a portal for visualisation of rare disease files.

Federation of data silos and repositories is increasingly seen as a necessity within the omics services, specially when inter-related to health initiatives. Local EGA is one of the approaches expected to be implemented across Europe and worldwide. The last steps of this community effort are to be presented during the talk, as well.

Finally, from the submitter perspective, we will cover both the new, powerful and intuitive Submitter Portal as well as the submission API, which enables programmatic metadata submission and standardisation.

COMBINING METAOMICS APPROACHES TO UNDERSTAND THE PROCESS OF HUMAN GUT MICROBIOTA DEVELOPMENT

*A. Rey-Mariño[1], S. Ruiz-Ruiz[1], N. Jiménez-Hernández[1], J. Pons[1], A. Artacho[1] and M. P. Francino[1,2]

[1] Unidad Mixta de Investigación en Genómica y Salud- Centro Superior de Investigación en Salud Pública (Generalitat Valenciana)--Institut Cavanilles de Biodiversitat i Biologia Evolutiva (Universitat de València), Valencia, Spain.
[2] CIBER en Epidemiología y Salud Pública (CIBERESP), Spain.

The intestinal microbiota is deployed along the development of the individual and its composition and functions differ depending on age. Thus, for example, during the first year of life the infant gut microbiota becomes more complex and increases its resemblance to that of the mother, that is, to that of an adult. Knowing about the changes that occur in the gut microbiota throughout life can help the diagnosis, treatment and prevention of diseases related to metabolic and inmune alterations. The combination of the different high-throughput metaomics is key to understand the process of intestinal microbiota development and at what level and to what extent these changes occur. To adress these issues we performed sequencing analysis to obtain measurements of microbial species, gene and gene transcript composition of the gut microbiota in a prospective cohort of 12 toddlers, 13 adolescents and 35 adults in order to establish at what age the microbiota reaches typical adult characteristics and to compare the degree of microbiota stability at different ages. We observe that the important differences that remain between the gut microbiota of infants and that of adults progressively subside during childhood and adolescence. At the taxonomic level, one main difference between toddlers, on one hand, and adolescents and adults, on the other, is the elevated abundance of *Bifidobacterium*. Our results show a directional change toward the taxonomic composition, functional composition and gene expression patterns of the adult microbiome. However, in adolescents, some gene abundances and expression levels still are found to differ from those of adults. As the prospective sampling progresses, we will perform temporal stability analyses to define the amount and speed of microbiota variation in the different age groups.

HYPERLINK "mailto:arey@gva.es" arey@gva.es

INTEGRATED ANALYSIS OF GERMLINE AND TUMOR DNA IDENTIFIES NEW CANDIDATE GENES INVOLVED IN FAMILIAL COLORECTAL CANCER

Marcos Díaz-Gay[1], Sebastià Franch-Expósito[1], Solip Park[2], Fran Supek[3], Jenifer Muñoz[1], Coral Arnau-Collell[1], Laia Bonjoch[1], Anna Gratacós-Mulleras[1], Paula A. Sánchez-Rojas[1], Clara Esteban-Jurado[1], Teresa Ocaña[1], Miriam Cuatrecasas[4], Maria Vila-Casadesús[5], Juan José Lozano[5], Genis Parra[6], Steve Laurie[6], Sergi Beltran[6], EPICOLON consortium, Antoni Castells[1], Luis Bujanda[7], Joaquín Cubiella[8], Francesc Balaguer[1], Sergi Castellví-Bel[1].

1. Gastroenterology Department, Institut d'Investigacions Biomèdiques August Pi I Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Hospital Clínic, Barcelona, Spain.
2. Systems Biology Program, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
3. Institut de Recerca Biomèdica (IRB Barcelona), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
4. Pathology Department, Hospital Clínic, Barcelona, Spain
5. Bioinformatics Platform, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Barcelona, Spain
6. CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
7. Gastroenterology Department, Hospital Donostia-Instituto Biodonostia, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Basque Country University, San Sebastián, Spain
8. Gastroenterology Department, Complexo Hospitalario Universitario de Ourense, Instituto de Investigación Sanitaria Galicia Sur, Ourense, Spain.

Presenter e-mail: diaz2@clinic.cat

BACKGROUND

Colorectal cancer (CRC) presents sometimes familial aggregation but no alterations in the known hereditary CRC genes. We aimed for the identification of new candidate genes potentially involved in germline predisposition to familial CRC.

METHODS

Integrated analysis of germline and tumor whole exome sequencing data was performed in eighteen unrelated CRC families. Deleterious single nucleotide variants (SNV) and loss of heterozygosity (LOH) were assessed as candidates for first germline or second somatic hits. Candidate tumor suppressor genes were selected in case that alterations were detected in both germline and somatic DNA, fulfilling Knudson's two-hit hypothesis. Somatic mutational profiling and signature analysis were also performed for all samples.

RESULTS

A series of germline-somatic variant pairs were detected. In all cases the first hit was presented as a rare SNV, whereas the second hit was either a different SNV (3 genes) or a LOH affecting the same gene (148 genes). ADCY8, ATM, BRCA2, ERCC2, IGF2R and SMARCA4 were among the most promising candidate genes for germline CRC predisposition.

CONCLUSIONS

Identification of new genes involved in familial CRC can be achieved by our integrated analysis. Further functional studies and replication in additional cohorts are required to confirm the selected candidates.

BEA: A WEB TOOL FOR BIOMARK GENE EXPRESSION ANALYSIS

AUTHORS AND INSTITUTIONS:
Beatriz Cadenas[1,2,3], Mayka Sanchez[1,4,5], Cristian Tornador[2,5,6]
M.Luz Calle[3]
Edgar Sanchez[1], Isaac Cebrián[2]

[1] Iron Metabolism: Regulation and Diseases Group, Josep Carreras Leukaemia Research Institute (IJC), Campus Can Ruti, Badalona, Barcelona, Spain.
[2]Whole Genix S.L, Barcelona, Spain
[3]Universitat de Vic-Universitat Central de Catalunya, Catalonia, Spain.
[4]Program of Program of Predictive and Personalized Medicine of Cancer (PMPPC), Institutd'Investigació Germans TriasiPujol (IGTP), Campus Can Ruti, Badalona, Barcelona, Spain.
[5]BloodGenetics S.L., Esplugues de Llobregat, Barcelona, Spain
[6]Fundación Teresa Moretó, Esplugues de Llobregat, Barcelona, Spain

Biomark HD system is a high-throughput technology that performs more than nine thousands of real-time quantitative PCR (qPCR) reactions simultaneously. Although numerous statistical tools are available in the public domain for the analysis of conventional qPCR experiments, this is not the case when many qPCR dataset from high-throughput experiments have to be analysed. Therefore, there is a need to provide a user-friendly software able to analyze Biomark output data.

To meet this need, we are developing BEA, Biomark Expression Analysis, a web tool designed to analyse Real Time qPCR data from Biomark HD System across multiple conditions and replicates. Analysis includes relative quantification using normalization against a reference gene ($\Delta\Delta_{CT}$ method), and appropriate statistical testing to identify differential expressed genes between tested groups, taking into account test of normality and number of groups. The web application is written in R using the Shiny platform. BEA has been tested using Bimark 96.96 dynamic array data from myelodisplastic syndrome patients.

BEA will help scientists perform gene expression analyses by interactively visualizing each step with customizable options and intuitive graphical user interface. The program provide comprehensive results visualization, downloadable plots and a summarizing report, and it would not require programming knowledge. BEA will be an open source software freely available online.

This research was supported by grant SAF2015-70412-R (MINECO), DJCLS R14/04 from Deutsche José Carreras leukämie Stiftung, 2017 SGR 288 GRC (GRE) Generalitat de Catalunya and from Fundació Internacional Josep Carreras and Obra Social “la Caixa” Spain to M.S.4
TOPIC: Omics/Bioinformatics
KEYWORDS: Biomark/ Web Tool/ High-throughput Gene Expression Analysis /

PREDICTION OF TRANSCRIPTION FACTOR BINDING BY STRUCTURAL MODELING

Oriol Fornes, Alberto Meseguer, Filip Årman, Jaume Bonet and Baldo Oliva

Knowledge of transcription factor (TF) binding sites is key to understanding how genes are regulated. Yet the binding preferences of most eukaryotic TFs remain unknown. In this scenario, the development of computational tools as a complement to experimental procedures is fundamental. Here, we introduce ModCRE, a homology modeling-based approach that combines structural information and protein binding microarray (PBM) data to predict the binding preferences of TFs and model TF-DNA interactions. Besides, ModCRE uses bacterial one-hybrid data to make specific predictions for TFs from the zinc fingers family. ModCRE was applied to the following tasks: 1) discriminating bound from unbound PBM 8-mers; 2) predicting JASPAR profiles from experimental methods excluding PBMs; and 3) modeling the structure of the INFß human enhanceosome. Thanks to its automated homology modeling pipeline for TF-DNA interactions, ModCRE can be applied on a large scale to complement the knowledge of gene regulatory networks.

A SEQUENCE-BASED DEEP LEARNING PROTEOCHEMOMETRICS MODEL FOR BINDING AFFINITY PREDICTION

Angela Lopez-del Rio[1,2,3], David Vidal[1], Alfons Nonell-Canals[1], Alexandre Perera-Lluna[2,3]
[1]Mind the Byte S.L., 08007, Barcelona, Spain.
[2]B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, 08028 Barcelona, Spain.
[3]Department of Biomedical Engineering, Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Esplugues de Llobregat, 08950 Barcelona, Spain.
angela@mindthebyte.com

In computer-aided drug discovery, classical Quantitative Structure-Activity Relationship (QSAR) modelling predicts biological activity for a specific target protein from ligand descriptors. However, it ignores protein information, overlooks possible cross-interactions with other targets and can lead to chemotype bias. Proteochemometrics is a conceptual extension of QSAR which considers information both from the target and the ligands, overcoming these issues.

Deep Learning (DL) is a branch of machine learning based on artificial neural networks. Many DL-based QSAR models have been reported, outperforming traditional machine learning methods. However, little has been published on DL-based proteochemometrics models.

In this work, we present the first proteochemometrics model which takes as protein descriptor the raw amino acid sequence, in order to capture its underlying biological information and long-range dependencies. For this, Long Short-Term Memory Networks are applied, a DL architecture specific for sequential data. The proposed model consists of an hybrid network merging the target and the ligand analysis blocks to give a classification on binding affinity as output. To build and test it, a curated benchmark of 51,404 ligand-target pairs from three different publicly available sources (ChEMBL, DUD and MUV) (Riniker, 2013) was randomly splitted 80/10/10 in training, validation and test sets, keeping an inactive/active imbalance as expected in a realistic setting.

The resulting model has an AUROC on test of 0.935 ± 0.003 and a BEDROC(α=20) of 0.952 ± 0.003. This performance is comparable to other published DL proteochemometrics models, but this architecture allows for further analysis on the underlying features related to binding affinity.

GUILDIFY V2.0: A TOOL TO IDENTIFY THE MOLECULAR NETWORKS UNDERLYING HUMAN DISEASES, THEIR COMORBIDITIES AND THEIR DRUGGABLE TARGETS

Joaquim Aguirre-Plans[1], Janet Piñero[2], Ferran Sanz[2], Laura I. Furlong[2], Narcis Fernandez-Fuentes[3], Baldo Oliva[1], Emre Guney[2]

1. Structural Bioinformatics Lab, Department of Experimental and Health Science, Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain
2. Integrative Biomedical Informatics Group, IMIM Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain
3. Department of Biosciences, Universitat de Vic-Universitat Central de Catalunya, Sagrada Familia 7, 08500 Vic, Catalonia, Spain

Presenter e-mail: joaquim.aguirre@upf.edu

The genetic basis of complex diseases involves alterations on multiple genes. Unravelling the interplay between these genetic factors is key to the discovery of new biomarkers and treatments. In 2014, we introduced GUILDify, a web server that searches for genes associated to diseases, finds novel disease-genes applying various network-based prioritisation algorithms and proposes candidate drugs. Here, we present GUILDify v2.0, a major update and improvement of the original method, where we have included protein interaction data for seven species and 22 human tissues and incorporated the disease-gene associations from DisGeNET. To infer potential disease relationships associated with multi-morbidities, we introduced a novel feature for estimating the genetic and functional overlap of two diseases using the top-ranking genes and the associated enrichment of biological functions and pathways (as defined by GO and Reactome). The analysis of this overlap helps to identify the mechanistic role of genes and protein-protein interactions in comorbidities. Finally, we provided an R package, guildifyR, to facilitate programmatic access to GUILDify v2.0 (http://sbi.upf.edu/guildify2).

CO-LOCATION OF PARALOGS IN TOPOLOGICALLY ASSOCIATING DOMAINS (TADS) MAY EXPLAIN WHY THE EFFECT OF SOME DELETERIOUS MUTATIONS IS SUPPRESSED

Natàlia Padilla[1], Xavier de la Cruz[1,2]
[1]Translational Bioinformatics, UAB, VHIR, [2]ICREA, Spain
[1]natalia.padilla@vhir.org

In recent years, a series of groundbreaking techniques have revealed a new layer of chromatin organization. Certain genomic regions which are highly self-interactive, organize the genome into local chromatin interaction domains, named Topologically Associating Domains (TADs). They are not only relevant from an organizational point of view; they also play a functional role, especially in the regulation of gene expression, changing the spatial proximity of genes and their cis-regulators.

In this work, we focused on studying the set of biomedically-relevant genes (BRG) and how they distribute along TADs, to see whether TAD-based regulation is a player that must be considered when deciphering the molecular origins of disease. We defined BRG as those genes for which at least one disease-causing mutation has been described in OMIM database. This produced a set of 5854 genes, which were mapped to a set of 3062 TADs (work of Dixon JR, et al 2012[1]). We found that 1664 TADs harboured at least one BRG and, more particularly, for a hundred TADs the number of BRG was higher than expected by chance.

Interestingly, in many cases the clustered BRG are paralogs, or come from the same protein family (in concordance with earlier work of Ibn-Salem J, et al 2017[2]). From a biomedical point of view, this is relevant because it may explain why sometimes a paralog compensates the effect of a coding deleterious variant present in another paralog of the same gene family. If both paralogs belong to the same TAD, we expect them to have a similar TAD-dependent gene expression regulation; therefore, co-expression of the undamaged paralog will compensate for the loss of function of the mutated paralog. We characterize in how many cases this can happen, by looking at the protein sequence divergence between paralogs, exploring their regulatory pattern and mapping the associated diseases. Finally, we complete our work by testing the robustness of this hypothesis to the dynamics changes of TAD boundaries across cell lines.

DYNBENCH3D, A WEB-RESOURCE TO DYNAMICALLY GENERATE BENCHMARK SETS OF LARGE HETEROMERIC PROTEIN COMPLEXES.

Bertoni M[1], Aloy P[2].

1. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain.
2. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain; Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain.

Multi-protein machines are responsible for most cellular tasks, and many efforts have been invested in the systematic identification and characterization of thousands of these macromolecular assemblies. However, unfortunately, the (quasi) atomic details necessary to understand their function are available only for a tiny fraction of the known complexes. The computational biology community is developing strategies to integrate structural data of different nature, from electron microscopy to X-ray crystallography, to model large molecular machines, as it has been done for individual proteins and interactions with remarkable success. However, unlike for binary interactions, there is no reliable gold-standard set of three-dimensional (3D) complexes to benchmark the performance of these methodologies and detect their limitations. Here, we present a strategy to dynamically generate non-redundant sets of 3D heteromeric complexes with three or more components. By changing the values of sequence identity and component overlap between assemblies required to define complex redundancy, we can create sets of representative complexes with known 3D structure (i.e., target complexes). Using an identity threshold of 20% and imposing a fraction of component overlap of <0.5, we identify 495 unique target complexes, which represent a real non-redundant set of heteromeric assemblies with known 3D structure. Moreover, for each target complex, we also identify a set of assemblies, of varying degrees of identity and component overlap, that can be readily used as input in a complex modeling exercise (i.e., template subcomplexes). We hope that resources like this will significantly help the development and progress assessment of novel methodologies, as docking benchmarks and blind prediction contests did. The interactive resource is accessible at https://DynBench3D.irbbarcelona.org.

ASSOCIATION BETWEEN BLOOD DNA METHYLATION AND GENE EXPRESSION IN CHILDREN

Ruiz Arenas, Carlos

DNA methylation is an epigenetic mechanism where a methyl group is added to cytosines placed in CG dinucleotides (CpGs). DNA methylation patterns can be modified by environmental exposures and these changes can eventually lead to diseases such as cancer or diabetes. Therefore, different epidemiological studies have performed a genome-wide evaluation of the methylation patterns using DNA methylation microarrays. However, typical epigenome-wide association analyses are usually difficult to interpret. These studies give a list of CpGs differently methylated but the effect of each individual CpG on gene expression is not known. To tackle this issue, we have used DNA methylation and gene expression microarray data from blood from 832 children of the Human Early Life Exposome (HELIX) project. We run 13,615,882 linear regressions between the CpGs and their nearby genes (0.5 Mb between CpG and gene transcription start site), adjusting for sex, age, cohort and cell type proportions. At Bonferroni correction, we found that 8,907 CpGs changed the expression of 3,790 genes, through 15,403 CpG -gene pairs. We identified some features that modify how a CpG regulates gene expression. Thus, CpGs that regulate gene expression tend to be placed in distal promoters and enhancers and but not in proximal promoters. These results will be available through a catalogue of cis expression quantitative trait methylation (cis eQTMs). All in all, this study sheds light on the regulation of gene expression during childhood and will help to interpret future epigenome-wide studies.

GENOMIC STRUCTURAL ALTERATIONS AS A DRIVING FORCE IN MPNST DEVELOPMENT

Bernat Gel[1,2], Ernest Terribas[1], Elisabeth Castellanos[1,2], Inma Rosas[1], Tapan Mehta[3,4], Peggy Wallace[5], Nancy Ratner[6], Ignacio Blanco[7], Juana Fernández-Rodríguez[8,2], Conxi Lázaro[8,2], Eduard Serra[1,2]

1 – Hereditary Cancer Group, Germans Trias i Pujol Research Institute (IGTP), Can Ruti Campus, Badalona, Barcelona, Catalonia, Spain

2 - CIBERONC

3 – Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, Alabama, USA

4 - Nutrition Obesity Research Center, University of Alabama at Birmingham, Birmingham, Alabama, USA

5 – Department of Molecular Genetics & Microbiology, University of Florida College of Medicine, Gainesville, Florida, USA

6 – Division of Experimental Hematology and Cancer Biology, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, Ohio, USA

7 - Clinical Genetics and Genetic Counseling Program, Germans Trias i Pujol Hospital, Can Ruti Campus, Badalona, Barcelona, Spain

8 - Hereditary Cancer Program, Joint Program on Hereditary Cancer, Catalan Institute of Oncology, IDIBELL campus, L'Hospitalet de Llobregat, Catalonia, Spain

Malignant peripheral nerve sheath tumors (MPNST) are soft tissue sarcomas with bad prognosis and lack of curative treatments. MPNST may arise from a preexisting benign plexiform neurofibroma (PNF), often after formation of a pre-malignant atypical neurofibroma (aNF).

We generated genomic structure (SNP-array), mutational (exome), transcriptomic (RNA-seq, microarray) and epigenomic (DNA methylation) data from a set of 15 MPNSTs, and also collected available data on MPNST, plexiform and atypical neurofibromas. We performed an integrative bioinformatic analysis to infer the mechanisms of MPNST development.

PNFs have no genomic structural alterations except those affecting chromosome 17q involved in the somatic inactivation of NF1. aNFs present also additional recurrent losses of the CDKN2A/B locus. Contrasting with these nearly-normal karyotypes, MPNSTs have hyperploid and highly rearranged genomes with somatic copy number alterations (SCNAs) affecting most chromosomes. However, MPNST genome structure is highly stable. In contrast to SCNAs, MPNSTs have a low number of point mutations, with no clear recurrently affected genes. Most point mutations are acquired after the genome reorganization.

This collective data suggest a model for MPNST origin, with a first progression towards a proliferative cell with reduced senescence due to the loss of *NF1* and *CDKN2A*/B, followed by a number of random catastrophic events of genomic alteration and the selection of a viable stable genomic combination.

Furthermore, SCNA have a profound impact on gene transcription levels and create regions with an accumulation of over- and under- expressed genes, transcriptional imbalances (TI). TIs mostly capture passenger gene expression but allow identification of genes with SCNA-independent expression regulation. The analysis of these genes provides insight into the biology of MPNSTs.

Gross genomic structural alterations are a driving force in MPNST biology and their genomic stability suggest a catastrophic event mediated by loss of senescence capacity as a probable origin.

MONITORING SCHWANN CELL DIFFERENTIATION IN AN IPSC-BASED MODEL THROUGH RNA-SEQ ANALYSIS

Miriam Magallón-Lorenz[1], Helena Mazuelas[1], Juana Fernández-Rodríguez[2], Yvonne Richaud-Patin[3], Imma Rosas[1], Ernest Terribas[1], Ignacio Blanco[4], Cleofé Romagosa[5], Elisabeth Castellanos[1], Ángel Raya[3], Conxi Lázaro[2], Meritxell Carrió[1], Eduard Serra[1] & Bernat Gel[1]

1- Hereditary Cancer Group, Germans Trias i Pujol Research Institute (IGTP)-PMPPC; Can Ruti Campus, Badalona, Barcelona, 08916; Spain

2- Hereditary Cancer Program, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona, 08098; Spain

3- Center of Regenerative Medicine in Barcelona (CMRB), Hospitalet de Llobregat, Barcelona, 08098; Spain

4- Clinical Genetics and Genetic Counseling Program, Germans Trias i Pujol University Hospital (HUGTiP),Can Ruti Campus, Badalona, Barcelona, 08916; Spain

5- Department of Pathology, Vall d'Hebron University Hospital (VHIR) Barcelona, Spain

Neurofibromatosis Type I (NF1) is an autosomal dominant disorder caused by heterozygous loss of *NF1*. Most NF1 patients develop neurofibromas, benign tumors of the peripheral nervous system caused by the complete inactivation of *NF1* in a cell of the Schwann cell (SC) lineage. Different types of neurofibromas exist. Plexiform neurofibromas (pNFs) have a risk of malignant transformation to a soft tissue sarcoma. In PNFs the inactivation of the *NF1* gene happens during development but the exact identity of the cell within the SC lineage receiving a second *NF1* hit is not known.

To study the differentiation process of SC, assess the impact of the *NF1* genotype in that process and to elucidate the neurofibroma cell-of-origin, we generated isogenic induced pluripotent stem cell (iPSC) lines from *NF1*-/- and *NF1* +/- pNF-derived primary cells. We then set-up a two-step differentiation protocol from iPSC to mature SC along the SC differentiation lineage using control *NF1*+/+ iPSCs.

In this work we analysed RNA-seq data from a SC differentiation time-course experiment. The analysis revealed robust expression levels and concordant expression of known markers along the SC lineage, validating the differentiation system.

Differential expression analysis between stages allowed us to identify new potential stage-specific markers including cell-surface markers. Enrichment analysis identified pathways and processes important for different stages of SC differentiation.

Comparison of *NF1*+/+ and *NF1*-/- cells revealed differences at the transcriptome level. We are currently performing co-expression analysis of known stage-specific genes and comparing our dataset with other relevant datasets such as regenerating nerve.

RNA-Seq analysis allowed us to confirm the robustness of our iPSC-based SC differentiation system and identify new stage-specific markers that could be used to better characterize the identity of the cells within pNFs. Further analyses are required to fully understand the impact of the *NF1* genotype on the differentiation process.

TAXONOMICAL AND FUNCTIONAL SIGNATURE FOR CROHN'S DISEASE

Pozuelo, Marta

Gut microbiota comprises bacteria, archaea, fungi and viruses that coexist along the gastrointestinal tract. Alterations in the composition of this gut microbiota, named dysbiosis, may lead to illnesses such as inflammatory bowel diseases (IBD). IBD is a chronic inflammatory disease divided in two subtypes: Crohn's disease (CD) and Ulcerative colitis (UC) with different clinical manifestations. Until now, studies relating gut microbiota to IBD, did not provide a microbial signature to discriminate CD or UC from healthy. In this study, we analyzed a Spanish cohort of CD (n=34) and UC patients (n=33) under remission and their healthy relatives (n=65), with a follow up of one year collecting fecal samples every 3 months until a flare appeared. We processed a total of 415 samples using 16S rRNA sequencing and 169 out of the 415 samples using shotgun sequencing for taxonomical and functional analyses. Based on 16S gene, we found a lower microbial diversity associated with a more unstable microbial community in CD patients compared to UC and healthy controls (HC). We proposed a microbial signature composed of bacterial genera, discriminating CD from HC and UC patients. Regarding shotgun sequencing, we also found greater dysbiosis in CD than in UC and HC with higher compositional similarity between UC and HC. Furthermore, at functional level, a similar trend was encountered showing significant differences that discriminate CD from UC or HC for functional categories such as carbohydrate metabolism or folding and sorting degradation. We did not find significant differences in functional categories between UC and healthy.

Our findings showed that CD could be differentiated from UC and HC based on 16S rRNA gene (taxonomically) as well as on shotgun data (taxonomically and functionally).

CNV SCREENING FROM NGS DATA IN ROUTINE GENETIC DIAGNOSTICS

José Marcos Moreno-Cabrera[1, 2, 3] (jmoreno@igtp.cat), Jesús del Valle[2,3], Eli Castellanos[1,3], Lidia Feliubadaló[2,3], Marta Pineda[2,3], Eduard Serra[1,3], Gabriel Capellà[2,3], Conxi Lázaro[2,3] and Bernat Gel[1,3]

[1] Hereditary Cancer Group, Program for Predictive and Personalized Medicine of Cancer - Germans Trias i Pujol Research Institute (PMPPC-IGTP), Campus Can Ruti, Badalona, Spain
[2] Hereditary Cancer Program, Joint Program on Hereditary Cancer, Catalan Institute of Oncology, IDIBELL campus in Hospitalet de Llobregat, Spain
[3] CIBERONC, Instituto de Salud Carlos III, Madrid, Spain

Next Generation Sequencing (NGS) is a key technology for detecting small variants in the genetic diagnostics of hereditary diseases. However, detection of larger variants as copy number variants (CNV) from NGS data remains a challenge. The gold standard for CNV detection in a genetic diagnostic setting is multiplex ligation-dependent probe amplification (MLPA), but the cost and time to perform it are drawbacks when implementing it in the diagnostics routine.

Tools adapted to detect CNVs from NGS panel data at single exon resolution have been recently published. The aim of this work is to identify and implement in the diagnostics routine a tool suitable to be used as a screening method prior to MLPA validation.

We selected five algorithms for performance evaluation: DECoN, CoNVaDING, panelcn.MOPS, ExomeDepth and CODEX2. These algorithms were evaluated over four cancer panel datasets from different sources (2 in-house, 2 external) for a total of 495 samples with 231 exon and multi-exon validated CNVs. The parameters for each algorithm were also optimized to maximize sensitivity. Code developed to perform the benchmark, CNVbenchmarkeR, is publicly available to help other labs to evaluate algorithms performance.

CoNVaDING, DECoN and panelcn.MOPS achieved 100% sensitivity over different datasets, with specificity ranging from 88.9% to 97.8%. DECoN performed the best in the in-house datasets, with 100% sensitivity and 93.3% specificity. Therefore, DECoN has been adopted as a screening method prior to MLPA validation in our I2HCP genetic diagnostics strategy for hereditary cancer. The inclusion of the screening step allowed us to increase the number of genes routinely tested for CNV, improving diagnostics yield. In addition, in the retrospective analysis of three previously untested genes over 1824 samples, we have identified, validated and reported 6 additional novel CNVs.

CONVOLUTIONAL DEEP NEURAL NETWORK BASED ANALYSIS OF ELECTROCARDIOGRAM SIGNALS

G. Riera-Solà[1], J. Garrido-Aguirre[2,3], A. López-del Río[2,3,4], A. Kravtsov[2], P. Caminal[2,3], P. Gomis[2,3], A. Perera-Lluna[2,3], J. Fonollosa[2,3]
1 Facultat de Matemàtiques i Estadística, Universitat Politècnica de Catalunya, Barcelona, Spain
2 Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial ESAII, Universitat Politècnica de Catalunya, Barcelona, Spain
3 Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain
4 Mind the Byte S.L., Barcelona, Spain

Presenter e-mail: jon.garrido@upc.edu

Electrocardiography (ECG) provides a simple, non-invasive, and inexpensive procedure widely used to diagnose heart diseases and obtain information regarding the heart condition. Over the years, large variety of automatic methods have been proposed for arrhythmia detection and classification. These methods rely on the extraction of features from the captured signals and require pre-processing steps for baseline removal, noise reduction, etc. Due to the nature of the Deep Convolutional Learning techniques, these techniques can circumvent customized feature extraction through learning a representation of the input.

In this work, we present an automatic classification model that does not require the definition of features beforehand. Our model is based on an autoencoder that includes two convolutional layers. A softmax activation is then used for beat classification. The model was trained and tested with a widely-used publicly available database, MIT-BIH.

Results showed that the proposed network successfully coded the heartbeats at a lower non-linear embedding, from which the original beats can be reconstructed. Our model showed good classification sensitivity compared to other classifiers that rely on signal pre-processing and the extraction of handcrafted features. Our results show that Deep Learning can be of utmost interest for the real-time classification of heartbeats and the detection of arrhythmia.

CHARACTERIZATION OF HIGH RISK NEUROBLASTOMA GROUP. POTENTIAL EPIGENETIC BIOMARKERS FOR TREATMENT RESPONSE ASSESSMENT

Alícia Garrido-Garcia[1], Soledad Gómez[1], Sara Pérez-Jaume[1], Laura Garcia-Gerique[1], Mariona Suñol[2], Gemma Bande[1], Maria Jesús Nagel[1], Jaume Mora[1], Cinzia Lavarino[1].
[1] Developmental Tumor Biology Laboratory, Hospital Sant Joan de Déu, Institut de Recerca Sant Joan de Déu, Barcelona, Spain.
[2] Department of Pathology, Hospital Sant Joan de Déu, Barcelona, Spain.

Presenter e-mail: agarridog@fsjd.org

Neuroblastoma (NB), the most frequently occurring solid pediatric tumor, accounts for 15% of cancer-related deaths in childhood. Probability of cure varies according to patient's age, extent of disease and tumor biology. High-risk NB tumors are chemotherapy-refractory tumors with low survival rates. These constitute a heterogeneous group of tumors, whereby patients can display response to treatment and long-term outcome or develop early progressive disease with poor outcome (ultra high-risk (UHR-NB). Genetics underlying this divergent clinical behavior is still greatly unknown. No reliable biomarkers for early treatment response assessment have been reported for high-risk NB.
Aim: We aim to define and characterize UHR-NB subgroup using DNA methylation and gene expression profiling, and investigate the effects of core (epi)genetic alterations on regulatory pathways. Thereby, identify potential biomarkers and therapeutic targets.
Methods: We analyzed DNA methylation and gene expression datasets of more than hundred high-risk NBs. Methylation data was annotated according to gene location, CpG islands, chromatin state categories. Bisulfite sequencing and pyrosequencing was performed for validation purposes. Cox-regression models were used to identify biomarkers. Survival curves were analyzed by Kaplan-Meier method and log-rank test.
Results: We observed two differential DNA methylation patterns within high-risk NB group. These methylation profiles were associated with divergent clinical evolution, clearly defining a group of patients with rapidly progressing, chemo-refractory tumors. Genomic functional analysis is ongoing. The varying methylation levels were not translated into gene expression, suggesting they may not have a functional impact. By analyzing and filtering DNA methylation data, we identified a reduced set of differentially methylated cytosines that defined UHR-NBs. Validation is currently ongoing on an independent cohort of NBs.
Conclusion: Our findings show that UHR-NB is defined by specific DNA methylation changes. We have identified a reduced set of biomarkers that represent a potential epigenetic classifier for this aggressive, chemoresistant subgroup of NB tumors.

DESIGN, DEVELOPMENT AND DEPLOYMENT OF A DJANGO-BASED PLATFORM FOR AUTOMATIC MEDULLOBLASTOMA STRATIFICATION FROM DNA METHYLATION PROFILES

Joshua Llano[123], Ander Sainz[123], Soledad Goméz[4], Cinzia Lavarino[4], Alexandre Perera-Lluna[123]

[1]B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain; [2]Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain; [3]Institut de Recerca Pediàtrica Hospital Sant Joan de Dèu, Esplugues de Llobregat, Barcelona, Spain; [4]Developmental Tumor Biology Laboratory, Institut de Recerca Sant Joan de Déu, Barcelona, Spain;

ABSTRACT

Background: Bioinformatics approaches are evolving rapidly and offer a wide range of tools and predictive models for a vast amount of human diseases and experimental platforms. However, the transfer and dissemination of these algorithms to a clinician audience can be challenging. For this purpose, we created PECO, a web application that allows bioinformaticians to deploy state-of-the-art classification models throughout a modular scheme. Clinicians without a computational specialization can apply the models in their clinical research through a user-friendly interface.

Method: PECO was programmed in Django, a web development Python module. PECO is scalable, modular and secure through a user authentication system. User data uploads are handled with MariaDB, a widely used open-source database service. On the other hand, PECO supports statistical models programmed in R using RPy2, in order to integrate the R Bioconductor repository and add versatility. Celery was selected as an asynchronous task queue, decoupling data processing tasks from the PECO frontend.

Results: PECO was used to deploy a multi-class classifier of the molecular subgrouping of medulloblastoma tumor cells. Clinicians can log in and upload DNA methylation data of their patients on three different input types: DNA methylation microarray, bisulfite pyrosequencing, and direct-bisulfite sequencing. The sample is processed under a minute and a report provides descriptive statistics and the predicted probabilities of each tumor class. Moreover, the researchers hold a private admin site where they can upload new algorithms and choose the active ones.

ONCODRIVECLUSTL: A SEQUENCE-BASED CLUSTERING METHOD TO IDENTIFY CANCER DRIVERS

Claudia Arnedo-Pac[1], Loris Mularoni[1], Ferran Muiños[1], Abel Gonzalez-Perez[1], Núria López-Bigas[1,2]

1. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain
Presenter e-mail: claudia.arnedo@irbbarcelona.org

The characterization of the genomic alterations driving tumorigenesis is one of the main goals in oncogenomics research towards the implementation of precision cancer medicine. Given the evolutionary behavior of cancer, computational methods based on signals of positive selection have been effectively applied, including clustering of somatic mutations. Remarkably, clustering has been shown to complement recurrence and functional impact signals in the detection of driver genes (Tamborero et al., 2013; Porta-Pardo et al., 2017). New clustering methods need, first, accurate background models reflecting the state-of-the-art knowledge in mutation rates; second, novel approaches in the measurement of clustering signals that enable us to expand drivers identification to the entire genome. To this aim, we have developed OncodriveCLUSTL, a new nucleotide sequence-based clustering algorithm to detect significant clustering signals in genomic regions. OncodriveCLUSTL is based on a local background model derived from the nucleotide context composition of the cohort under study. Using TCGA exome sequencing cohorts, our method is able to identify *bona fide* cancer drivers, outperforming the existing OncodriveCLUST algorithm (Tamborero, Gonzalez-Perez, y Lopez-Bigas, 2013) and complementing other methods based on different signals of positive selection. We also show the applicability of OncodriveCLUSTL to non-coding regions using whole genome sequencing data.

Porta-Pardo,E. et al. (2017) Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nat. Methods*, 14, 782-788.
Tamborero,D., et al. (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.*, 3, 2650.
Tamborero,D., Gonzalez-Perez,A., y Lopez-Bigas,N. (2013) OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29, 2238-2244.

ANALYSIS OF COMMUNICATION NETWORKS IN CYTOCHROME C OXIDASE

Anne Sophie Hartmann[1,2,3,4,*], Alexandre Perera Lluna[2,3,4], Petra Imhof[1]
[1]Institute of Theoretical Physics, Freie Universität Berlin, Germany
[2]B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain
[3]Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid,Spain
[4]Institut de Recerca Pediàtrica Hospital Sant Joan de Dèu, Esplugues de Llobregat, Barcelona, Spain

An important part of the respiratory chain enabling aerobic respiration is the electron-pumping enzyme Cytochrome c Oxidase (CcO). Research has shown each proton enters the protein through one of two channels called the D- and K-channel. These are hypothesized to depend on the protonation of key residues D132, E286 (D-channel), and K362, E101 (K-channel) indicating communication between residues. Communication between protein residues in different metrics can be expressed using communication graphs.

Communication graphs using a hydrogen bond metric and a generalized correlation metric for different protonations of key residues in the PR redox state of CcO were analyzed. A modified version of the Manhattan distance was used to define a fast and simple similarity measure on pairs of graphs sharing a node set.

Furthermore, a diffusion-based label propagation algorithm implemented in the R package diffuStats was used to investigate the communication of key residues inside the protein. A linear model relating diffusion score and protonation of the key residues was constructed for each node to examine the dependence of communication on the protonation state.

The similarities of all possible protonation state pairs were compared, showing that pairs where E286 or K362, respectively, are unprotonated in both states tend to be more similar. Comparison of graphs considering only selected D-channel residues showed strong dependence on the protonation of E286.
These observations are hypothesized to be due to the opening/closing of the channels.

* annehartmann@physik.fu-berlin.de

THE CHALLENGE OF PHYLOGENETIC INFERENCE IN VERY OLD GENE FAMILIES: A COMPUTER SIMULATION STUDY

Paula Escuer, Julio Rozas & Alejandro Sánchez-Gracia

Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Av. Diagonal 643, Barcelona 08028, Spain.

One of the most frequent approaches for studying the origin and evolution of gene families is determining the phylogenetic relationships across family members. These phylogenies allow characterizing the diversification pattern, classifying them into subfamilies, or even establishing possible functional homologies from phylogenetic groupings. Nevertheless, in the analysis of very old families, the inconsistence and uncertainty introduced in the multiple sequence alignments by distantly related copies certainly affects phylogenetic inference. To better understand the effects of introducing highly divergent sequences in these phylogenetic analyses, we carried out a computer simulation study of the performance of a range of tree inference methods, including conventional MSA-based (MAFFT or hmmalign coupled with FastTree, RAxML or Q-TREE) and alignment-free methods (PaHHM-Tree, Sachmo, ACS and LZ). We simulated gene family trees using divergence times and turnover rates similar to those estimated for the oldest arthropod chemoreceptor families (gustatory and ionotropic receptors); these families originated long before the diversification of major arthropod lineages (Hexapoda, Myriapoda and Chelicerata; > 600 Mya), and show a highly dynamic birth-and-death evolution. We then generated pseudo-observed sequences emulating family members by evolving amino acid sequences across these trees. For the analysis, we used standard tree topology distance metrics to compare the simulated (real) topologies and those inferred by the different benchmarked methods from simulated sequences. We found that although maximum likelihood methods based on MSA outperform alignment-free methods, none of tested approaches was capable to correctly reproduce real topologies, especially for internal tree branches. These findings are very relevant for researchers interested in understanding the origin, evolution and function of old gene families and highlight the necessity of developing new, more suited approaches to address these important questions.

ANALYSIS OF GERMLINE *DE NOVO* MUTATION RATES ON EXONS AND INTRONS

Miguel Rodriguez-Galindo[1,2*], Sònia Casillas[1,3], Antonio Barbadilla[1,3]

[1]Institute of Biotechnology and Biomedicine (IBB), Autonomous University of Barcelona (UAB), 08193 Cerdanyola del Valles, Barcelona, Spain
[2]Present address: Center for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain
[3]Department of Genetics and Microbiology, Autonomous University of Barcelona (UAB), 08193 Cerdanyola del Valles, Barcelona, Spain

Presenting author: miguel.rodriguez@crg.eu

A main assumption underlying molecular population genetics is that the mutation rate is constant across close genic regions, that is, it does not depend on sequence function. Challenging this assumption, a recent study has found a reduction in the mutation rate in exons compared to introns in somatic cells due to an enhanced exonic mismatch repair system activity. If this reduction happens also in the germline, it can compromise studies of population genomics, including the detection of the footprint of selection.

We compiled and analyzed ten whole-genome germline mutation datasets from recent studies to consolidate a medium-density but high-quality *de novo* mutation map across the human genome. Then we tested the possible reduction in the exonic germline mutation rates through the different approaches that were employed to analyze somatic cells. We found no reduction in the mutation rate in exons compared to introns in the germline, in contrast to what has been previously described in somatic cells. Therefore, there is no evidence of an enhanced mismatch repair system activity in exons with respect to adjacent introns in germline cells.

Our findings may point to a dichotomous nature of germline and soma with respect to mutational and DNA repair processes.

PPAXE FACILITATES HUMAN-CURATION OF PROTEIN-PROTEIN INTERACTIONS FILTERED FROM THE SCIENTIFIC LITERATURE

S. Castillo-Lara, J.F. Abril
Computational Genomics Lab;
Genetics, Microbiology & Statistics Dept.;
Universitat de Barcelona; Institut de Biomedicina (IBUB).
Barcelona, Catalonia, Spain.

Protein-protein interactions (PPIs) are crucial to build models for understanding many biological processes. Although several databases hold many of these interactions, exploring them, selecting those relevant for a given subject, and contextualizing them can be a difficult task for researchers. Extracting PPIs directly from the most recent scientific literature sources can be very helpful for providing such context, as the sentences describing these interactions may give insights to researchers in helpful ways.

We have developed a python module and a web application, PPaxe, that allows users to extract PPis and protein occurrence from a given set of PubMed and PubMed Central articles, based on abstracts and full-texts respectively. PPaxe tokenizes and annotates the components of the sentences with StanfordCoreNLP and then distills a number of features that are analyzed by a Random Forest classifier (trained over AImed, LLL-challenge and BioInfer curated datasets).

Finally, it presents the results of the analysis in different ways to help researchers export, filter and analyze the interactions retrieved easily. Among its outputs, users can play with an interactive graph built from the reconstructed interaction network.

PPaxe web demo is freely available at https://compgen.bio.ub.edu/PPaxe. Further materials can be found on the tool's website.

QSUTILS: AN R PACKAGE TO STUDY VIRAL QUASISPECIES COMPLEXITY WITH NGS DATA.

Mercedes Guerrero-Murillo[1], Josep Gregori i Font[1,2,3], Francisco Rodríguez-Frias[2,4],Josep Quer Sivila[1,2]

[1]Liver Unit, Internal Medicine, Liver Disease Laboratory, Vall d'Hebron Institut Recerca-Hospital Universitari Vall d'Hebron (VHIR-HUVH), Universitat Autònoma de Barcelona, 08035 Barcelona, Spain
[2]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd) del Instituto de Salud Carlos III, 28029 Madrid, Spain
[3]Roche Diagnostics, Sant Cugat del Vallès, Spain
[4]Liver Pathology Unit, Departments of Biochemistry and Microbiology, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona (UAB), Barcelona, Spain

Summary: RNA and DNA viruses that replicate by low fidelity polymerases generate viral quasispecies, a collection of closely related viral genomes. This variability contributes greatly to the adaptive potential of the virus. We define complexity of a viral quasispecies as the intrinsic property that quantifies the diversity and frequency of haplotypes, independently of the population size that contains them. Quasispecies complexity can describe viral behaviour by predicting viral disease progression and/or response to treatment; hence, it has an obvious interest for clinical reasons. The complexity can be estimated through diversity indices, which may be classified as incidence-based, focused on the number of observed entities irrespective of their abundances; abundance-based, indices that take into account the observed or estimated abundance of each entity and; functional, that are computed on differences among traits of the observed entities. Part of the diversity indices are adapted from ecology.

We developed and R package, that we have named "QSutils", intended for use with quasispecies data obtained by next-generation sequencing (NGS) of highly mutated viral populations. QSutils offers a set of utility functions for viral quasispecies analysis, there are three main types: (1) data manipulation and exploration: functions useful for converting reads to haplotypes and frequencies, repairing reads, intersecting strand haplotypes, and visualizing haplotype alignments; (2) diversity indices: functions to compute diversity and entropy, in which incidence, abundance, and functional indices are considered; (3) data simulation: functions useful for generating random viral quasispecies data.

Availability: https://github.com/VHIRHepatiques/QSutils

Contact: mercedes.guerrero@vhir.org

ASSESSING THE PERFORMANCE OF NGS BIOINFORMATICS TOOLS FOR THE DETECTION OF SOMATIC MUTATIONS IN AUTOINFLAMMATORY DISEASES

M. Solis-Moruno*[1,2], A. Mensa-Vilaro[3], L. Batlle-Maso[1,2], T. Marques-Bonet[1], JI. Arostegui[3], F. Casals[2]

[1] Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Doctor Aiguader 88, Barcelona, Spain

[2] Genomics Core Facility, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, 08003 Barcelona, Spain

[3] Department of Immunology, Hospital Clínic-IDIBAPS, Barcelona, Spain

*manuel.solis@upf.edu

Introduction: Autoinflammatory diseases represent a privileged scenario for the study of somatic variation. First, somatic mutations are suspected to cause autoinflammatory diseases in a considerable fraction of patients. Second, DNA from blood is easily obtained. And third, separating different cell populations is possible with standard flow cytometry procedures. However, detection of somatic mutations from NGS presents some difficulties. Low frequency somatic mutations are at risk of being undetected. Also, most of the algorithms for somatic mutation analyses are optimized for cancer studies, where a tumour sample is compared with the healthy tissue. Thus, higher coverages and the modification of experimental designs and pipelines are needed to detect these variants.

Materials and Methods: We performed whole exome sequencing (WES) to a mean coverage ~245X in a total of 16 different samples belonging to 10 different individuals. We have matched samples from peripheral blood and oral mucosa for 5 of them and, for one of, from peripheral blood and urine.

Results: We found 15 out of 16 variants, all but one at low frequency (2.8%). All others (ranging from 2.7%-31.3%) were detected by, at least, both VarDict and VarScan2. Estimated frequencies by experimental methods and by the NGS approach were similar. Conclusion: Next generation sequencing is a tool with potential application to detect somatic mutations in autoinflammatory diseases.

ANALYSIS OF VIROLOGIC OUTCOME MEASURES DURING ANALYTICAL TREATMENT INTERRUPTIONS IN CHRONIC HIV-1 INFECTED PATIENTS

Csaba FEHÉR[1,2,3,§], Lorna LEAL[2,3], Monserrat PLANA[3], Nuria CLIMENT[3], Alberto CRESPO GUARDO[3], Esteban MARTÍNEZ[2,3], Pedro CASTRO[2,3,4], Vicens DÍAZ-BRITO[2], Beatriz MOTHE[5.6,7], Juan Carlos LÓPEZ BERNALDO DE QUIRÓS[8,9], Josep María GATELL[2,3,10], Felipe GARCÍA[2,3], Patrick ALOY[1,11]

1 Institute for Research in Biomedicine (IRB Barcelona), the Barcelona Institute for Science and Technology, Barcelona, Spain.
2 Infectious Diseases Department, Hospital Clinic of Barcelona – HIVACAT, University of Barcelona, Barcelona, Spain.
3 Retrovirology and Viral Immunopathology Laboratory, AIDS Research Group, August Pi i Sunyer Biomedical Research Institute (IDIBAPS) – HIVACAT, Hospital Clínic of Barcelona, University of Barcelona, Barcelona, Spain
4 Medical Intensive Care Unit, Hospital Clinic of Barcelona, Barcelona, Spain
5 IrsiCaixa AIDS Research Institute, Badalona, Spain
6 Infectious Diseases Department, Hospital Germans Trias i Pujol, Badalona, Spain
7 University of Vic – Central University of Catalonia, Barcelona, Spain.
8 HIV/Infectious Diseases Unit. Hospital General Universitario Gregorio Marañón. Madrid. Spain
9 Instituto de Investigación Sanitaria Gregorio Marañón. Madrid. Spain
10 Senior Global Medical Director. ViiV Healthcare. Barcelona. Spain
11 Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain.

Controlled interruptions of antiretroviral therapy (ART) – a.k.a. analytical treatment interruptions (ATI) – are an inherent part of research on immune therapies against HIV. However, there is no widespread agreement on a "gold standard" parameter that best defines responders to the immunological interventions studied. Our aims were to analyze the most commonly used virological end-point parameters, to establish correlations between them, to evaluate possible confounding factors, and to propose a resuming parameter candidate for future ATI studies.

We carried out a retrospective analysis of 334 ATI episodes in 249 HIV-1 infected patients. We analyzed quantitative [baseline viral load (VL), set point, delta set point, VL and delta VL at given weeks after ATI, peak VL, delta peak VL, and the area under the rebound curve (AUC)], and temporal parameters [time to rebound (TtR), set point, peak, and certain absolute and relative VL thresholds]. Potential confounding factors evaluated were sex, age, number of previous ATIs, time of known HIV infection, time on ART, and immunological interventions.

Most patients had detectable VL by week 6. Median TtR was 2 weeks and median time to set point was 8 weeks. The drop of VL at set point with respect to baseline values was >1 log10 copies/mL in 13.9% of the cases. TtR and baseline VL were correlated with most temporal and quantitative parameters. According to a multivariate analysis baseline VL and the use of immunological interventions were factors independently associated with TtR.

We concluded that TtR could be an optimal surrogate marker of response in ATI studies, since it is an early parameter that correlates well with the majority of other virological end-points. Caution is warranted in the use of delta set point as an efficacy end-point in single-arm ATI studies. Baseline VL should be taken into account in the design of future ATI studies.

COMPARATIVE GENE EXPRESSION HIGHLIGHTS THE IMPORTANCE OF AGE-RELATED INFLAMMATORY RESPONSE IN ALZHEIMER'S DISEASE MOUSE MODELS

Sergi Bayod[1], Eduard Pauls[1], Teresa Juan-Blanco[1], Samira Jaeger[1], Miquel Duran[1], Camille Stephan-Otto[1], Víctor Alcalde[1], Patrick Aloy[1,2]

[1]Institute for Research in Biomedicine (IRB Barcelona), Parc Científic de Barcelona, Baldiri Reixac 10, 08028 Barcelona, Spain
[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

Presenter e-mail: sergi.bayod@irbbarcelona.org

Alzheimer's disease (AD) is the most prevalent form of dementia, which affects around 47 million people worldwide. Besides amyloid-ß (Aß) and tau deposition, a growing body of evidence shows that oxidative stress, vascular dysfunction and inflammatory processes are also involved in the development of the disease. It is known that these processes lead to gene expression changes. However, how these changes contribute or result as a consequence of the disease are poorly understood. Here, we performed a comparative gene expression profiling from the hippocampus of three AD mouse models (3xTg, NL-F and NL-G-F) at various ages representative of AD clinical stages. Using RNA-Seq technology, we estimated the expression of 131021 unique transcripts that encode for 52551 genes, 21950 out of them are protein-coding genes. Our data indicate that the lysosome system, osmotic stress and the spliceosome are affected at early stages of the disease. Later, the accumulation of Aß in the mouse hippocampus triggers the activation of immune response (innate response, complement and inflammation), the increment of microglial and astrocyte markers or apoptosis. Genes involved in neurotransmitters-dependent synapses (dopaminergic, glutamatergic or cholinergic synapses) are downregulated along with AD progression. On the other hand, the NL-G-F model develops an intense neuroinflammatory response linked to Aß accumulation earlier (from 3 months onwards) than the other mouse models. Our results provide a comprehensive gene expression dataset of three AD mouse models, which helps to identify new targets and to propose new therapeutic opportunities with the aim to stop or slow down the progression of the disease.

RNA-SEQ ON NON-MODEL ORGANISMS

Authors: Jorge Espinosa[1], Jordi Durban[3], Vincent L. Viala[2], Juan J. Calvete[3].
jdurban@ibv.csic.es
[1.] Departament de Genètica, Universitat de València
[2.] Instituto Butantan, Brazil
[3.] Laboratori Venòmica Evolutiva i Traslacional, Institut Biomedicina València (CSIC)

Snakebite envenoming is a neglected tropical disease and a major public health problem in many tropical and subtropical countries that disproportionately affect lower socioeconomic segments of society. It occurs in at least 1,8 - 2,7 million people worldwide which cause 81,000 – 138,000 deaths per year.

Venoms research has been continuously enhanced by advances in technology. In particular, the emergence of 'omic' technologies at the turn of the twenty-first century has revolutionized biological research. The breakthrough experienced by venom research in the last decade is due to the development and application of omics technology to the qualitative and quantitative profiling of the venom gland mRNA expression (venom gland transcriptomics) and the precise identification of the components expressed in the venom (venomics). Both approaches have been fueled by advances on sequencing technologies and improvements on **RNASeq**. However, in the absence of a genomic reference, several approaches have been developed in order to analyze mRNA expression data from Next Generation Sequencing technologies which are not as straightforward as those developed when a genome reference sequence is available.

In the present work, we evaluated the performance of the *de novo* RNASeq assembly pipeline of the *Evolutionary and Translacional Venomics lab* and provided hints about the integration of third generation sequencing technology in order to solve the structural organization of some genes expressed in the venom gland.

EMIRNA: A COMPREHENSIVE PIPELINE FOR DISCOVERY AND ANNOTATION OF MICRORNAS IN ANIMAL SPECIES

Emilio Mármol-Sánchez[1*], Susanna Cirera[2], Marcel Amills[1]
[1]Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Universitat Autònoma de Barcelona, Bellaterra 08193.
[2]Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark.

*Corresponding author: emilio.marmol@cragenomica.es

MicroRNAs (miRNAs) are a group of small non-coding RNAs widely recognized as key post-transcriptional regulators in many relevant biological processes. Although the functional role of several miRNAs has been thoroughly studied in humans and rodents, other relevant species still lack an accurate and complete miRNA annotation. This lack of knowledge makes it difficult to infer the functional roles of these miRNAs as well as to establish the list of putative target genes.
Computational prediction of novel miRNA candidates by applying Machine Learning algorithms constitutes an emerging and active research field for improving miRNA annotation. In the present study we have developed a new comprehensive pipeline for pre-miRNA recognition and homology comparison between species. A positive set of pre-miRNAs and negative hairpin microRNA-like non-coding sequences were selected from 10 relevant animal species, followed by sequence, structural and statistical features extraction for training a Support Vector Machine (SVM) classifier. The performance of this pipeline with regards to detecting novel miRNA sequences in the porcine genome was subsequently tested. The extensive list of human miRNAs that is currently available was used to identify orthologous non-annotated pre-miRNA candidates in the porcine genome based on sequence alignment and genome topology similarities.
By using this pipeline, we were able to identify 53 novel pre-miRNA candidates in the porcine genome, 19 of which were detected in smallRNA-Seq data for *gluteus medius* skeletal muscle in 48 sequenced Duroc pigs. Twelve of these candidates were selected for further verification using miRspecific qPCR in samples from liver and *longissimus dorsi* muscle from 7 Göttingen minipigs. A total of 9 miRNAs (ssc-miR-483-3p, ssc-miR-219a-3p, ssc-miR-10395-3p, ssc-miR-200a-3p ssc-miR-302d-3p, ssc-miR-6516, ssc-miR-502-3p, ssc-miR-484-5p and ssc-miR-3613-3p) were profiled in these tissues confirming the prediction of our pipeline.
The eMIRNA pipeline demonstrated good performance and applicability on pre-miRNA detection in multiple animal species and allowed annotation of relevant novel miRNAs in the porcine genome.

TRANSLATIONAL DATA SCIENCE IN PEDIATRIC ONCOLOGY: A PRACTICAL EXAMPLE

Soledad Gómez, Alicia Garrido, Laura Gerique, Isadora Lemos, Sara Pérez-Jaume, Mariona Suñol, Jaume Mora, Cinzia Lavarino.

Email: sgomezg@fsjd.org

The primary aim of our group is to translate research rapidly into routine clinical practice. To this end, we have developed an optimized workflow that starts from Data Science across bench to bedside. Here, we describe an example of our research in Medulloblastoma (MB), the most common malignant pediatric brain tumor.
Recent integrated genomic studies have identified four core subgroups of MB with distinctive clinicopathological and molecular features that proved to be associated with clinical outcome. These molecular subgroups exhibit distinctive transcriptional and epigenetic features that define relevant clinical subgroups of patients. The clinical relevance and utility of the molecular classification has increased, not only for refining patients' prognosis but also for the discovery of therapeutic targets and the design of clinical trials.
We have developed a novel approach for subgroup classification of MB tumors based on a reduced panel number of epigenetic biomarkers. The specific characteristics of our panel enable the analysis using diverse molecular techniques that can be implemented in clinical diagnostic practice.
The DNA methylation-signature was defined by re-analyzing 1,576 samples (913 MB, 457 non-MB tumors, 85 normal tissues) comprising previously published DNA methylation microarray data (HumanMethylation BeadChip450K, HM450K). A first *in silico* validation of the classifier was performed using published HM450K data sets. On bench, we tested our classifier on a cohort of 121 DNAs from MB samples (DNA from fresh frozen and FFPE samples) using diverse molecular techniques. The results showed that our classifier is robust, accurate and reproducible, and it can be applied to DNA from both frozen and FFPE MB specimens using different molecular approaches. These features make it technically simple, easy to interpret and a rapid way for the classification of patients with MB. Our approach represents a potentially practical and cost-effective epigenetic classifier for most centers treating children with brain tumors.

RPGENET V2 – ENHANCED NAVIGATION THROUGH THE RETINITIS PIGMENTOSA INTERACTION NETWORK

Rodrigo Arenas-Galnares, Sergio Castillo-Lara, Vasileos Toulis,
Gemma Marfany, Roser Gonzàlez-Duarte, Josep F. Abril

Genetics, Microbiology and Statistics Dept. and the Institute of Biomedicine (IBUB); University of Barcelona; Av. Diagonal 643, 08028, Barcelona, Catalonia, Spain

Retinitis pigmentosa (RP) is a genetic visual disorder characterized by the death of photoreceptor cells that affects 1 in every 4000 people around the world, although molecular causes are still unknown for 60% of diagnosed patients. RP is highly heterogeneous with over 200 known genes associated to the disease and requires new bioinformatic tools to assist in proper diagnosis and in research of the disorder.
RPGeNet (Boloc et al, 2015) is a curated interaction network interface developed on top of a skeleton network, which was defined by the shortest paths between all the known genes associated with RP. This interface can handle expression and variation data from different sources in order to provide further evidences to propose novel candidate causative genes located within paths connecting driver genes.
RPGeNet.v2 relies on top of a new graph database engine, running a neo4j manager, that facilitates fast queries across different interaction levels derived from the skeleton network. Furthermore, this new design allows new types of queries, such looking for the relationships with driver genes for any gene contained in the whole interaction network that integrates data from literature searches (via PPaxe), and interactions from BioGRID and STRING high-throughput databases.

METABOWISE: CONTEXT-BASED ANNOTATION OF LC-MS FEATURES THROUGH DIFFUSION IN GRAPHS

Maria Barranco-Altirriba[123], Pol Solà-Santos[123], Sergio Picart-Armada[123], Alexandre Perera-Lluna[123]

[1]B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain;
[2]Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain;
[3]Institut de Recerca Pediàtrica Hospital Sant Joan de Dèu, Esplugues de Llobregat, Barcelona, Spain;

Background: Liquid Chromatography coupled to Mass Spectrometry (LC-MS) is a widely used analytical platform for untargeted metabolomics, a global profiling of small molecules in biological samples. LC-MS data workflow consists of a set of preprocessing methods followed by peak-to-metabolite annotation. Metabolite annotation is still considered a major bottleneck, due to the inability to prioritize between metabolite candidates for a given peak. To that end, we introduce MetaboWISE (Metabolomics Wise Inference of Speck Entities), a context-based annotation algorithm for LC-MS data leveraging biochemical networks to prioritize metabolite annotation in terms of biological plausibility.

Methods: MetaboWISE is a two-stage algorithm. First, a matching algorithm queries peak mass-to-charge ratio values to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database with the possibility of applying an adduct filter that only keeps quasi-molecular adducts. The second stage prioritizes the candidate metabolites using a mass-based precision filter, an adduct probability filter and diffusion scores. Diffusion is applied on the unique peak-metabolite annotations in a biochemical metabolite network, based on the KEGG Reactant pair (RPair) annotations. MetaboWISE is validated on a public dataset (MetaboLights, MTBLS20) that characterizes the adult human urinary metabolome.

Results: When applying diffusion to the dataset filtered by the adducts likelihood, the diffusion scores achieved are significantly higher for the metabolites present in the reference dataset, labelled as Positive, with respect to the rest of candidate metabolites, labelled as Negative, ($p = 6.07 \cdot 10^{-08}$, one-sided Wilcoxon test). In addition, when ordered by diffusion scores, the positives are over-represented in the top 50% metabolites of a peak, with respect to other candidate metabolites ($p = 6.973 \cdot 10^{-04}$, one-sided Fisher's exact test).
Conclusions: This study demonstrates that KEGG RPair network has predictive power in LC-MS annotation, specifically through the use of network diffusion to prioritize candidate metabolites.

INTEGRATING HIGH-RESOLUTION OPTICAL MICROSCOPY AND COARSE-GRAINED MODELS OF CHROMOSOME SEGMENTS

Pablo Romero Marimon [1], Marie Victoire Negembor [2], Diana Buitrago [1], Jürgen Walther [1], Pablo D. Dans [1], Isabelle Brun-Heath [1], Rafael Lema [1], Pia Cosma [2], and Modesto Orozco [1].

[1] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Barcelona, Spain.
[2] Centre of Genomic Regulation (CRG). The Barcelona Institute of Science and Technology. Barcelona, Spain.

The full sequence of the human genome has been unveiled almost 20 years ago, although our comprehension of gene function is still limited. The linear sequence of DNA provided invaluable information about genes, regulatory elements, and their distribution along chromosomes. However, to fully understand gene function and gene regulation the linear genome should be placed in the right context: the cell nucleus. Inside the nucleus, genes and regulatory elements are organized forming complex three-dimensional (3D) structures that change over time. In this contribution, super-resolution imaging of specific genes was integrated alongside molecular modeling to unravel the 3D conformation of genes inside the nucleus. We present a series of 2D/3D fitting procedures to connect super-resolution images of a 2.3 Mbp segment of Human chromosome 12 (Chr12) and two individual genes with two particle-based coarse-grained (CG) models of chromatin. Immunolabeling of H3 histones of the whole nucleus of fibroblast cells (IMR90) was combined with STORM microscopy to obtain the localization of nucleosomes with nanometric precision. Simultaneously, oligopaint strategy was used to label and visualize the DNA sequence corresponding to the genes GAPDH and NANOG. In parallel, we developed a top-down CG model to represent Chr12 based on HiC-biased simulations which results in a family of 3D structures resembling the chromosomal path that are fitted to the immuno-STORM localizations. To represent the genes at higher resolution and reproduce oligo-STORM data, we used a bottom-up CG model of chromatin with basepair resolution and nucleosome positions derived from MNase-seq data which is connected to a Monte Carlo code to sample the conformational space. By fitting the experimental localizations to the CG models we identified complex structural arrangements and differential gene compaction that correlate with different states of transcription, and which results in the first models of individual Human genes in 3D at near atomic resolution.

REFERENCE-FREE RECONSTRUCTION AND ERROR CORRECTION OF TRANSCRIPTOMES FROM NANOPORE LONG-READ SEQUENCING

Ivan de la Rubia [1] , Joel Indi [1 2] , Eduardo Eyras [1 3]

1. Pompeu Fabra University, Barcelona, Spain
2. Instituto de Medicina Molecular, Universidade de Lisboa , Lisbon, Portugal
3. Catalan Institution of Research and Advanced Studies, Barcelona, Spain

Email: ivan.delarubia@upf.edu

Disease states generally present specific RNA transcripts that do not exist in normal cells. These transcript isoforms may not exist in the reference annotation, and short-read sequencing data may not recover them accurately either due to the limitation of the read length or to the lack of an appropriate genome reference. Long-read sequencing technologies offer the potential to obtain the actual transcriptome operating in cells. However, accurate analysis of long-reads remains challenging due to error rates and the fact that transcript splicing variants differ from each other by short sequence stretches; hence, new analysis methods are needed.We have developed RATTLE, a new method for the reference-free reconstruction of transcriptomes from Nanopore sequencing reads. RATTLE uses a new k-mer based similarity measure to cluster reads and quantify transcripts, performs error correction, and delineates alternative transcript isoforms, from long reads without the need of a genome reference. Using experimental and simulated data, we show that RATTLE outperforms other methods at clustering, and it achieves a detection of known splice-sites similar to reference-based methods. Additionally, read correction with RATTLE improves the proportion of reads mapped to the genome, and recovers accurately gene and transcript abundances. Direct quantification of transcriptomes coupled to our tools SUPPA (Trincado et al. 2018) and SPADA (Climente-González et al. 2017), leverages long-read technologies for the study of transcriptome dynamics without the need of a reference genome using our tool

RATTLE enables the improved characterization of transcriptomes operating in cells across multiple conditions and disease states directly from long-read sequencing, opening up the application of quantitative transcriptomics in cell models, non-model organisms, and individuals for which a genome reference is not available. Our method, together with the mobility of Nanopore technology, will facilitate the systematic implementation of long-read transcriptomics in clinical and field work.

SYSTEMS-BIOLOGY MECHANISTIC EVALUATION OF THE ROLE OF GRK2 IN COLORECTAL CANCER RELAPSE ACCORDING TO DISEASE STAGES AND GRK2 GENE EXPRESSION LEVELS

Raquel Valls[1], Teodoro Vargas[2], Ana Ramírez de Molina[2], Simón Perera[1], Petronila Penela[3], Federico Mayor Jr[3], José Manuel Mas[1]

[1] Anaxomics (Barcelona)

[2] Unit of Molecular Oncology and Nutritional Genomics of Cancer, IMDEA-Food

[3] Departamento de Biología Molecular and Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM); IIS La Princesa

Abstract

BACKGROUND: Colorectal cancer (CRC) is one of the most common cancers worldwide, and its recurrence remains a major issue [1]. GRK2 is a signaling node in cell growth and proliferation-related processes [2,3], which can influence cancer progression in a cell-type and tumor-specific way [4]

METHODS: Association of GRK2 mRNA levels with progression free survival (PFS) was analysed by Cox regression analysis and Kaplan-Meier curves in gene expression data from GSE17538 and GSE39582 [5-7] of CRC patients. Anaxomics's TPMS technology, consisting in the molecular modeling and analysis of human pathophysiology through Systems Biology and Artificial Intelligence [8-10], was applied to identify relevant pathways and to suggest mechanistic hypotheses regarding GRK2's role. GRK2-interactome and functionally linked processes were integrated with gene expression data and assessed through a gene set enrichment analysis (GSEA) [11] and a molecular mechanism analysis.

RESULTS: Stage-2, high-GRK2 CRC patients had decreased PFS and 2-fold higher risk of relapse. The inverse is found in advanced CRC patients. GSEA results support that GRK2 fosters oncogenic pathways in CRC, especially in stage-2 patients; and also supports a dual role for GRK2 with some different stage-dependent enriched pathways (e.g. GPCR-related pathways). The mechanistic analysis identified differential effector-to-effector pathways leading to GRK2-triggered CRC in either same- or different-stage patients, normally or over-expressing GRK2. Most importantly, it underlines the higher difference between same-stage patients with discordant GRK2 expression that between different-stage patients concordant for GRK2 expression levels.

References

[1] Kanwar, S.S., A. Poolla, and A.P. Majumdar, *Regulation of colon cancer recurrence and development of therapeutic strategies.* World J Gastrointest Pathophysiol, 2012. **3**(1): p. 1-9.

[2] Rivas, V., et al., Role of G protein-coupled receptor kinase 2 in tumoral angiogenesis. Mol Cell Oncol, 2014. **1**(4): p. e969166.

[3] James D Robinson, J.A.P., The diverse roles of G protein-coupled receptor kinase 2 (GRK2): a focus on regulation of receptor tyrosine kinases (RTKs). Receptors & Clinical Investigation, 2014. **Vol 1, No 3**.
[4] Nogués L, Palacios-García J, Reglero C, Rivas V, et al.G protein-coupled receptor kinases (GRKs) in tumorigenesis and cancer progression: GPCR regulators and signaling hubs. Semin Cancer Biol. 2018 Feb;48:78-90. PMID: 28473253
[5] Smith JJ, Deane NG, Wu F, Merchant NB et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 2010 Mar;138(3):958-68. PMID: 19914252
[6] Freeman TJ, Smith JJ, Chen X, Washington MK et al. Smad4-mediated signaling inhibits intestinal neoplasia by inhibiting expression of β-catenin. Gastroenterology 2012 Mar;142(3):562-571.e2. PMID: 22115830
[7] Marisa L, de Reyniès A, Duval A, Selves J et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med 2013;10(5):e1001453. PMID: 23700391
[8] Perera, S., Artigas, L., Mulet, R., Mas, J. M., & Sardón, T. (2014). Systems biology applied to non-alcoholic fatty liver disease (NAFLD): treatment selection based on the mechanism of action of nutraceuticals. Nutrafoods, 13(2), 61-68.
[9] Iborra-Egea, O., Gálvez-Montón, C., Roura, S., Perea-Gil, I., Prat-Vidal, C., Soler-Botija, C., & Bayes-Genis, A. (2017). Mechanisms of action of sacubitril/valsartan on cardiac remodeling: a systems biology approach. NPJ systems biology and applications, 3(1), 12.
[10] Romeo-Guitart, D., Forés, J., Herrando-Grabulosa, M., Valls, R., Leiva-Rodríguez, T., Galea, E., ... & Coma, M. (2018). Neuroprotective Drug for Nerve Trauma Revealed Using Artificial Intelligence. Scientific reports, 8(1), 1879.
[11] Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

ASSESSMENT OF KINSHIP DETECTION USING RNA-SEQ DATA

Blay N[1], Casas E[1], Galván-Femenía I[2], de Cid R[2], Vavouri T[1,2]
[1] Josep Carreras Leukaemia Research Institute, Barcelona, Spain
[2] Institut de Recerca Germans Trias I Pujol, Barcelona, Spain

Contact: nblay@carrerasresearch.org

Kinship detection is a common practice in genetic analysis, both to detect new relationships and to confirm known relationships. Currently, reconstruction of pedigrees is based on genotyping arrays, whole genome sequencing, whole exome sequencing and microsatellites. In the case where only RNA-Seq data is available for multiple individuals in a pedigree, kinship detection could be useful to detect labeling mistakes. The problem with this type of data is the low proportion of the genome that covers, and the differential allelic expression (i.e. due to imprinting); providing a low quantity of SNPs and with some genotyping errors.

We have assessed the use of RNA-Seq data to determine and represent kinship relationships between individuals, through pairwise identity by descent (IBD) estimation of high quality SNPs. We developed a pipeline through which we obtained high quality SNPs after successive filters to minimize allelic imbalance, sequencing errors, mapping errors and genotyping errors, and used these SNPs to calculate pairwise IBD estimates. The method provides a graphic representation of those IBD estimates and also a representation of the detected familial relationships. We compared the results of this pipeline using RNA-Seq data of a human family and with simulated data from 1000 Genomes Project. This method is able to detect labeling mistakes and to identify up to second degree relationships.

SYSTEMATIC ASSESSMENT OF GENETIC CONTEXT-DEPENDENT GENE ESSENTIALITY

Jolanda van Leeuwen[1,*], Carles Pons[2,*], Patrick Aloy[2], Brenda J. Andrews[1,3], and Charles Boone[1,3]

[1]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario M5S 3E1, Canada
[2]Institute for Research in Biomedicine (IRB Barcelona), the Barcelona Institute for Science and Technology, Barcelona, Catalonia, Spain
[3]Department of Molecular Genetics, University of Toronto, 160 College Street, Toronto ON M5S 3E1, Canada
*These authors contributed equally to this work.

The genetic background in which a mutation occurs often modulates its functional impact. In the most extreme case, gene essentiality may differ between individuals of the same species. To identify genes whose essentiality depended on the genetic context, we systematically isolated spontaneous mutations that could overcome the lethality of deletion alleles of essential genes in *Saccharomyces cerevisiae*. Out of the 728 essential genes tested, we identified 124 with genetic context-dependent essentiality. These dispensable genes were enriched for gene duplicates and for transport and signaling functions, but depleted for members of protein complexes and for genes involved in protein degradation and RNA processing. Most suppressor mutations were on genes functionally related to the dispensable gene and they often affected essential genes in cases of gain-of-function mutations. Dispensable genes exhibited particular phylogenetic properties, being more likely to be absent, duplicated, or non-essential in other species than non-dispensable genes. Our work illustrates the distinct features of context-dependent essential genes, and reveals the mutations relevant for the cellular rewiring needed to overcome the lethality of essential deletions. These results can potentially help to interpret the differential gene essentiality between individuals and across species.

CAP-CLS: FULL-LENGTH LONG NONCODING RNA ANNOTATION WITH 5' CAP-SELECTED RNA CAPTURE COUPLED WITH LONG-READ SEQUENCING

Silvia Carbonell Sala[1], Julien Lagarde[1], Barbara Uszczynska-Ratajczak[2], Hiromi Nishiyori-Sueki[3], Piero Carninci[3], Rory Johnson[4], Roderic Guigo[1] & The GENCODE Consortium

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

[2]Centre of New Technologies, Warsaw, Poland

[3]RIKEN Center for Integrative Medical Sciences, Yokohama Campus, Japan

[4]Department of Clinical Research, University of Bern, Bern, Switzerland

silvia.carbonell@crg.eu

LncRNAs are RNA transcripts longer than 200 nucleotides that are capped and often spliced, but that do not encode any identifiable peptide product. Several lncRNAs are linked to human diseases, but their function is still unknown for the most part. LncRNA expression is in general tissue specific and a large number of lncRNAs are very lowly expressed.

Accurate annotation of genes and their transcripts is a foundation of genomics. However, reference gene collections remain incomplete—many gene models are fragmentary, and thousands more remain uncatalogued, particularly for long noncoding RNAs (lncRNAs).

The GENCODE project aims to annotate with high accuracy all loci encoding lncRNAs in the human and mouse genomes. To build the exhaustive and high-confidence catalog of lncRNAs we employ Capture Long-read Sequencing (CLS).

Recently we demonstrated the superiority of the CLS method compared to short read sequencing methods to annotate lncRNAs. Still, many transcripts are incomplete at their 5' ends. To remedy that, we adapted the CapTrap 5'-end enrichment protocol to our CLS workflow (Cap-CLS). Preliminary results obtained on the Oxford Nanopore platform show that Cap-CLS considerably improves the 5'-end completeness of sequenced transcripts when compared to traditional cDNA library construction methods. Thus, the Cap-CLS method will greatly facilitate the high-throughput characterization of mammalian full-length transcriptomes.

UNRAVELING THE MOLECULAR MECHANISMS INVOLVED IN ADAPTATION OF ADIPOSE TISSUE TO COLD

Jordi Rodó, Miquel Garcia, Estefania Casana, Victor Sacristan, Claudia Jambrina, Sergio Muñoz, Cristina Mallol, Xavier Leon, Sara Darriba, Ignasi Grass, Sylvie Franckhauser, Veronica Jimenez and Fatima Bosch

Center of Animal Biotechnology and Gene Therapy and Department of Biochemistry and Molecular Biology, School of Veterinary Medicine, Universitat Autònoma de Barcelona and Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Spain

Background and aims: Type 2 diabetes (T2D) and obesity are strongly associated and constitute a major health problem. Novel and safe approaches are needed to prevent and combat the current T2D-obesity epidemic. To unravel factors capable of inducing browning of white adipose tissue (WAT), activation of brown adipose tissue (BAT) and/or improving glucose metabolism we performed an exhaustive, comparative transcriptomic analysis of mRNA and lncRNA isolated from brown and white adipose depots of adult mice exposed to cold.

Results: Principal component analysis (PCA) suggested that the gene expression profile of BAT or eWAT exposed to 23ºC or 4ºC did not differ significantly. However, cold exposure leads to a strong gene expression changes in iWAT samples. Differentially expressed genes and lncRNA were then evaluated in each fat pad. Specifically, 243, 247 and 50 genes were upregulated in iBAT, iWAT, and eWAT, respectively, whereas 87, 218 and 672 were downregulated in the same tissues. Functional analysis of this data further demonstrated that iWAT exposed to cold resembled BAT. iWAT also showed the highest number of enriched pathways and gene ontologies with significant statistically differences. Moreover, both pathway enrichment and gene ontology results indicated that major differences were due to metabolism- and mitochondrial-related genes. A second bioinformatic pipeline using several complementary approaches such as logical set relation, pattern matching, gene functional, co-expression and interaction networks analyses allow us to identify novel cold-induced factors capable of improving energy and glucose metabolism. The role of the most robust factors was then tested in mice fed a high-fat diet (HFD) after gene transfer using adeno-associated viral (AAV) vectors.
Conclusion: This study contributes to better understand the molecular mechanisms underlying adaptation of adipose tissue to cold and identifies novel factors able to enhance non-shivering thermogenesis.

GENOMIC ANALYSIS OF POLYPURINE REVERSE-HOOGSTEEN HAIRPINS DIRECTED AGAINST SURVIVIN IN HUMAN CELLS: OFF-TARGET EFFECTS AND TOXICITY STUDIES

Alex J. Félix[1], Carlos J. Ciudad[1] and Véronique Noé[1]

[1]Department of Biochemistry and Physiology, School of Pharmacy, and Institute of Nanoscience and Nanotechnology, University of Barcelona, Barcelona, Spain.

Presenter e-mail: ajife93@gmail.com

PolyPurine Reverse Hoogsteen (PPRH) hairpins are new pharmacological tools used for gene silencing that have been applied for a number of gene targets. We had previously reported that PPRHs against survivin were able to decrease specifically the viability of PC3 prostate cancer cells. The mechanism of cell death was due to an increase in apoptosis since survivin is an antiapoptotic gene. This effect was specific for prostate cancer cells given that it was not effective in HUVEC non-tumoral cells. These PPRHs were efficient both *in vitro* and *in vivo*. In the present work, we performed a functional pharmacogenomics study on the effects of specific and unspecific hairpins against survivin. Incubation of PC3 cells with the specific HpsPr-C-WT led to 244 differentially expressed genes when applying the $p < 0.05$, $FC > 2$, Benjamini-Hochberg filtering. Importantly, the unspecific or control Hp-WC did not originate differentially expressed genes using the same settings. Gene Set Enrichment Analysis (GSEA) showed that the differentially expressed genes clustered very significantly within the gene sets of Regulation of cell proliferation, Cellular response to stress, Apoptosis and Prostate cancer. Network analyses using STRING identified important interacting gene-nodes within the response of PC3 cells to treatment with the PPRH against survivin, mainly POLR2G, PAK1IP1, SMC3, SF3A1, PPARGC1A, NCOA6, UGT2B7, ALG5, VAMP7
and HIST1H2BE, the former six present in the Gene Sets detected in the GSEA. In addition, hepatotoxicity and nephrotoxicity *in vitro* studies were performed in HepG2 and 786-O cell lines, respectively. The unspecific hairpin did not cause toxicity in cell survival assays (MTT) and produced minor changes in gene expression for selected genes in RT-qPCR arrays specifically developed for hepatic and renal toxicity scr

THE REGULATORY GENOME OF DROSOPHILA REGENERATION

Cecilia C. Klein[1,2], Elena Vizcaya[1], Florenci Serras[1], Roderic Guigó[2,3] and Montserrat Corominas[1]
[1]Departament de Genètica, Microbiologia i Estadística and Institute of Biomedicine (IBUB), Universitat de Barcelona. Barcelona, Catalonia, Spain, [2]Centre for Genomic Regulation (CRG). Barcelona, Catalonia, Spain, [3]Universitat Pompeu Fabra. Barcelona, Catalonia, Spain

Presenter e-mail: Cecilia.klein@crg.eu

One of the key questions in regenerative biology is to unveil the regulatory regions capable to trigger tissue recovery. Regeneration is the ability to reconstruct missing parts. The capacity to regenerate varies greatly, not only between species, but also between tissues and organs, as well as from one developmental stage to another in the same species. *Drosophila* imaginal discs show a high regenerative capacity after genetically induced cell death. We performed genome-wide chromatin landscape analyses (ATAC-Seq and RNA-Seq) at different time points of *Drosophila* imaginal disc regeneration to study the transcriptional programs as well as the regulatory elements responsible for tissue regeneration. We identified sets of upregulated genes located close to one another in the linear genome (clusters). Open chromatin regions that presented higher accessibility in regeneration compared to controls (namely damage-responsive regulatory elements: DRREs) were classified according to their position relative to the closest transcription start site of a gene. DRREs were validated using enhancer reporter fly lines. Since spatial chromatin organization connects active enhancers to target promoters to regulate gene expression, we confirmed individual interactions between DRREs and clusters of co-regulated genes by Chromatin Conformation Capture. Moreover, DRREs contained conserved binding motifs for transcription factors that are upregulated and required for regenerating organs in fly, zebrafish and mouse. Our findings indicate there is global co-regulation of gene expression where genes localized in genomic clusters may be regulated by the same elements. Furthermore, we found a regeneration program driven by the cooperation among regulatory elements acting exclusively within damaged tissue, with enhancers co-opted from other tissues and other developmental stages, as well as with endogenous enhancers that show increased activity after injury. Such elements host binding sites for regulatory proteins that include a core set of conserved transcription factors that may control regeneration across metazoans.

THE CAUSES AND CONSEQUENCES OF LOSING THE RESPIRATORY CHAIN COMPLEX I

Miquel Àngel Schikora Tamarit*, Marina Marcet-Houben*, Toni Gabaldón Estevan*
*Centre de Regulació Genòmica. Comparative Genomics Lab.

Mitochondrial oxidative phosphorylation (OXPHOS) is one of the most essential and conserved systems across the whole tree of life. Recent evidence has shown that OXPHOS factors can act beyond energy production in cells. One of the cornerstones of this pathway is Complex I, which also plays a role in drug response, oxidative stress and pathogenesis. Furthermore, mutations on this complex are associated with cancer, neurodegenerative and mitochondrial disease in humans.

One of the most puzzling observations of the field is that Complex I has been lost multiple independent times in eukaryotic species, mostly in fungi. This represents a recurrent phenomenon of convergent evolution, whose implications on cell physiology and behavior remain obscure. Interestingly, a lot of Complex I-lacking organisms have adapted to a pathogenic lifestyle, including the causative agents of malaria, microsporidiosis and some types of plant parasitism. Inferring the causes and consequences of losing Complex I could lead to a better understanding of its role on multiple cellular systems, and eventually clarify the infectivity mechanisms of pathogens that lack this “essential” component.

To this end, we have performed a comparative phylogenomics study of fungal organisms to search for genomic features correlated to Complex I loss. Our current results (based on a few comparisons of fungal species) suggest that the upregulation of drug resistance and nutrient starvation survival are associated with Complex I loss. We have designed a machine-learning-based approach to validate these findings across all fungal genomes. Our hypothesis is that this will reveal the factors (in terms of genes and/or pathways) related to the causes (the evolutionary events that led to the loss) and consequences (the impact on cell physiology and perhaps adaptation to new environments) of losing such an important building block of eukaryotic life.

CHEMOSENSITIZATION TO METHOTREXATE TREATMENT IN COLON CANCER CELLS USING POLYPURINE REVERSE HOOGSTEEN HAIRPINS

Laia Fargas Codina[1], Alex J. Félix[1], Véronique Noé[1] and Carlos J. Ciudad[1]

[1]Department of Biochemistry and Physiology, School of Pharmacy, and Institute of Nanoscience and Nanotechnology, University of Barcelona, Barcelona, Spain.

Presenter e-mail: laiafargas4@gmail.com

Cancer chemotherapy with methotrexate (MTX), an inhibitor of dihydrofolate reductase, leads to drug resistance which impairs the success of the treatment. Previously, we had analyzed by functional genomics the differentially expressed genes in HT29 colon cancer cells resistant to MTX. In the present work we analyzed by Human exon expression microarrays (GeneChip Gene 1.0 ST Array System) the human transcriptome response to short incubations with MTX (8, 24 and 48 hours) using the GeneSpring 14.9 software, selecting those genes with FC >2 and $p<0.05$. Then, we compared the two sets of genes using Venn-diagrams to find out the differentially expressed genes in common between resistant cells and cells treated for a short period of time. The rationale was that those genes would be crucial in the process of development of resistance to this drug. Our aim was to inhibit the expression of the genes in common using Polypurine reverse Hoogsteen hairpins (PPRHs) together with MTX, to increase the chemosensitivity towards MTX in colon cancer cells. Specific PPRHs were design against these targets and were transfected individually in HT29 cells with the cationic liposome DOTAP and treated with MTX. HT29 cells were incubated for two days with PPRHs against genes such as Caveolin 1 (*CAV1*), the *SLC4A11* transporter, the aldo-ketoreductase1C2 (*AKR1C2*) and the cell adhesion protein MCAM. The silencing of these genes significantly increased MTX sensitivity compared to MTX alone. In conclusion, this study identified *CAV1*, *SLC4A11*, *AKR1C2* and *MCAM* as genes that could be targeted in adjuvant MTX therapy to avoid the resistance to this chemotherapeutic agent.

UNRAVELLING THE MYSTERIES OF TRANSCRIPTION WITH DRNA LONG READ SEQUENCING

Authors: Jèssica Gómez Garrido[1], Beatriz Martin Mur[1], Regina Antoni[1], Javier Gutierrez[1], Marc Dabad[1], Fernando Cruz[1], Simon Heath[1,2], Ivo Gut[1,2], Marta Gut[1,2], Tyler Alioto[1,2].

1. CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Baldiri Reixac 4, Barcelona 08028, Spain

2. Universitat Pompeu Fabra (UPF), Barcelona, Spain

Presenting author: Jèssica Gómez Garrido (jessica.gomez@cnag.crg.eu)

Over the past decade, RNA-seq has been the go-to tool for detecting transcripts, quantifying differential expression and studying alternative splicing. However, Illumina sequencing has the distinct drawback of generating only short reads, with which it is very difficult, if not impossible, to accurately reconstruct entire transcripts, especially in the face of pervasive alternative splicing and lack of reference genome. Newer single molecule long-read sequencing methods like PacBio Iso-seq or Oxford Nanopore's cDNA and/or Direct RNA sequencing have the advantage of being able to generate full-length transcripts, albeit at the expense of lower throughput and higher error rate. Moreover, Direct RNA sequencing also provides a new tool for directly assessing base modifications that occur in transcripts. We present here the work that we have done to incorporate Direct RNA reads into our de novo genome annotation pipeline, allowing us to easily determine real isoforms and give less weight to aberrant constructions that result from assembling the Illumina reads. Furthermore, we also highlight the added value that long reads bring to our effort to discover new isoforms in already annotated genomes and their potential usefulness for the study of events such as intron retention or fusion transcripts. In conclusion, our purpose here is to summarize the effort that we make to stay on top in the use of new sequencing technologies and take advantage from them in the transcriptomic and genomic analysis that we normally perform.

PAN-CANCER ANALYSIS OF AMINO ACID CHANGES AND THEIR POTENTIAL EFFECT ON PROTEIN STRUCTURE

A. Diéguez-Docampo[1*], M.D. Stobbe[1] and I.G. Gut[1,2]

[1] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain.
[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain.

[*]Author presenting poster: andrea.dieguez@cnag.crg.eu

The Pan-Cancer Analysis of Whole Genomes consortium has brought together whole-genome sequencing data for 2,583 cancer genomes, covering 37 tumour types, from the International Cancer Genome Consortium and The Cancer Genome Atlas projects. Our goal is to do a pan-cancer analysis of all protein structure changes caused by somatic substitutions in protein-coding genes found in this dataset. By analysing such a large data set we can extract information and uncover patterns that may be missed when looking at only one or a few samples individuals. We first focused on missense mutations, which produce an amino acid change in the protein sequence. To predict their potential effect we evaluated different features, such as the location of the mutated residue together with its resulting chemical change, the free energy change of protein folding using FoldX [1] and the conservation of the mutated amino acid using Consurf [2]. As a case study we analysed 90 cancer genomes from the Spanish Chronic Lymphocytic Leukemia Genome Project [3], in which 1172 somatic missense mutation were detected. For 94% of these a protein structure of sufficient quality and coverage is missing, which is required for the energy predictions. Without this feature conclusions on the changes in protein structure is more limited. This shows the challenge of studying mutations in detail on a structural level. For 69 mutations, however, for which we did have all the information we predict that 28 are likely damaging the protein, while the others seem not to be problematic. We will use these results to study the impact of these mutations on the biological function of the specific protein. Mutations can for example disrupt functional domains or active sites causing a loss or gain of protein function. This information will help prioritize interesting mutations, e.g. those with important implications for clinical outcomes.

[1] Schymkowitz, et al., Nucleic Acids Res. 2005
[2] Ashkenazy et al., Nucleic Acids Res. 2016
[3] Beekman et al., Nat. Med. 2018

CNAPP: A WEB-BASED TOOL FOR INTEGRATIVE ANALYSIS OF GENOMIC COPY NUMBER ALTERATIONS

Sebastià Franch-Expósito[1]*, Laia Bassaganyas[2]*, Maria Vila-Casadesús[3], Eva Hernández-Illán[1], Roger Esteban-Fabró[2], Marcos Díaz-Gay[1], Juan José Lozano[3], Antoni Castells[1], Josep M. Llovet[2,4,5], Sergi Castellví-Bel[1], Jordi Camps[1,6]

1 Gastrointestinal and Pancreatic Oncology Team, Institut D'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clínic de Barcelona, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Universitat de Barcelona, Barcelona, Catalonia, Spain

2 Liver Cancer Translational Research Group, Liver Unit, Institut D'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clínic, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Universitat de Barcelona, Barcelona, Catalonia, Spain

3 Bioinformatics Unit, CIBEREHD, Barcelona, Catalonia, Spain

4 Mount Sinai Liver Cancer Program, Division of Liver Diseases, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, USA

5 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

6 Unitat de Biologia Cel•lular i Genètica Mèdica, Departament de Biologia Cel•lular, Fisiologia i Immunologia, Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra, Catalonia, Spain

* equal contribution

Presenter e-mail: sebasfranch@gmail.com

Copy number alterations (CNAs) are a hallmark of cancer. Large-scale cancer genomic studies have already established the CNA landscape of most human tumor types and some CNAs are recognized as cancer-driver events. However, their precise role in tumorigenesis as well as their clinical and therapeutic relevance remains undefined, thus computational and statistical approaches are required for the biological interpretation of the data. Here, we describe CNApp, a user-friendly web tool that offers sample- and cohort-level association analyses, allowing a comprehensive and integrative exploration of CNAs with clinical and molecular variables. CNApp generates genome-wide profiles, calculates CNA levels by computing broad, focal and global CNA scores, and uses machine learning-based predictions to classify samples by using segmented data from either microarrays or next-generation sequencing. In the present study, using copy number data of well-annotated 10,635 genomes from The Cancer Genome Atlas spanning 33 cancer subtypes, we showed that patterns of CNAs classified tumor types according to their tissue-of-origin and that broad and focal CNA scores correlated positively in those samples showing low levels of chromosome and arm-level events . Moreover, CNApp allowed a robust description of CNAs in hepatocellular carcinoma further confirming previous results identified with other methods. Finally, we established machine learning-based models to predict colon cancer molecular subtypes and microsatellite instability based on broad and focal CNA scores and specific genomic imbalances. In summary, CNApp facilitates data-driven research and provides a unique framework to comprehensively assess CNAs and perform integrative analyses to enable identification of relevant functional implications. CNApp is hosted at http://bioinfo.ciberehd.org/ and the source code is freely available at GitHub (https://github.com/ait5/CNApp).

MULTITRAIT GENOME ASSOCIATION ANALYSIS FOR HUMAN SKIN AND HAIR VARIATION IN THE GCAT COHORT

Iván Galván-Femenía[1], Mireia Obón-Santacana[1,2], David Piñeyro[3], Marta Guindo-Martinez[4], Xavier Duran[1], Anna Carreras[1], Raquel Pluvinet[3], Juan Velasco[1], Laia Ramos[3], Susanna Aussó[3], Josep M Mercader[5,6], Lluís Puig[7], Manuel Perucho[8], David Torrents[4,9], Víctor Moreno[2,10], Lauro Sumoy[3] and Rafael de Cid[1]1. Genomes for Life - GCAT lab Group, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Crta. de Can Ruti, Camí de les Escoles s/n. Can Ruti Campus, 08916 Badalona, Spain
2. Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO), IDIBELL and CIBERESP. Avinguda de la Gran via, 199, 08908 L'Hospitalet de Llobregat, Barcelona, Spain
3. High Content Genomics and Bioinformatics Unit, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Crta. de Can Ruti, Camí de les escoles s/n. Can Ruti Campus, 08916 Badalona, Spain
4. Barcelona Supercomputing Center (BSC-CNS). Joint BSC-CRG-IRB Research Program in Computational Biology, Carrer de Jordi Girona 29-31, 08034 Barcelona, Spain
5. Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, US.
6. Diabetes Unit and Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, US.
7. Banc de Sang i Teixits, Carrer del Taulat, 106, 08005 Barcelona, Spain
8. Cancer Genetics and Epigenetics Group, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Crta. de Can Ruti, Camí de les escoles s/n. Can Ruti Campus, 08916 Badalona, Spain
9. ICREA, Catalan Institution for Research and Advanced Studies, Barcelona, Passeig de Lluís Companys, 23, 08010 Barcelona, Spain
10. Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain
Presenter email: igalvan@igtp.cat

Heritability estimates have revealed an important contribution of genetic variants for most common traits. However, SNP analysis by single-trait genome-wide association studies (GWAS) has failed to achieve their total heritability. In our study (Galván-Femenía et al. "Journal of Medical Genetics", 2018), we applied a multitrait GWAS approach to discover additional factor of the missing heritability of human anthropometric variation. We analysed 15 million genetic variants from the Genomes For Life - Cohort study of the Genomes of Catalonia (GCAT) at baseline (N=4988) (Obón-Santacana et al. "BMJ Open", 2018). In single-trait SNP association we confirmed variants in *IRF4* ($p=2.8\times10^{-57}$), *SLC45A2* ($p=2.2\times10^{-130}$), *HERC2* ($p=2.8\times10^{-176}$), *OCA2* ($p=2.4\times10^{-121}$) and *MC1R* ($p=7.7\times10^{-22}$) associated with hair, eye and skin colour, freckling, tanning capacity and sun burning sensitivity and the Fitzpatrick phototype score, all highly correlated cross-phenotypes. In this study, we further explore the missing heritability and present new results on human skin and hair colour traits by applying the multi-trait GWAS approach. We replicate the results from the GCAT cohort with GWAS summary statistics from the UK Biobank (N=361,194) (http://www.nealelab.is/uk-biobank).

NEXTFLOW MEETS PROVENANCE USING RESEARCH OBJECT SPECIFICATION

Edgar Garriga Nogales[1,2], Paolo Di Tommaso[1], Cedric Notredame[1]

1 Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain
2 Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
Presenter e-mail: edgar.garriga@crg.eu

With computation becoming an increasingly important component of medical data analysis, the issue of provenance has become central so as to ensure traceability and reproducibility of data analysis, and, ultimately, of diagnostics. Provenance refers to information about entities, activities, and people involved in producing a piece of data or its analysis, which can be used to form assessments about its quality, reliability or trustworthiness. Some of the main goals of the provenance are: defining the activities and entities of the workflow. Being able to know when they started and when they ended, track who created the file, which process used the file and who supervised the procedure including the full parameters used to ran the pipeline.
Proper management of provenance will allow biologists to access details of any particular workflow execution, compare results produced by different executions or plan new experiments more efficiently. Reproducibility is one of the core principles for any scientific workflow and remains a challenge, meanwhile, provenance should be tracked and used to capture all these requirements supporting reusability of existing workflows.

In this project, we combined Nextflow and ResearchObject. Nextflow is a framework based on the data flow programming model, which simplifies the writing of parallel and distributed pipelines while allowing developers to focus on the application. One of the strengths of Nextflow is the integration of Docker and Singularity, allowing self-contained and reproducible computational pipelines. Another of the benefits of Nextflow is to be polyglot. The main three objectives of ResearchObject are: Identity, providing a unique identifier to the project. Aggregation, allowing the author to wrap all the useful elements for the project. And the Annotation, capturing an extra metadata to know the relation between elements, when and how they were produced. The results are fully traceable workflow output.

FBONTO: AN ONTOLOGY TO REPRESENT FOOD INTAKE DATA AND ASSOCIATE IT WITH METABOLOMIC DATA

Pol Castellano Escuder[1,2], Cristina Andrés-Lacueva[1], Alexandre Sánchez-Pla[2]

[1]Biomarkers and Nutritional & Food Metabolomics Research Group, Department of Nutrition, Food Science and Gastronomy, University of Barcelona. [2]Statistics and Bioinformatics Research Group, Department of Genetics, Microbiology and Statistics, University of Barcelona.

MOTIVATION:

Currently, nutrition research generates a lot of complex data hard to analyze and associate with other data or omics such as metabolomics. By itself, metabolomics is closely linked to nutrition. However, it is still difficult to associate these two types of data. One reason for this difficulty is the heterogeneity found in the information provided by participants in nutritional studies about what they have eaten. In order to manage this heterogeneity we have decide to build an ontology, describing foods in a hierarchichal way that enables a common description of food intake. This ontology contains **formal naming, definition of the categories, properties, and relations between the concepts of two types of data, food and related metabolites.**

METHODS:

This ontology has been created using Protégé 5.5.0 version. As is common in ontologies, FBOnto (Food-Biomarker Ontology) has been written in OWL language (Web Ontology Language).

RESULTS:

The ontology presented is called FBOnto and is made of two sub-ontologies. The first ontology describes foods: from simple such as fruits and vegetables to more complex foods such as *ratatouille*. The second ontology describes metabolites, grouped in their different chemical classes. The nodes or elements of these two sub-ontologies are connected by the properties of each one, so that if a metabolite is in different foods, it will connect with all of them. This ontology allows us to visualize data in a bidirectional way, going from metabolomics to nutritional data or vice versa. In addition, this ontology can also be useful for other analyses such as enrichment analysis or a novel concept, food-enrichment analysis.

RELATING BULK TRANSCRIPTOMIC VARIATION WITH PATTERNS IN HISTOPATHOLOGICAL IMAGES

Vasilis Ntasis[1,4], Manuel Muñoz[1,2], Roderic Guigó[1,3]
[1] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Catalonia
[2] Universitat Politècnica de Catalunya. Departament d'Estadística i Investigació Operativa. 08034 Barcelona, Catalonia
[3] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Catalonia
[4] Department of Biology, University of Crete, Greece

The Genotype Tissue Expression (GTEx) project has generated bulk RNAseq data for more than 900 human individuals across 53 different tissues, totalling almost 17400 samples. For each of these samples, there is a high resolution histopathological image of the tissue slice available. With these unique resources, we aim to relate transcriptomic variation with changes in phenotypes derived from patterns in histopathological images.

Histological images are very large in terms of size, making it difficult for computational pipelines to process them as single units, thus, they have to be divided into patches. Moreover, a significant portion of each image is background, which should be excluded from downstream analyses. In order to efficiently segment tissue content from each image, we designed a pipeline that involves the use of a Canny edge detector and a graph-based segmentation method. This allows us to select only relevant patches from the high-resolution images.

Once the images have been preprocessed, we use them to train a convolutional neural network to perform classification. We observe that our model is able to discriminate with high accuracy to which tissue each image belongs to. As part of this process, the model learns a low dimensional representation of the information contained in each image. When performing clustering of these low-dimensional feature vectors, samples cluster together by tissue, indicating that these low-dimensional embeddings are accurate representations of tissue morphology. Our current work focuses on relating these learned features to changes in gene expression.

With these integrative approaches that encompass RNAseq data and *image phenotypes*, we aim to open up the path to a deeper understanding of how a pathological condition affects the human body.

HIGHLY SENSITIVITY MICROSATELLITE INSTABILITY ANALYSIS IS USEFUL FOR THE DETECTION OF MISMATCH REPAIR DEFECTS IN NORMAL TISSUE OF BIALLELIC MISMATCH REPAIR MUTATION CARRIERS

Fátima Marín*[1], Núria Bonifaci*[1], Maribel González-Acosta[1], Ben Puliafito[1], Anna Fernández[1], Daniel Rueda[2], Katharina Wimmer[3], Conxi Lázaro[1], Marta Pineda[1], Gabriel Capella[1]
[1]Hereditary Cancer Program, Catalan Institute of Oncology, ICO-IDIBELL-CIBERONC, 08908, Hospitalet de Llobregat, Barcelona, Spain.
[2]Familial Cancer Counselling Unit, Oncology Division, Doce de Octubre University Hospital, 28041, Madrid, Spain
[3]Division of Human Genetics, Medical University Innsbruck, Peter-Mayr-Straße 1, 6020 Innsbruck, Austria

*Both authors contributed equally to this study
Email addresses: fmnieto@iconcologia.net; nbonifaci@iconcologia.net

Aim: Lynch syndrome (LS), caused by germline monoallelic mutations in MMR genes, is mainly characterized by early adult-onset colorectal and endometrial tumors. Constitutional Mismatch Repair Deficiency (CMMRD), caused by biallelic mutations in the same genes, is characterized by the development of hematological malignancies, brain and colorectal tumors during childhood and adolescence. As a result of MMR deficiency, tumors exhibit microsatellite instability (MSI) and/or loss of MMR protein expression. MMR deficiency has been also described in non-neoplastic tissues from these patients.

The aim of our work is to preliminarily evaluate the putative clinical impact of the assessment of MSI at high sensitivity in peripheral blood cells in the identification of hereditary cancer syndromes associated with MMR deficiency.

Methods: Blood samples from 11 CMMRD and 48 LS patients and 37 healthy individuals were included in this study. Two stable and 3 MSI tumor samples were included as controls. A custom subexome NGS panel based on HaloPlex-HS technology, which included 277 mononucleotide repeat (MNR) markers, was sequenced at high depth. A custom pipeline was developed to detect MNR indels at high sensitivity. Data from 22 healthy controls were used to establish the baseline instability at each MRN locus. Frequencies of each allele length of individual MNR in cases were compared against the baseline frequency. An MSI score was calculated per sample, representing the percentage of instable MNR.

Results: The percentage of instable MNR markers is significantly higher in all CMMRD blood samples and MSI-tumors DNA when compared to negative controls (blood from healthy controls and DNA from stable tumors). In contrast, our approach is not sensitive enough to differentiate between LS and control blood samples.

Conclusions: Our approach might result into a diagnostic tool for CMMRD diagnosis, especially in cases with a suggestive phenotype and in the absence of identified pathogenic MMR mutations.

RESMARKERDB: A DATABASE OF BIOMARKERS OF RESPONSE TO ANTIBODY THERAPY IN BREAST AND COLORECTAL CANCER

Judith Pérez-Granado[1], Janet Piñero[1], Laura I. Furlong[1].

[1]Research Group on Integrative Biomedical Informatics (GRIB), Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Universitat Pompeu Fabra (UPF), Barcelona.

Presenter e-mail: jperez2@imim.es

The development and clinical efficacy of therapeutic monoclonal antibodies for breast and colorectal cancer have contributed greatly to the improvement of patients' outcomes by individualizing treatments according to their genetic background (Chiavenna et al., 2017). Responding patients, however, may become resistant to treatment in advanced stages. In other cases, patients may be resistant to treatment even though they are molecularly characterized to be responsive (Pruneri et al., 2016; Bronte et al., 2015). Although several databases characterize biomarkers of drug response, there is a need of resources that offer this information to the user in a harmonized manner.

Here, we present ResMarkerDB, a centralized repository that gathers information of biomarkers of response to FDA-approved therapeutic monoclonal antibodies in breast and colorectal cancer. ResMarkerDB was developed as a user-friendly web interface to show data in an organized way and facilitate exploration of current knowledge of these biomarkers. It integrates information from available public repositories and new data extracted and curated from the literature. All data are downloadable and were homogenized and standardized following community-based standards and available ontologies. ResMarkerDB allows prioritizing biomarker data and its response to therapy in a specific cancer type according to evidence supporting their association and potential clinical usefulness. The source of these associations is varied and includes publications and guidelines. Different levels of evidence are considered too, from pre-clinical to distinct clinical phases.

ResMarkerDB database currently features 266 biomarkers of diverse nature: 45 non-coding genes and 211 coding genes; almost 180 gene variants, more than 40 copy number alterations and 70 alterations in gene expression, among others. These alterations are mapped to more than 100 distinct genes, and are involved in more than 500 biomarker-drug-tumor associations. ResMarkerDB aims to enhance translational research efforts in identifying existing and new actionable biomarkers of drug response in cancer. This new tool is available at http://resmarkerdb.org.

FUNDING: We received support from ISCIII-FEDER (PIE15/00008, CP10/00524, CPII16/00026), the EU H2020 Programme 2014-2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and is supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. The DCEXS is a "Unidad de Excelencia María de Maeztu", funded by the MINECO (ref: MDM-2014-0370).

TRANSCRIPTIONAL PROFILING OF HUMAN EPITHELIAL CELLS UPON INFECTION BY ACINETOBACTER BAUMANNII AS A METHOD TO IDENTIFY POTENTIAL ANTIMICROBIAL CANDIDATES

Javier Macho Rendón*, Núria Crua Asensio*, Marc Torrent Burgas
Systems Biology of Infection Lab, Department of Biochemistry and Molecular Biology, Universitat Autònoma de Barcelona, 08193, Cerdanyola del Vallès, Spain.
* Both authors have contributed equally to this work

Antimicrobial resistance is the capacity of microorganisms to resist antimicrobials such as antibiotics and antivirals. The exposure to widely used drugs and the lack of new antimicrobials has triggered the emergence of multidrug-resistant bacteria, which is nowadays one of the biggest threats to human health and public health systems. This has motivated initiatives to discover and develop new molecules that can serve as more effective antimicrobials. In this study, we analyzed changes in the transcriptional profile of human epithelial cells with after infection with *Acinetobacter baumannii* to search for promising new targets that could be related with the infection response. Functional analyses derived from the outcome of time-resolved differential expression analysis allowed us to identify new genes with strong changes in the expression profile at different time-points of the infection compared to the control. Functional studies reveal that these genes may participate in signaling pathways related to the infection and immune response. The combination of these results with host-pathogen protein-protein interaction networks will pave the way to discover new antimicrobial candidates for multidrug-resistant bacteria.

ACQUISITION / LOSS OF PROTEINS AND MUTATIONS: TWO GENETIC MECHANISMS DRIVING THE EVOLUTION OF SKAPE PATHOGENS IN THE SAME DIRECTION.

Oscar Conchillo-Solé[1], Daniel Yero[1,2], Isidre Gibert[1,2], Xavier Daura[1,3]

[1] Institut de Biotecnologia i de Biomedicina (IBB), Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra (Cerdanyola del Vallès) Barcelona, Spain. [2] Dept de Genètica i de Microbiologia, UAB. [3] Catalan Institution for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain.

oscar.conchillo@uab.cat; xavier.daura@uab.cat

Currently, bacteria developing antibiotic resistance are becoming a huge problem. Microorganisms gain resistance by two, apparently, independent and unrelated genetic mechanisms of evolution: acquiring / loosing genes and fixing mutations on them. We have constructed phylogenetic trees for the most important antibiotic resistant species causing nosocomial infections (known as ESKAPE pathogens). For each species two trees were generated, each one reflecting one of the previously mentioned mechanisms; one was obtained by multiple sequence alignments of its coreproteome and the other taking into account how many proteins are shared between strains. Through the comparison of both trees we have observed that, although they are obtained from different data sets, they both group the same strains in the same clusters. Here we can infer how both genetic strategies converge and drive the evolution of such organisms in the same direction. Consequently, it suggests that organisms that share the same proteins tend to share the same mutations too. This points in the direction that, in order to find the clues for developing novel drugs and strategies to fight these pathogens, both mechanisms must be considered at the same time.

Acknowledgments

This work has been supported by Spanish MICINN/FEDER (BIO2015-66674-R) and Catalan AGAUR (2009SGR-00108).

MODELLING EMBRYONIC DECISION MAKING AT SINGLE-CELL LEVEL

Laura Mora Bitria[1], Néstor Saiz[2], Anna-Katerina Hadjantonakis[2] and Jordi Garcia-Ojalvo[1]

Affiliations:
[1]Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, 08003, Spain
[2]Developmental Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, 10065, USA

Abstract:
One of the main goals of developmental biology is to decipher how individual cells make fate decisions. Here we focus on the decision between the epiblast and primitive endoderm fates, that takes place in the growing mammalian preimplantation embryo. Single-cell expression data from mouse embryos have provided mechanistic insights on how this decision is made. Informed by experimental observations, we aim to model computationally the interplay between the biochemical pathways that dictate cell-fate decisions and the mechanical constraints established by the growing embryo. In particular, we want to establish the role of FGF signalling in coordinating spatially the two cell types emerging from the second fate decision of the developing embryo. We show that a mutual inhibition circuit with auto-activation loops between the cell fate markers Gata6 and Nanog is sufficient to recover the four cellular states that coexists during the second cell fate decision within the inner cell mass. Finally, incorporation of FGF/ERK-signalling in the previous model together with the mechanics of the developing embryo permits to recapitulate the dynamics of experimental observations.

BCHAR: BACKGROUND CHARACTERIZATION OF LC-MS SIGNAL BASED ON GAUSSIAN MIXTURE MODELS

Pol Solà-Santos[123], Sergio Picart-Armada[123], Maria Barranco i Altirriba[123], Alexandre Perera-Lluna[123]

[1]B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain;
[2]Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain;
[3]Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain;

Background: Tandem of Liquid Chromatography (LC) coupled to Mass Spectrometry (MS) is a widely used analytical technique on metabolomics. LC properly separates the chemical constituents of the body fluid by their affinities towards a packed column while MS identifies the masses through magnetic deflection and measures their relative presence in the sample in terms of intensity. Among others, equipment aging contributes to a background intensity baseline that must be removed, motivating an intensive production of computational tools during the last years. Here we present bChar, an unsupervised clustering method based on Gaussian mixture models able to characterize whether a given intensity is distinguishable from noise.

Methods: bChar fits an unsupervised Gaussian mixture model to the scan intensities to characterize density distributions of signal and noise subpopulations. Goodness of fit is tested against the null model of one population. *p*-values are computed by fitting the null model to bootstrapped intensities. Probability of belonging to one of the populations is computed using Bayes' Theorem. The method is applied to the fatty acid amide hydrolase knockout dataset (faahKO, Bioconductor) and evaluated against the intensity filter implemented on XCMS, one of the most widely used metabolomics processing tool.

Results: It has been demonstrated that LC-MS intensity matrices fit better under the assumption of the existence of two subpopulations in all but one sample (p-value<0.05, Bonferroni correction). In addition, bChar has been able to estimate noise and signal distributions per-sample in a non-supervised manner showing convergence with the optimal filters implemented on XCMS.

Conclusions: This study demonstrates that LC-MS signal can be modeled as a Gaussian mixture under the hypothesis of the existence of two subpopulations. In addition, this characterization allows to give a per-peak probability of being true signal enhancing all the downstream analysis.

PHARMSCREEN: MOLECULAR OVERLAYS DERIVED FROM 3D HYDROPHOBIC SIMILARITY.

Javier Vazquez,[†,‡] Alessandro Deplano,[†] Albert Herrero,[†] Tiziana Ginex, [‡] Enric Gibert,[†] Enric Herrero,[†] and F. Javier Luque[‡]

[†] Pharmacelera, Plaça Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain
[‡] Department of Nutrition, Food Science and Gastronomy, Faculty of Pharmacy and Institute of Biomedicine (IBUB), University of Barcelona, Av. Prat de la Riba 171, Santa Coloma de Gramenet E-08921, Spain
Email address of presenting author: j.vazloz@gmail.com

In drug discovery projects the receptor structure is not always available. Accordingly, ligand-based virtual screening of chemical libraries is a valuable strategy to address this obstacle. In this approach, determine the potential overlay between small molecules is the key step. The proper alignment is influenced by several factors, including the quality of the physico-chemical descriptors utilized to account for the molecular determinants of biological activity. Relying on the hypothesis that the variation in maximal achievable binding affinity for an optimized drug-like molecule is largely due to desolvation, we explore here a novel strategy for the 3D alignment of small molecule that exploits the partitioning of molecular hydrophobicity into atomic contributions in conjunction with information about the distribution of hydrogen-bond (HB) donor /acceptor groups in a given compound. The computational procedure is calibrated by using a dataset of 402 molecules pertaining to 14 distinct targets taken from the literature, and validated against the AstraZeneca test that comprises 121 experimentally derived sets of molecular overlays. The results point out the suitability of the MST based-hydrophobic parameters for generating molecular overlays, as correct predictions were obtained for 94%, 79%, and 54% of the molecules classified into easy, moderate and hard sets, respectively. Moreover, the results point out that this accuracy is attained at a much lower degree of identity between the templates used by the combination of electrostatic / steric fields. These findings supports the usefulness of the hydrophobic / HB fields to generative complementary overlays that may be valuable to rationalize the structure-activity relationships of drug-like compounds.

SYSTEMATIC ANALYSIS OF BIASES IN SMALL RNA SEQUENCING AFFECTING ISOMIR DATA ANALYSIS

Antonio Luna de Haro1, Raquel Pluvinet1, and Lauro Sumoy1

1. High Content Genomics and Bioinformatics Unit, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Crta. de Can Ruti, Camí de les escoles s/n. Can Ruti Campus, 08916 Badalona, Spain

Presenter email: lsumoy@igtp.cat

Next generation sequencing has revealed that a majority of mature miRNAs have sequences that diverge from the canonical reference (i.e. the miRNA sequence registered in miRBase). This observation has led to the widely accepted notion of the existence of miRNA isoforms, termed isomiRs. In recent years isomiRs have been shown to be derived from imprecise cutting during the enzymatic processing of hairpin pre-miRNA precursors, leading to terminal insertions and deletions. Additional variability is introduced by non-template addition of bases at the ends, and by editing of internal bases, primarily through conversion of cytosines into inosines resulting in sequences which differ from the template DNA sequence. Since sequencing reads have inherent errors, we hypothesized that some component of the isomiR variation could be due to systematic technical noise. In order to address whether base differences found in isomiRs are true biological variations or the result of synthetic artifacts from library preparation or errors in sequencing, we have compared paired end reads with single end reads from smallRNA sequencing, using a dataset generated from circulating RNA extracted from serum of healthy individuals and lung cancer patients. We have detected non-negligible systematic differences between single and paired end data which primarily affect putative internally edited isomiRs, and at a much smaller frequency terminal length changing isomiRs . This is relevant for the identification of true isomiRs in small RNA sequencing datasets. Differential expression analysis of cancer versus healthy appears to be only slightly affected by these artifactual reads, since most are detected at very low counts and filtered out by commonly used count thresholding steps prior to statistical inference and the universal filtering applied by DESeq2. We conclude that potential artifacts derived from library preparation and data processing could result in an overestimation of abundance and diversity of miRNA isoforms. Efforts in annotating the isomiRnome should take this into account.

HIGH-THROUGHPUT ANALYSIS OF LNCRNA KNOCKDOWN EXPERIMENTS IN PILOT PHASE OF FANTOM6 PROJECT

Ramil Nurtdinov[1] and Roderic Guigó[1,2]

1. Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Catalonia, Spain.

2. Universitat Pompeu Fabra (UPF), Barcelona 08002, Catalonia, Spain.

Presenting author: ramil.nurtdinov@crg.eu

Our group actively participates to FANTOM collaboration. The current 6-th round is dedicated to lncRNAs and their function in the cell. The collaboration used antisense oligonucleotides to systematically suppress approximately four hundreds of lncRNAs. The knockdown effect was quantified using cell proliferation assays and corresponding transcription changes were characterized using CAGE (Cap Analysis Gene Expression) profiling. Our group, among many others, participates to analysis of these data. For each lncRNAs five different regions were targeted by antisense oligonucleotides and in many cases different knockdowns resulted in different growth rates.

We developed an algorithm that clusters the knockdown experiments for each lncRNAs based on growth rates. Quite frequently the algorithm produced two clusters one of which shows substantial alteration of growth rate, while the second cluster includes control experiments. This difference indicates that the proper target position of antisense oligonucleotides is crucial for their function indicating possible role of alternative splicing or other post-transcriptional modification.

Clustering of the phenotyping effect of different knockdowns has an effect on subsequent analyzes. First, it allows to ignore the experiments with inefficient designs or merge them with control experiments. Second, it allows to treat all the experiments with substantial effect as different replicates to increase the statistical power of any downstream analyzes.

REVERSION OF TRANSCRIPTOMIC SIGNATURES IN NOVEL ALZHEIMER'S DISEASE CELLULAR MODELS GENERATED BY GENE EDITING

Eduardo Pauls[1], Miquel Duran-Frigola[1], Víctor Alcalde[1], Sergi Bayod[1], Samira Jaeger[1] and Patrick Aloy[1,2]

1. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain.
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain.
Presenter e-mail: eduardo.pauls@irbbarcelona.org

Recurrent failure of clinical trials in the Alzheimer's disease (AD) field highlights the need for new approaches to target complex diseases. Using CRISPR/Cas9-mediated gene editing we have generated neuron-like cells harboring mutations known to cause AD (familial form) in the APP and PSEN1 genes. As expected, the presence of AD mutations affected the secretion of the amyloid-beta peptide, increasing the ratio between of the more toxic, long form of the peptide (1-42) and the short form (1-40; Ab42/Ab40 ratio) which is a hallmark of familial AD mutations. Moreover, genome-wide gene expression analysis of mutated cells indicated dysregulation of pathways previously linked to APP or PSEN1 protein function and AD. Taking advantage of the LINCS L1000 repository contained in our Chemical Checker framework followed by several filtering steps we identified a list of 35 compounds with the potential to revert the AD-specific signatures. We then selected 3 of them based on their effect inhibiting amyloid beta secretion, suggesting a direct relationship with AD phenotype. When we compared the transcriptomic signatures of AD mutant cells treated with CMP1, 2 or 3 we observed a clear, significant reversion of the AD transcriptomic patterns. The combination between appropriate cellular models and the Chemical Checker framework may serve a systematic starting point for drug discovery in complex diseases.