V Jornada de Bioinformàtica i Genòmica

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

Patrocinada per:

PROGRAMA I RESUMS DE LES COMUNICACIONS

HOSPITAL UNIVERSITARI VALL D'HEBRON

Sala d'Actes del Pabelló Docent

Passeig de la Vall d'Hebron 119-129

Barcelona

20 de desembre de 2017

COMITÈ ORGANITZADOR:

Xavier de la Cruz (ICREA, VHIR)
Lauro Sumoy (PMPPC-IGTP)
Mario Cáceres (ICREA, UAB)
Roderic Guigó (CRG-UPF)
Ana Ripoll (UAB, BIB)

SUPORT:

Mariàngels Gallego (SCB)
Maite Sánchez (SCB)
Eva Alloza (BIB)

Programa

8:30 – 9:15 Registration

9:15 – 9:30 Wellcome and opening of the symposium
Dra. Laia Arnal (Directora de Desenvolupament de Negoci del Institut de Recerca Vall d'Hebron (VHIR))
Dra. Ana Ripoll (Presidenta Bioinformatics Barcelona Association - BIB)
Dr. Marc Martí-Renom (Tresorer de la Societat Catalana de Biologia)

9:30 - 10:15 **Invited Lecture: Mateo Valero** (Barcelona Supercomputing Center-Centro Nacional de Supercomputación). From classical to runtime aware computer architectures.
Chair: Ana Ripoll (UAB, BIB)

SESSION I. CHROMATIN STRUCTURE AND FUNCTION
Chair: Marc Martí-Renom (CNAG-CRG)

10:15 - 10:30 **Emanuele Raineri** (CNAG-CRG). Inference of genomic spatial organization from methylation samples.
10:30 - 10:45 **Marco Di Stefano** (CNAG-CRG). Exploring the time dependent structural rearrangements of SOX2 locus in mouse using the TADdyn tool.
10:45 - 11:00 **Elena Álvarez de la Campa** (VHIR). Are TGFß-responsive genes confined into specific TADs?

11:00 - 11:30 Coffee break

SESSION II. SEQUENCE VARIABILITY: EVOLUTION AND DISEASE
Chair: Xavier de la Cruz (ICREA, VHIR)

11:30 - 11:45 **Joan Frigola** (IRB Barcelona). Reduced mutation rate in exons due to differential mismatch repair.
11:45 - 12:00 **Roger Mulet** (UAB & Erasmus University Medical Center). PopHuman: the human population genomics browser.
12:00 - 12:15 **Bernat Gel** (IGTP). A data analysis pipeline and platform for the genetic diagnostics of hereditary cancer.

12:15 - 13:00 **Invited Lecture: Mauno Vihinen** (Lund University). Accurate prediction methods for variation interpretation.

13:00 – 13:05 Presentation of the Interuniversity PhD Programme in Bioinformatics.
Dr. Xavier Daura (ICREA & UAB, PhD programme coordinator)

13:05 - 14:00 Lunch

13:30 - 14:30 Gathering of clinical bioinformaticians

14:00 - 14:30 Poster viewing with authors

SESSION III. FROM APPLICATIONS TO THE BASICS

Chair: Alex Sánchez (VHIR, UB)

14:30 - 14:45 **Lourdes Peña-Castillo** (Memorial University of Newfoundland). Type 1 diabetes patients sub-group discovery associated with complications.

14:45 - 15:00 **Enrique Marcos** (IRB Barcelona). De novo computational design of proteins for targeting small-molecules and nucleosomal dna.

15:00 - 15:15 **Teresa Juan-Blanco** (IRB Barcelona). Rationalizing drug response in cancer cell lines.

15:15 - 15:30 **Albert Sorribas** (UdL). Quantitative design principles of yeast metabolism during adaption to heat shock.

15:30 - 15:45 **Eduard Ocaña** (IBE). Origin and evolution of eukaryotic nitrate assimilation: its occurrence in unicellular relatives of animals.

15:45 - 16:00 **Hafid Laayouni** (IBE). Network topology and the evolution of enzyme-coding genes.

16:00 - 16:30 Coffee break

SESSION IV. UNDERSTANDING GENE REGULATION

Chair: Lauro Sumoy (PMPPC-IGTP)

16:30 - 16:45 **Albert Pla Planas** (University of Oslo). A deep learning approach for miRNA target prediction: Exploring the importance of pairing beyond seed region.

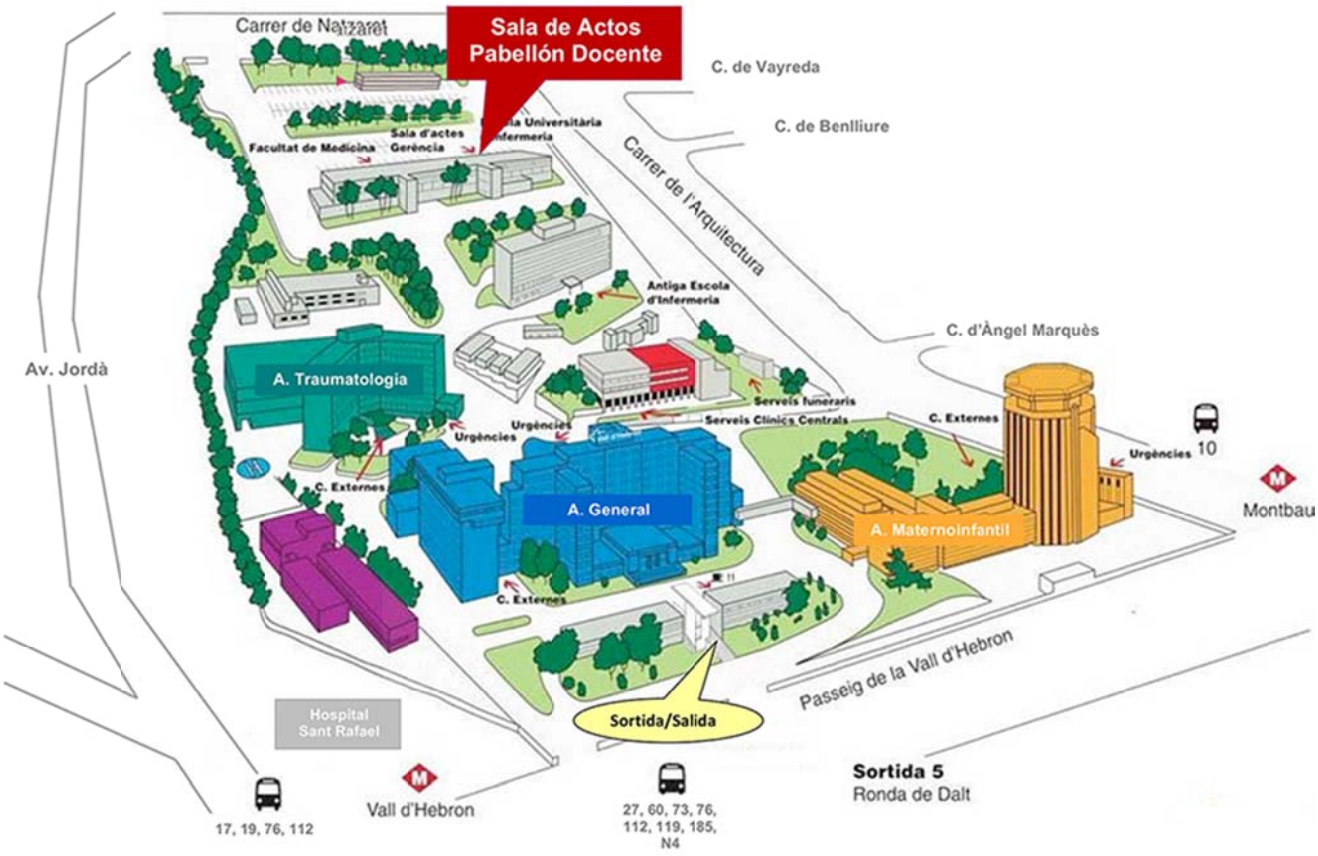
16:45 - 17:00 **Sílvia Pérez-Lluch** (CRG). Natural non-coding antisense transcription along development and evolution.

17:00 - 17:15 **Ivan Erill** (University of Maryland). Comparative genomics analysis of prokaryotic regulatory networks with CGB.

17:15 - 18:00 **Invited Lecture: Eileen Furlong** (European Molecular Biology Laboratoty (EMBL)). Genome regulation during developmental transitions: Generating robustness and precision.

18:00 - 19:00 Poster viewing with authors and beer session

19:00 *Genes* award to the best oral communication and poster and end of the symposium.
Dr. Roderic Guigó (Coordinador Secció de Biologia Computacional I Bioinformàtica de la Societat Catalana de Biologia)
Dr. Mario Cáceres (Coordinador Secció de Genòmica i Proteòmica de la Societat Catalana de Biologia)


Sala de Actos
Pabellón Docente
Carrer de Natzaret
C. de Vayreda
C. de Benlliure
Carrer de l'Arquitectura
C. d'Àngel Marquès
Av. Jordà
Facultat de Medicina
Sala d'actes
Gerència
Antiga Escola d'Infermeria
A. Traumatologia
Urgències
C. Externes
Serveis funeraris
Serveis Clínics Centrals
A. General
A. Maternoinfantil
Urgències
10
Montbau
Passeig de la Vall d'Hebron
Sortida/Salida
Hospital Sant Rafael
Vall d'Hebron
17, 19, 76, 112
27, 60, 73, 76, 112, 119, 185, N4
Sortida 5
Ronda de Dalt

Oral presentations

INFERENCE OF GENOMIC SPATIAL ORGANIZATION FROM METHYLATION SAMPLES

Emanuele Raineri[1], François Serra[1], Renee Beekman[2], Roser Vilarrasa-Blasi[2], Iñaki Martin-Subero[2], Marc A. Marti-Renom[1], Ivo Gut[1], Simon Heath[1]

1 CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain.
2 Biomedical Epigenomics Group, Institut d'Investigacions Biomédiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.
Presenter email: emanuele.raineri@cnag.crg.eu

The 3D configuration of the genome in the nucleus is a cell-type specific feature and represents an important layer to understand its regulation. However, typical approaches to characterize the spatial structure of the genome, such as the different Chromosome Conformation Capture methods, are still cumbersome and expensive, making them not broadly available. Therefore, inferring the 3D structure based on linear genetic and DNA methylation information would be an important resource to exploit the potential of such data as generated by large scale consortia. Here, we show a method to predict the values of the first eigenvector of the HiC matrix for a sample (and hence the positions of the A and B compartments) using only the GC content of the sequence and a single whole genome bisulfite sequencing (WGBS) experiment on the same sample. We train and test our model on 10 datasets for which we have both WGBS and HiC data; we then run the model on 206 samples produced by the Blueprint consortium (which include the 10 used for training) and use chromatin data to confirm that the predicted compartments match the active and repressed genomic regions. Our model takes into account the differences between chromosomes and reconstructs the eigenvector with enough accuracy that it can be used to look at structures which might be intermediate between A and B. Potential applications include the study of how the boundaries between active and repressed regions change in cancer.

EXPLORING THE TIME DEPENDENT STRUCTURAL REARRANGEMENTS OF SOX2 LOCUS IN MOUSE USING THE TADDYN TOOL

Marco Di Stefano[1,2,3], Ralph Stadhouders[2,3], Thomas Graf[2,3], Marc A. Marti-Renom[1,2,3,4]

1. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain.
2. Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.
3. Universitat Pompeu Fabra (UPF), Barcelona, 08010, Spain.
4. ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain.
Presenter e-mail: marco.distefano@cnag.crg.eu

Recent Chromosome Conformation Capture (3C) experiments revealed the genomic organization during reprogramming from B cells to pluripotent stem cells in mouse (Stadhouders, R., Vidal, E. *et al*. 2017, Nature Genetics, in press). These studies suggested that Sox2 region changes its organization before transcriptional activation. Here, we integrate 3C interaction data and molecular dynamics simulations to unveil the relationship between the time dependent structural rearrangements of Sox2 locus and its activation during reprogramming. We found that Sox2 region undergoes a major structural transition mainly driven by the "caging" of the Sox2 locus inside a small structural domain. Interestingly, the domain formation insulates the Sox2 and its super-enhancer region from the rest of the structure promoting specific interactions between them. This cage effect stabilizes the local dynamics of Sox2 locus during transcription activation.

TGFß-RESPONSIVE GENES ARE CONFINED INTO SPECIFIC TADS?

Elena Álvarez de la Campa[1], Raquel Fueyo[2], Claudi Navarro[2], Simona Iacobucci[2], Sara de la Cruz-Molina[2], Álvaro Rada-Iglesias[2], Xavier de la Cruz[1] and Marian Martínez-Balbás[2]

1. Vall d'Hebron Institute of Research (VHIR), Passeig de la Vall d'Hebron, 119; E-08035 Barcelona, Spain. Institut Català per la Recerca i Estudis Avançats (ICREA). Barcelona 08018, Spain.
2. Department of Molecular Genomics. Instituto de Biología Molecular de Barcelona (IBMB), Consejo Superior de Investigaciones Científicas (CSIC), Barcelona 08028, Spain.
3. Center for Molecular Medicine Cologne (CMMC), University of Cologne, Robert-KochStrasse 21, 50931 Cologne, Germany.
Presenter e-mail: elena.alvarezdelacampa@gmail.com

Topological associated domains (TADs) are a three-dimensional chromosome structures stable across different cell types and highly conserved across species. An important biological property of TADs, is that they restrict the accessibility of transcription factors to their gene-specific binding sites, and regulate the potential enhancer-promoter interactions occurring at a specific location and at a given moment. Specific subsets of genes and enhancers are activated by signaling cascades that govern tissue transcriptional programs, like for the TGFß-pathway (Long *et al*., 2016). Within this context, we know now that enhancers and the genes they regulate are normally located within the same TAD (Dixon *et al*.; 2016). In this work, we are interested in these two characteristics (transcription factor accessibility and TAD co-location of enhancers and genes) of the chromatin-based regulation of gene expression, and study how TGFß-pathway regulates transcription from a topological point of view. Firstly, we use HiC techniques, we demonstrate that the transcriptional program regulated by TGFß in a neural stem cell model is confined into topologically restricted domains (TADs). We present a subset of 55 TADs that are enriched in genes responding to the TGFß pathway, meaning that the genes that these responsive genes are not randomly localized in the chromatin 3D-context. Secondly, we test the idea that the TGFß enriched TADs should also be enriched in TGFß-responsive enhancers (Fueyo *et al*.; 2017). More precisely, we show that the average number of enhancers falling within TADs significantly enriched in TGFß-upregulated genes is higher than for those hosting downregulated ones.

REDUCED MUTATION RATE IN EXONS DUE TO DIFFERENTIAL MISMATCH REPAIR

Joan Frigola[1,2*], Radhakrishnan Sabarinathan[1,2*], Loris Mularoni[1,2], Ferran Muiños[1,2], Abel Gonzalez-Perez[1,2], Núria López-Bigas[1,2,3,†]

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.

* These authors contributed equally to the work
†Corresponding author e-mail: nuria.lopez@irbbarcelona.org
Presenter e-mail: joan.frigola@irbbarcelona.org

While recent studies have revealed higher than anticipated heterogeneity of mutation rate across genomic regions, mutations in exons and introns are assumed to be generated at the same rate. Here we find fewer somatic mutations in exons than expected based on their sequence content, and demonstrate that this is not due to purifying selection. Moreover, we show that it is caused by higher mismatch repair activity in exonic than in intronic regions. Our findings have important implications for our understanding of mutational and DNA repair processes, our knowledge of the evolution of eukaryotic genes, and practical ramifications for the study of the evolution of both tumors and species.

POPHUMAN: THE HUMAN POPULATION GENOMICS BROWSER

Roger Mulet[1*], Sònia Casillas[1], Pablo Villegas-Miron[2], Sergi Hervas[1], Esteve Sanz[3], Daniel Velasco[1], Jaume Bertranpetit[2], Hafid Laayouni[2,4] and Antonio Barbadilla[1,3]

1. Institut de Biotecnologia i de Biomedicina and Department de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain.
2. Institute of Evolutionary Biology (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, PRBB, Barcelona, Spain.
3. Servei de Genòmica i Bioinformàtica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain.
4. Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, Barcelona, Spain.
*Present Address: Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands.
Presenter email: r.muletlazaro@erasmusmc.nl

The 1000 Genomes Project (1000GP) represents the most comprehensive world-wide nucleotide variation data set so far in humans, providing the sequencing and analysis of 2504 genomes from 26 populations and reporting >84 million variants. The availability of this sequence data provides the human lineage with an invaluable resource for population genomics studies, allowing the testing of molecular population genetics hypotheses and eventually the understanding of the evolutionary dynamics of genetic variation in human populations.
Population genomics analyses of the 1000GP data can be largely facilitated by (i) making an inventory of parameter values along the chromosomes that capture the evolutionary properties of the available sequences, and (ii) allowing the query and visualization of these estimates in a genome browser designed specifically for this data. Here we present PopHuman, a new population genomics-oriented genome browser that allows the interactive visualization and retrieval of an extensive inventory of population genetics metrics. These have been computed using a novel pipeline that faces the unique features and limitations of the 1000GP data, and include a battery of nucleotide variation measures, divergence and linkage disequilibrium parameters, as well as different tests of neutrality. All metrics have been estimated in non-overlapping windows along the chromosomes as well as in annotated genes for all 26 populations of the 1000GP. PopHuman is open and freely available at http://pophuman.uab.cat/.
The PopHuman database and browser go a step forward in the description and analysis of the most comprehensive human diversity data to date from a population genomics perspective. Furthermore, we aim to extend PopHuman with novel metrics of transcriptomic and epigenomic variation, not only across individuals and species but also during the lifetime of an individual and/or in different parts of the body. In this way, PopHuman will become a pioneer population multi-omics browser advancing the upcoming population–omics synthesis.

A DATA ANALYSIS PIPELINE AND PLATFORM FOR THE GENETIC DIAGNOSTICS OF HEREDITARY CANCER

Bernat Gel[1,3], José Marcos Moreno-Cabrera[1,2,3], Elisabeth Castellanos[1,3], Inma Rosas[1], Eva Tornero[2], Jesús del Valle[2], Marta Pineda[2], Lidia Feliubadaló[2], Gabi Capellá[2,3], Conxi Lázaro[2,3] and Eduard Serra[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, L'Hospitalet de Llobregat, Spain.
3. CIBERONC, Instituto de Salud Carlos III, Madrid, Spain.
Presenter e-mail: bgel@igtp.cat

The I2HCP is an NGS strategy to globally analyze hereditary cancer genes with diagnostic quality. It consists of a custom gene panel with 135 genes associated to hereditary cancer, a diagnostics oriented data analysis pipeline and a data analysis platform called Pandora. The strategy has been implemented into the routine diagnostics of hereditary cancer at the Catalan Institute of Oncology (ICO) and at the Germans Trias i Pujol Research Institute (IGTP). So far it has been used to study more than 1700 individuals.
The analysis pipeline has been developed focusing on the specificities of a clinical setting: adding additional quality controls, adjusting the sensitivity/specificity, aiming for total accountability and reproducibility, etc. but uses standard and well tested tools. Data and metadata produced by the pipeline is stored in a centralized database.
To manage the diagnostics workflows in our labs, we have developed Pandora, a data analysis platform that plays a central role as the coordinator of the diagnostics processes. Pandora automates repetitive tasks such as downloading the data from the sequencers or launching the analysis pipeline and offers a rich web-based platform for the users to further analyse the pipeline results: validate and classify the variants, identify low-coverage regions, etc. Users usually work only with the variants in the set of genes determined by the clinical suspicion of each individual, but if needed and allowed by the patient's consent, they can open their view and study variants in any gene from the panel. There are currently more than 3400 manually classified variants in the database and for each new sample less than 5% of the variants have to be classified.
The I2HCP data analysis pipeline and the Pandora analysis platform have played a key role in the transition of our genetic diagnostics labs from Sanger to NGS.

TYPE 1 DIABETES PATIENTS SUB-GROUP DISCOVERY ASSOCIATED WITH COMPLICATIONS

S. Sadra Mirhendi[1], Sharon Smith[3], Leigh-Anne Newhook[3], and Lourdes Peña-Castillo[1,2]

1. Department of Computer Science, Memorial University of Newfoundland, St. John's, Canada.
2. Department of Biology, Memorial University of Newfoundland, St. John's, Canada.
3. Department of Pediatrics, Memorial University of Newfoundland, St. John's, Canada.
Presenter e-mail: lourdes.pena@crg.eu, lourdes@mun.ca

Type 1 diabetes mellitus (T1DM) is one of the most common chronic diseases in childhood and results from autoimmune destruction of pancreatic ß-cells, leading to insulin deficiency. Using a large amount of heterogenous data including demographic, clinical and genetic data from 197 T1DM patients, we aim to identify T1DM patients sub-groups associated with higher risk of developing T1DM complications or comorbidities. To do this, we applied two state of the art unsupervised machine learning approaches, namely Generalized Low Rank Models (GLRM) and Similarity Network Fusion (SNF), to integrate this data and identify patients sub-groups by clustering. Patients sub-groups identified were evaluated in terms of over-enrichment of patients with a specific complication in a sub-group. Our results indicate that is possible to identify T1DM patients sub-groups with higher risk of developing complications such as hyperglycemia, hypoglycemia, nerve damage, thyroid disease, and dyslipidemia.

DE NOVO COMPUTATIONAL DESIGN OF PROTEINS FOR TARGETING SMALL-MOLECULES AND NUCLEOSOMAL DNA

Enrique Marcos[1,2], Benjamin Basanta[2], Tamuka M.Chidyausiku[2], Gaetano T. Montelione[3], David Baker[2], Modesto Orozco[1]

1. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac 10, 08028, Barcelona, Spain.
2. Institute for Protein Design, University of Washington, Seattle, WA, 98195, USA.
3. Department of Biochemistry and Molecular Biology, The State University of New Jersey, Piscataway, NJ, 08854, USA.
Presenter e-mail: emarcos82@gmail.com

Current strategies to engineer proteins for biocatalysis, molecule biosensing or therapeutics, rely on finding existing proteins having already a similar function or, at least, a suitable geometry and enough stability to tolerate mutations to achieve the new function. This dependence on existing protein structures can be a limitation for certain applications and, instead, computationally designing proteins, de novo, with custom-made structures should be more effective. Here we have used Rosetta to develop two de novo design approaches to tackle different biotechnological and biomedical problems. For novel custom-made small-molecule binders and enzymes, we have developed a computational approach to de novo design thermostable protein folds with cavities formed by curved ß-sheets, which were validated with NMR and X-ray crystallography [1]. We explored stabilization strategies based on disulfide bonds and homodimer interfaces that allow balancing the incorporation of cavity-forming functional mutations that are detrimental for protein stability. This approach allows us to control the size and shapes of protein cavities that can be used to install novel ligand-binding and catalytic sites into stable scaffolds. For novel chromatin-based drugs and research, we have computationally designed helical bundles and TAL-like proteins to target the exposed face of nucleosomal DNA with specificity. This may allow to block the access of certain types of pioneer transcription factors, which confer faster activation of gene expression and can promote the growth of cancer cells, to nucleosomal DNA. As no structural information is available for such specific nucleosomal DNA binding, we have de novo designed protein-nucleosome interfaces that have not been observed in nature yet.

References:
1. E. Marcos, B. Basanta, T.M. Chidyausiku, *et al*. “Principles for designing proteins with cavities formed by curved ß sheets”. *Science* 355, 201-206 (2017).

RATIONALIZING DRUG RESPONSE IN CANCER CELL LINES

Teresa Juan-Blanco[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. Institició Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain.
Presenter e-mail: teresa.juanblanco@irbbarcelona.org

Cancer cell lines (CCLs) play an important role in the initial stages of drug discovery allowing, among others, for the high-throughput screening of drug candidates. Many polymorphisms in genes encoding drug-metabolizing enzymes, transporters and drug targets, as well as disease-related genes have been linked to altered drug sensitivity. Yet, identifying the correlation between this variability and pharmacological responses remains challenging. Here, we propose a system biology method to identify the mechanisms that may affect drug sensitivity in CCLs. In particular, we exploited somatic mutations, gene expression and drug response data provided by the Cancer Cell Line Encyclopedia. We integrated these molecular profiles with protein-protein interaction data to detect groups of CCLs with similarly perturbed network regions and that present similar drug responses. Furthermore, we identified genes expression signatures that might be responsible for the differential drug response and that are beneficial for predicting drug sensitivity in CCLs.

QUANTITATIVE DESIGN PRINCIPLES OF YEAST METABOLISM DURING ADAPTION TO HEAT SHOCK

Tania Pereira[1,2],*, Ester Vilaprinyo[1,2],*, Gemma Belli[1,2], Enric Herrero[1], Baldiri Salvado[1,2], Albert Sorribas[1,2], Gisela Altés[1,2], Rui Alves[1,2]

1.Institute of Biomedical Research of Lleida IRBLleida, 25198, Lleida, Catalunya, Spain.
2. Departament de Ciències Mèdiques Bàsiques, University of Lleida, 25198, Lleida, Catalunya, Spain.
* These authors contributed equally to this work

Summary: Microorganisms evolved adaptive responses to survive stressful challenges in ever changing environments. Understanding the relationships between the physiological/metabolic adjustments allowing cellular stress adaptation and gene expression changes being used by organisms to achieve those adjustments may significantly impact our ability to understand and/or guide evolution. Here, we studied those relationships during stress adaptation in *Saccharomyces cerevisiae*, focusing on heat shock responses.
We combined dozens of independent experiments measuring whole genome gene expression changes during stress response with a simplified kinetic model of central metabolism. We identified physiological variables and genes whose changes permit adaptation to heat shock and desiccation/rehydration.
Further, we identified the quantitative ranges for those changes that specifically allow yeast to adapt to each of the two stresses. Our approach is scalable to other adaptive responses and could assist in developing biotechnological applications to manipulate cells for medical, biotechnological, or synthetic biology purposes.

Keywords: Biological Design Principles/Systems Biology/Computational Biology/Multilevel modelling/Integrative Biology

ORIGIN AND EVOLUTION OF EUKARYOTIC NITRATE ASSIMILATION: ITS OCCURRENCE IN UNICELLULAR RELATIVES OF ANIMALS

Eduard Ocaña-Pallarès[1], Sebastian R. Najle[1,2], Iñaki Ruiz-Trillo[1,3,4], Claudio Scazzocchio[5,6]

1. Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, Barcelona 08003, Catalonia, Spain.
2. Instituto de Biología Molecular y Celular de Rosario (IBR) CONICET and Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Ocampo y Esmeralda s/n, Rosario S2000FHQ, Argentina.
3. Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Avinguda Diagonal 645, Barcelona 08028, Catalonia, Spain.
4. ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Catalonia, Spain.
5. Department of Microbiology, Imperial College, London, United Kingdom.
6. Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France.

Presenter e-mail: ed3716@gmail.com

The incorporation of nitrogen from the atmosphere is energetically demanding, nitrate being the most oxidized form that can be assimilated by some eukaryotes. We used phylogenetic inference and sequence-similarity networks to study the origin, evolution and distribution of the gene families specifically involved in the incorporation of nitrate among an updated sampling of eukaryotic genomes. We show that the main family of nitrate transporters and the two families of nitrite reductases (NIR) described in eukaryotes are of bacterial origin, while the nitrate reductase (NR) was originated through the fusion of three different genes. The resulting phylogenetic trees and the patchy distribution suggest that horizontal gene transfer (HGT) played an important role in the evolution of these gene families. Among the recently available genomes analyzed, we detected the presence of nitrate-related gene families among three unicellular relatives of animals that belong to Teretosporea. In two of them, the canonical C-terminal region of the NR is replaced by the N-terminal duplicated domain of the NIR gene, which is found downstream to NR. In the third teretosporean investigated, we found the transporter and NIR genes but not the NR, an unexpected pattern present also in a distantly related species from the red algae group. Interestingly, both genomes are the only in our sampling containing an uncharacterized putative molybdopterin oxidoreductase physically linked to nitrate-related genes, possibly acting as a NR. Finally, two experimental approaches were carried out: (i) we cultured the three teretosporean organisms on different media to check if they are able to grow on nitrate as a sole nitrogen source and (ii) we quantified with RT-qPCR the expression of nitrate-related genes to understand their regulation in response to different nitrogen sources. Results and implications will be discussed.

NETWORK TOPOLOGY AND THE EVOLUTION OF ENZYME-CODING GENES

Hafid Laayouni[1,2], Begoña Dobon[1], Ludovica Montanucci[1], Juli Peretó[3] and Jaume Bertranpetit[1]

1. IBE Evolutionary Biology Institute (CSIC-UPF), Universitat Pompeu Fabra, PRBB, Doctor Aiguader, 88, Barcelona, Catalonia, Spain.
2. Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, 08003 Barcelona, Catalonia, Spain.
3. Institute of Systems Biology, Department of Biochemistry and Molecular Biology, University of Valencia, Valencia, Spain.
Presenter e-mail: hafid.laayouni@upf.edu

Metabolic networks comprise thousands of enzymatic reactions functioning in a controlled manner and have been shaped by natural selection. Thanks to the genome data, the footprints of adaptive (positive) selection are detectable, and the strength of purifying selection can be measured. This has made possible to know where, in the metabolic network, adaptive selection has acted and where purifying selection is more or less strong and efficient. We have carried out a comprehensive molecular evolutionary study of all the genes involved in the human metabolism. We investigated the type and strength of the selective pressures that acted on the enzyme-coding genes belonging to metabolic pathways during the divergence of primates and rodents. Then, we related those selective pressures to the functional and topological characteristics of the pathways. The effect of pathway topology and functional category on gene evolution seems to vary depending on the time-scale of the action of natural selection, and some local trends could be replicated at a global scale.

A DEEP LEARNING APPROACH FOR MIRNA TARGET PREDICTION: EXPLORING THE IMPORTANCE OF PAIRING BEYOND SEED REGION

Albert Pla[1], XiangFu Zhong[1,2], Fatima Heinicke[1,2], and Simon Rayner[1,2]

1. Department of Medical Genetics, University of Oslo, Kirkeveien 166 (bygg 25), 0407 Oslo, Norway.
2. Department of Medical Genetics, Oslo University Hospital, Kirkeveien 166 (bygg 25), 0407 Oslo, Norway.
Presenter e-mail: a.p.planas@medisin.uio.no

MicroRNAs (miRNAs) are a family of small non-coding RNAs that regulate gene expression by binding to partially complementary regions within their target genes. Computational methods play an important role in predicting potential miRNA targets, but typically they only identify approximately 80% of known bindings. This is a consequence of a core assumption of target prediction tools, which assume that it is the miRNA seed region (nt 2 to 8) that defines the key interactions between a miRNA and its target. Nonetheless, recent studies support the importance of nucleotide pairing beyond seed region, indicating that the entire miRNA is involved in the targeting process and pointing the need of a more flexible prediction methodology.
To investigate the role of non-canonical sites in the targeting process we adopted a novel approach based on Deep-Learning (DL) which rather than basing predictions in current assumptions (e.g seed region), investigates the entire miRNA and 3'UTR mRNA target nucleotides. We used more than 150,000 experimentally validated human miRNA:gene targets to train a DL network that automatically learns a set of features describing the targeting process. The consideration of the whole miRNA:mRNA transcript allows us to study the impact of mutations in the target site region and how miRNA isoform variations can affect the strength and functionality of targets.
Results show that this unbiased approach recognizes the seed region as a key feature in the targeting process, but also demonstrates that pairings beyond seed region also play an important role. In addition, thermodynamic analysis of the result suggests a link between low site accessibility energy and functionality in non-canonical targets, whilst this link appeared weaker in canonical sites. Regarding the study of isomiR targets, we found there was a major impact in variations affecting the miRNA seed region, but we also identified changes in the target behavior when variations involved the miRNA 3'end

NATURAL NON-CODING ANTISENSE TRANSCRIPTION ALONG DEVELOPMENT AND EVOLUTION

Sílvia Pérez-Lluch[1,2], Alessandra Breschi[1,2], Cecilia Klein[1,2], Marina Ruiz-Romero[1,2], Emilio Palumbo[1,2], Amaya Abad[1,2], Carme Arnan[1,2] and Roderic Guigó[1,2,3].

1. Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Catalonia, Spain.
2. Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain.
3. Institut Hospital del Mar d'Investigacions Mediques (IMIM), Barcelona 08003, Catalonia, Spain.

Presenter e-mail: silvia.perez@crg.eu

The genome of *Drosophila melanogaster* is estimated to encode over two thousand lncRNAs; however, only few of them have a characterized function. Natural antisense transcripts (NATs) are fully processed lncRNAs which overlap protein coding genes on the opposite strand with or without exonic complementarity. Several roles in genomic regulation are reported for NATs in metazoa, including gene expression regulation of the overlapping protein coding gene, DNA methylation, chromatin modifications and RNA editing. Here, we have identified 855 lncRNAs overlapping 873 protein coding genes in antisense orientation, forming 953 sense-antisense (SA) pairs in the fruit fly genome. By analysing the transcriptome of different imaginal tissues at 3rd instar larvae, we have explored the relationship between NATs expression and alternative transcript usage across fly larval samples. Of the 376 SA expressed pairs involving a protein coding gene with multiple isoforms, *blistered*/*CR44811* is the one showing a highest correlation between changes in coding gene isoform usage and NAT expression. *blistered* (*bs*) gene encodes for two main isoforms: a short one, expressed mainly in the wing where the NAT *CR44811* is also expressed, and a long one, expressed in the other tissues where the NAT is silent. *CR44811* CRISPR mutant flies show a dramatic change in the *bs* isoform usage in larval and pupal wings, as well as a strong phenotype in the adult, indicating impairment of the proper wing development. Manual annotation of the *bs* locus using available RNAseq data from other species, has allowed us to align both isoforms of the coding gene as well as the lncRNA along development. We have been able to track the presence of the two isoforms of *bs* and until crustaceans, indicating that the usage of the two proteins is not restricted to wing development in flies.

COMPARATIVE GENOMICS ANALYSIS OF PROKARYOTIC REGULATORY NETWORKS WITH CGB

Sefa Kiliç[1], Ivan Erill[2]

1. Google Cloud Platform, Seattle, WA (USA)
2. Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD (USA)
Presenter e-mail: erill@umbc.edu

Transcriptional regulatory networks are the main mechanism governing gene expression and adaptation to environmental changes in Bacteria, yet our understanding of how these networks evolve over time to incorporate new functions and respond to novel challenges, such as virulence determinants and antibiotics, remains fairly limited. He were report on the development of a generic comparative genomics system for the analysis of transcriptional regulatory networks in prokaryotic genomes (CGB) and its application to the analysis of stress response systems in different bacterial groups. Based on a formal Bayesian inference framework, the CGB modular system enables the integration of complete and draft genome sequences for fast comparative analysis of regulatory regions and performs ancestral state reconstruction of regulatory loci to discern the evolutionary trajectory of the regulatory network. Our results illustrate the flexibility of the CGB platform to deal with draft genomic sequence and distributed sources of experimental information on the specificity of the transcription factor-binding motif within a bacterial clade. Using the bacterial SOS response to DNA damage as a model transcriptional network, our analysis provides further insights into the complex evolutionary history of a conserved regulatory system that integrates mechanisms involved in antibiotic resistance and persistence across the Bacteria domain. Our work also identifies the fundamental regulatory elements of bacterial pathogenesis mediated by the type III secretion system.

Posters

TOWARDS A BEST 'PER GENE' METHOD TO IDENTIFY PATHOGENIC VARIANTS ASSOCIATED WITH METABOLIC DISORDERS

Josu Aguirre Gómez[1], Xavier de la Cruz Montserrat[1,2]

1. Translational Bioinformatics, Vall d'Hebron Research Institute (VHIR), Vall d'Hebron University Hospital, Passeig de la Vall d'Hebron, 119-129,08035 Barcelona, Catalonia, Spain.
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain.
Presenter e-mail: josu.aguirre@vhir.org

To improve the usage of Next-Generation Sequencing (NGS) tools in the clinical we need to finely adapt them to the problem at hand. For example, different in silico pathogenicity predictors have performances that vary, even for the same gene. In this context, selecting a suboptimal predictor for a given gene may affect the diagnostics yield for the diseases associated to this gene. In this work we address this problem, trying to identify the best gene predictor for a set of 58 genes underlying known metabolic disorders. At a technical level, we followed the standard steps in a bioinformatics project: (i) build a training set of pathogenic/neutral variants; (ii) test different tools for this set of variants; and (iii) explore, for which gene, which is the best pathogenicity predictor. We used the training dataset to test the performance of 4 known in silico tools: PolyPhen-2, SIFT, PON-P2 and CADD. For the 58 genes, we could identify the best predictor in each case, although in some cases it had a very low coverage. Overall, we find that most frequently, PolyPhen-2 is the best pathogenicity predictor and also has the highest coverage. We also observe that in some cases only using the best gene predictor we obtain the correct answer for a specific variant. In summary, using the best, per gene pathogenicity predictor can result in a noticeable improvement in the annotation of variants obtained from sequencing methods.

PREDICTION OF DRUG COMBINATIONS USING DRUG TARGETS AND PROTEIN-PROTEIN INTERACTIONS

Joaquim Aguirre-Plans[1], Emre Guney[2], Jordi Mestres[3], Narcís Fernandez-Fuentes[4], Baldo Oliva[1]

1. Structural Bioinformatics Lab, Department of Experimental and Health Science, Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain.
2. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), 08028 Barcelona, Catalonia, Spain.
3. Research Group on Systems Pharmacology, Research Program on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain.
4. Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth SY23 3DA, UK.
Presenter e-mail: joaquim.aguirre@upf.edu

Drug combinations have become the standard treatments against complex diseases, because they are able to tackle compensatory signaling pathways that cause drug resistance, and they are often associated to synergistic effects. However, most of the computational methods to predict drug combinations rely on gene expression data before and after drug administration (1), which is scarce. Here, we explore the possibilities that drug targets and protein-protein interactions provide.
We present three different methods (target-based named dcTargets, network-based named dcGUILD and structure-based named dcStructure) to predict drug combinations using supervised machine learning algorithms. dcTargets uses only drug target information, dcGUILD expands this information using the protein-protein interaction network (2), and dcStructure is target-independent and uses a similarity score of the SMILES of the drugs. We compare the three methods and characterize the cases in which it is better to use drug target or protein-protein information.
In terms of Area Under the ROC curve values, we observe an increase in dcTargets and dcGUILD as we have more drug target information, reaching values over 0.9 when we have 9 targets or more. We also see that the expansion of the drug target information using dcGUILD significantly improves the prediction of drug combinations when the targets of the drugs are from unrelated biological processes.
(1) Madani Tonekaboni,S.A. *et al*. (2016). Br. Bioinform.
(2) Guney,E. and Oliva,B. (2012). PLoS One.

AIR: ARTIFICIAL INTELLIGENCE RNASEQ

Riccardo Aiese Cigliano, Andreu Paytuví-Gallart, Ermanno Battista, Fabio Scippacercola, Walter Sanseverino

Sequentia Biotech SL, Carrer Comte d'Urgell 240, 08036 Barcelona, Spain.
Presenter e-mail: raiesecigliano@sequentiabiotech.com

In the field of genomics, sequencing technologies have drastically changed in the last few years and the output of complex data generated has outpaced the solutions available for analysis, integration and interpretation. RNA Sequencing has emerged as the number one technique in transcriptomics and thus the solution we propose is based on this. A.I.R.: Artificial Intelligence RNASeq is the first easy to use SaaS (Software as a Service) built with solid scientific methods. AIR is able to perform a robust DEG and GEOA analysis with different statistics to solve three important obstacles in the genomics field simultaneously: the informatics problem (specifically data storage, automatization of results and duration of analysis); the scientific problem (data interpretation and data integration, as well as providing new bioinformatics and statistical functions); the social problem (the lack of availability of skilled bioinformaticians). The overall objective of this project is to introduce a disruptive innovation that will allow researchers to perform transcriptomics data analysis easily, quickly and affordably. AIR is accessible at http://transcriptomics.cloud

DISTINCT GENETIC VULNERABILITIES-BY-STRESSFUL LIFE EVENTS INTERACTIONS PROPOSE ADDITIONAL RISK FOR DEPRESSIVE SYMPTOMS

Aleix Arnau-Soler[1], Mark J. Adams[2], Toni-Kim Clarke[2], Donald J. MacIntyre[2], Keith Milburn[3], Lauren B. Navrady[2], Generation Scotland[4], Caroline Hayward[5], Andrew M. McIntosh[2,6], Pippa A. Thomson[1,6]

1. Medical Genetics Section, University of Edinburgh, Centre for Genomic and Experimental Medicine and MRC Institute of Genetics and Molecular Medicine, Edinburgh, UK.
2. Division of Psychiatry, Deanery of Clinical Sciences, University of Edinburgh, Royal Edinburgh Hospital, Morningside Park, Edinburgh EH10 5HF, UK.
3. Health Informatics Centre, University of Dundee, Ninewell Hospital & Medical School, Dundee, UK.
4. A collaboration between the University Medical School and NHS in Aberdeen, Dundee, Edinburgh and Glasgow, Scotland, UK.
5. Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK.
6. Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK.
Presenter e-mail: aleix.arnau.soler@igmm.ed.ac.uk

Stressful life events (SLE) may contribute to liability of depression through interaction with known genetic risk factors or through interaction with novel genetic factors specific to SLE. Polygenic risk scores (PRS), derived from genome-wide association studies (GWAS), reflect an individual's genetic vulnerability. We used PRS for depression and for a measure of genetic contribution to stress-sensitivity, derived from a genome-wide interaction study, to test the variance in depressive symptoms explained by both PRS-by-SLE interactions.
Genetic vulnerability-stress models were tested using PRS weighted by the largest meta-GWAS of depression (PRS_D), or PRS weighted by stress-sensitivity effect (PRS_{SS}) using the STRADL cohort (Stratifying Depression and Resilience Longitudinally; N=4,919), which has data on depressive symptoms and preceding 6 months SLE.
PRS_D-by-SLE interaction significantly contributed to risk of depressive symptoms (R^2=0.08%, P=4.91x10^{-2}), the effect being stronger in females (R^2=0.19%, P=1.84x10^{-2}) and not significant in males (P=6.37x10^{-2}). PRS_{SS}-by-SLE interaction significantly contributed in males (R^2=0.60%, P=3.38x10^{-4}), but not in females (P=0.14) or the full cohort (P=7.75x10^{-2}). The PRS_{SS}-by-SLE interaction in males explained almost as much as the variance explained by main PRS_D effects (R^2=0.66%, P=2.09x10^{-4}) and are additive.
In conclusion, genetic vulnerability-by-stress interactions predict additional risk for individuals with high genetic predisposition and reported SLE, with a potential distinction between sexes. Incorporation of SLE data and genetic predictors based on environmental adversity/stress response should improve clinical prediction of stress-related diseases and may help to identify individuals for targeted support.

CREATION OF A MICROENVIRONMENT-BASED DIAGNOSIS TOOL FOR BREAST CANCER PATIENTS

Rosa Barcelona Cabeza[1], Andreu Paytuví-Gallart[1], Elisa Rivas[2], Alexandre Calon[2], Riccardo Aiese Cigliano[1], Walter Sanseverino[1]

1. Sequentia Biotech SL, Carrer Comte d'Urgell 240, 08036 Barcelona, Spain.
2. Hospital del Mar Medical Research Institute (IMIM), Dr Aiguader, 88, 08003, Barcelona, Spain.
Presenter e-mail: rbarcelona@sequentiabiotech.com

Breast cancer is the second most common cancer worldwide, being the most frequent cancer among women, and the fifth leading cause of cancer death. Therapies are proposed based on the molecular classification of the disease, however, these classification does not fully reflect breast cancer heterogeneity and patients receiving the same diagnosis and treatment can have different outcomes.
We aim to develop a multifaceted molecular classification of the disease based both on cancer cell attributes and stromal and immune features using microarray expression data from 32 breast cancer cohorts from the Gene Expression Omnibus (GEO) repository from the National Center for Biotechnology Information (NCBI). Therefore, this classification will be used to highlight possible information crucial for decision-making regarding immunotherapy and personalized medicine.
The first step will consist on refine existing clinical molecular classification of samples from GEO, mostly performed by immunohistochemistry techniques which may leads to misclassification. In order to do so, we will perform a k-means unsupervised and a supervised classification, allowing to correlate the phenotypic and expression data. Then, relatively to the expression of transcriptomic markers an abundance score will be calculated by immune and stromal population for each molecular subtype. After the corresponding differential expression analysis, we will obtain an immune and stromal profile for each new refined molecular subtype that can be integrated with an immunohistochemistry approach to improve the cancer diagnosis and better understand the patient profile.

APPLICATION OF CLINICAL EXOME SEQUENCING PANEL IN EARLY ONSET IMMUNODEFICIENCY PATIENTS

Batlle-Masó L [1], Mensa-Vilaró A [2,3], Solís-Moruno M [1], Tormo M ¤, Marquès-Bonet T [1], Aróstegui JI [2,3], Casals F ¤

1 Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, 08003 Barcelona, Catalonia, Spain.
2 Functional Unit of Clinical Immunology Hospital Sant Joan de Déu-Hospital Clínic, Barcelona, Catalonia, Spain.
3 Immunology Department. Biomedical Diagnostics Center, Hospital Clínic-IDIBAPS, Barcelona, Catalonia, Spain.
4 Servei de Genòmica, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, 08003 Barcelona, Catalonia, Spain.
Presenter e-mail: laura.batlle@upf.edu

Primary immunodeficiencies are usually difficult to diagnose due to their high phenotypic heterogeneity and variable expression. Diagnosis difficulties can be resolved using genetic testing. Next-generation sequencing is a cost-effective approach to identify likely causative genetic variants. This identification can lead to a better understanding of the disease and increase the number of diagnosed patients. The aim of the study is to use next generation sequencing to detect genetic variants likely to be causative of the disease in pediatric patients. For that, we performed target sequencing (TruSight® One panel, 4811 genes) in 36 samples from patients with clinical suspicion of immune disease. After that, bioinformatics analysis was done to detect likely causative genetic variants. We corroborate our findings using Sanger sequencing in the most relevant candidates. We found likely causative genetic variants of the disease in 15 patients. In three cases we proposed a relationship between the genetic alteration and the observed phenotype. However, further studies and functional validation are needed. We conclude that target sequencing is a suitable approach to detect likely causative variants in primary immunodeficiency patients.

ROMA POPULATION ANCESTRY FROM A WHOLE GENOME SEQUENCE PERSPECTIVE

Erica Bianco[1], Carla Garcia-Fernandez[1], Begoña Dobon-Berenguer[1], Mihai G. Netea[2], Jaume Bertranpetit[1], David Comas[1]

1. Departament de Ciències Experimentals i de la Salut, Institut de Biologia Evolutiva (CSIC-UPF), Universitat Pompeu Fabra, Barcelona, Spain.
2. Department of Internal Medicine, Radboud University Medical Center, Nijmegen, the Netherlands.
Presenter e-mail: erica.bianco@upf.edu

Roma people (aka Gypsies) are the largest minority in Europe. Previous linguistic and genetic studies showed that Roma ancestors left the Northwest part of the Indian subcontinent ~1.5kya. As other Indian populations, Roma genomes exhibit West Eurasian and South Asian ancestry components. However, uniparental studies showed that Roma admixed with Europeans, after their arrival in Europe ~1kya. Therefore, Roma West Eurasian ancestry has two origins: an ancient component, already present before the out of India; and a recent component, due to recent admixture with Europeans. To our knowledge, the distinction between these two components has not been addressed yet.
Using whole genome sequencing data of 46 Roma volunteers, we analyzed Roma demographic history and determined the proportion of West Eurasian ancestry due to recent admixture with Europeans.
We approached Roma complex admixture pattern by fitting different scenarios to real data. We found that the best fit scenarios reflect previous knowledge: Roma are the result of an ancient admixture between ancestral West Eurasians and ancestral South Asians followed by a more recent admixture with Europeans. In the best fit scenarios, Roma ancestors had >50% of West Eurasian ancestry, which increased to ~80% in present day Roma. Best fit scenarios must also include extensive drift (as a result of serial bottlenecks or founder effects).
Our preliminary results confirm that Roma have a complex demographic history and at least two main admixture events occurred. Roma ancestors were already the result of the admixture between West Eurasian and South Asian ancestries. After the out of India, they further admix with Europeans, increasing their West Eurasian ancestry component of >25%.

SiNoPsis: SINGLE NUCLEOTIDE POLYMORPHISMS SELECTION AND PROMOTER PROFILING

Daniel Boloc[1], Natalia Rodríguez[1], Teresa Torres[1], Susana García[1], Anna Gortat[1], Patricia Gassó[1,2,3], Amalia Lafuente[1,2,3], Sergi Mas[1,2,3], Miquel Bernardo Arroyo[2,3,4,5]

1. Department of Pathological Anatomy, Pharmacology and Microbiology, University of Barcelona, Barcelona, Spain.
2. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.
3. Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Madrid, Spain.
4. Barcelona Clinic Schizophrenia Unit (BCSU), Psychiatry Service, Hospital Clínic de Barcelona, Barcelona, Spain.
5. Department Psychiatry and Clinical Psychobiology, University of Barcelona, Barcelona, Spain.
Presenter e-mail: danielboloc@gmail.com

The selection of a single nucleotide polymorphism (SNP) using bibliographic methods can be a very time-consuming task. Moreover, a SNP selected in this way may not be easily visualized in its genomic context by a standard user hoping to correlate it with other valuable information. Here we propose a web form built on top of Circos that can assist SNP-centered screening, based on their location in the genome and the regulatory modules they can disrupt. Its use may allow researchers to prioritize SNPs in genotyping and disease studies. SiNoPsis is bundled as a web portal. It focuses on the different structures involved in the genomic expression of a gene, especially those found in the core promoter upstream region. These structures include transcription factor binding sites (for promoter and enhancer signals), histones and promoter flanking regions. Additionally, the tool provides eQTL and linkage disequilibrium (LD) properties for a given SNP query, yielding further clues about other indirectly associated SNPs. Possible disruptions of the aforementioned structures affecting gene transcription are reported using multiple resource databases. SiNoPsis has a simple user-friendly interface, which allows single queries by gene symbol, genomic coordinates, Ensembl gene identifiers, RefSeq transcript identifiers and SNPs. It is the only portal providing useful SNP selection based on regulatory modules and LD with functional variants in both textual and graphic modes (by properly defining the arguments and parameters needed to run Circos).

EPIGENOME-WIDE ASSOCIATION STUDY IN CHILDHOOD OBESITY

Pol Castellano-Escuder[1], Maria Jesús Leal-Lewitt[1], Marta Ramon-Krauel[2], Carles Lerin[1], Judith Cebrià[1], Ruben Díaz[2], Josep C Jiménez-Chillarón[1]

1. Fundació Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain.
2. Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain.

Presenter e-mail: polcaes@gmail.com

Childhood obesity is one a major Public Health issue. Overweight/obese children have a high risk of being obese as adults and develop other co-morbidities, including type 2 diabetes, cardiovascular disease and several types of cancer. It has been proposed that epigenetic mechanisms might be involved in mediating long-term metabolic dysfunction.

Here we analyzed DNA methylation profiles (Infinium MethylationEPIC BeadChip 850K) in whole blood from 26 obese (zBMI > 2) and 12 control lean pre-pubertal children (zBMI < 1). 109 CpG sites appeared differentially methylated between the two groups (methylation change >10% and FDR value <0.05). 5 out of the 109 targets, were located within the *SPATC1L* (Spermatogenesis and Centriole Associated 1 Like) gene. Next, we performed a Two Sample Mendelian Randomization test, which allows determining whether a particular CpG site is causal or consequence for the disease. Strikingly, we found that 2 of the 5 CpG sites mapping the *SPATC1L* locus are causal for multiple disorders, including not only childhood obesity, but also type 2 diabetes, ulcerative colitis and inflammatory bowel disease. *SPATC1L* expression was similar in whole blood of obese and lean subjects. However, the contiguous genes *COL6A2* (Collagen Type VI Alpha 2 Chain), *LSS* (Lanosterol Synthase), *YBEY* (YbeY Metallopeptidase) and *C21orf58* were differentially expressed between the two groups. Together, we show that childhood obesity is associated to a small change in DNA methylation (109 CpG sites). In our dataset, only 2 CpG sites appeared to play a causative role in the development of the disease. We hypothesize that these CpG sites might mediate disease risk by modulating the expression of physically close genes.

EFECT OF AGE AND OVARIAN RESERVE ON THE TRANSCRIPTOME OF HUMAN OOCYTES

Cornet-Bartolomé D[1,2], Barragán M[1], Pons J[3], Ferrer-Vaquer A[1], , Schweitzer[4], Hubbard J[5], Auer H[5], Rodolosse A[3], Grinberg D[2], Vassena R[1]

1. Clínica EUGIN, Travessera de les Corts 322, 08029 Barcelona, Spain.
2. Departament de Genètica, Facultad de Biologia, Universitat de Barcelona, Spain.
3. Functional Genomics Core, Institute for Research in Biomedicine (IRB) Barcelona, Parc Científic de Barcelona, Baldiri Reixac 10, 08028 Barcelona, Spain.
4. Thermo Fisher Scientific, 3450 Central Expressway, Santa Clara, CA 95051, USA.
5. Functional GenOmics Consulting, Bellavista 53, 08753 Pallejà, Spain.
Presenter e-mail: dcornet@eugin.es

Age and ovarian reserve both affect oocyte developmental competence, i.e. its ability to sustain early embryonic development. Ovarian reserve decreases with age, and oocyte competence decreases with it. The molecular mechanisms underlying a diminished ovarian reserve are poorly characterized. After maturation, oocytes are mostly transcriptionally quiescent, and developmental competence prior to embryonic genome activation relies on maternal RNA and proteins. Total RNA of 36 oocytes from 30 women undergoing oocyte donation was independently isolated, amplified, labeled, and hybridized on HTA 2.0 arrays (Affymetrix). Data were analyzed using TAC software, in four groups, each including nine oocytes, according to the woman's age and antral follicular count (AFC): Young with High AFC; Old with High AFC; Young with Low AFC and Old with Low AFC. qPCR was performed to validate arrays. A set of 30 differentially expressed mRNAs and 168 non-coding RNAs (ncRNAs) were differentially expressed in relation to age and/or AFC. Few mRNAs have been identified as differentially expressed transcripts, and among ncRNAs, a set of Piwi-interacting RNAs clusters (piRNAs-c) and precursor microRNAs (pre-miRNAs) were identified as increased in high AFC and old groups, respectively. Our results indicate that age and ovarian reserve are associated with specific ncRNA profiles, suggesting that oocyte quality might be mediated by ncRNA pathways. We show, for the first time, that several non-coding RNAs, usually regulating DNA transcription, are differentially expressed in relation to age and/or ovarian reserve. Interestingly, the mRNA transcriptome of in vivo matured oocytes remains remarkably stable across ages and ovarian reserve, suggesting the possibility that changes in the non-coding transcriptome might regulate some post-transcriptional/translational mechanisms which might, in turn, affect oocyte developmental competence.

iSkyLIMS, A FRIENDLY ENVIRONMENT TO FACILITATE THE INCORPORATION OF MASSIVE SEQUENCING TO A GENOMICS CORE FACILITY

Luis Chapado[1], Sara Monzón[1], Pedro J. Sola[1], Ana Hernández[1], Ángel Zaballos[2] , Isabel Cuesta[1]

1. Bioinformatics Unit, Core Scientific and Technical Units, Institute of Health Carlos III, Carretera Majadahonda- Pozuelo km2, Majadahonda, Madrid, Spain.
2. Genomics Unit, Core Scientific and Technical Units, Institute of Health Carlos III, Carretera Majadahonda- Pozuelo km2, Majadahonda, Madrid, Spain.

Presenter e-mail: chapado.l@gmail.com

The introduction of massive sequencing (MS) in genomics facilities has meant an exponential growth in data generation, requiring a precise tracking system, from library preparation to fastq file generation, analysis and delivery to the researcher. Software designed to handle those tasks are called Laboratory Information Management Systems (LIMS), and its software has to be adapted to their own genomics laboratory particular needs.
iSkyLIMS is born with the aims to help on the wet laboratory tasks, and implements a workflow that guides genomics labs on their activities from library preparation to data production, reducing potential errors associated to high throughput technology, and facilitating the quality control of the sequencing. Also, iSkyLIMS connects the wet lab with dry lab facilitating data analysis by bioinformaticians.
ISkyLIMS is an open-source software that run on a linux distribution with a web based interface for user interaction, which has been implemented using Django Framework running on Python 3.6, and a MySQL database to store the processed data generates by the Illumina sequencer.
The sequencing runs inside iSkyLIMS are handled in a state machine concept where each run is passing through all possible states, from initial until completed state. Data from MS Illumina platform are fetched, processed and stored in a database which is the base for the quality analysis, statistics information, and reports done afterwards. Keeping all this information centralized and running on a virtual environment allows iSkyLIMS to be a scalable system to fulfill the future needs, since the number of runs increase with the incorporation of MS to the routine of a clinical research laboratory.

METAGENOMICS ANALYSIS OF OIL CONTAMINATED SOILS BY 16S GENE SEQUENCING REVEALED DEEP CHANGES IN MICROBIAL COMMUNITY STRUCTURE

Di Tomaso K[1,2], Paytuví-Gallart A[3], Bonomo MG[1], Calabrone L[1], Bufo SA[1], Aiese Cigliano R[3], Salzano G[1]

1. Department of Sciences
2. Ph.D. school in Applied Biology and Environmental Safeguard, Università degli Studi della Basilicata, Potenza, Italy.
3. Sequentia Biotech SL, Carrer Comte d'Urgell 240, 08036 Barcelona, Spain.
Presenter e-mail: katiaditomaso@gmail.com

Relations between soil biodiversity and ecosystem functions depend on structural and functional diversity of species. Microorganisms represent the major part of soil community concerning total biomass, nutrient transformation and degradation of toxic compounds. Many bacterial species are non-cultivable out of their natural habitat and can be studied within their environment with metagenomics analysis, sequencing DNA from all species.
In this work, microbial community structure after an oil pipeline installation was studied by 16S gene sequencing. DNA was extracted from agricultural and forest soils, collected at 0-20 cm and 20-40 cm depth in 2013, 2014 (year of the installation), 2015 and 2016. A library of 16S amplicon was prepared and sequenced with MiSeq, producing 300 bp paired-end reads. On average 266,168 reads were retained after processing with Trimmomatic-0.33.
Metagenomics analysis was performed using GAIA, an innovative pipeline able to map processed reads against the NCBI database. After mapping, GAIA uses a Lowest Common Ancestor (LCA) algorithm to bin reads into OTUs with high accuracy and then estimates alpha and beta diversities. Based on different similarity thresholds, reads are binned into the taxonomic levels domain, phylum, family, genus and species if their identity is equal or higher to 70%, 73%, 85%, 93% and 97%, respectively. GAIA identified a total of 56 phyla, 485 families, 1,190 genus and 23,232 species. Alpha diversity estimation revealed a decrease of richness in all samples in 2014 and beta diversity was higher in these soils. The most abundant OTUs were *Proteobacteria*, *Actinobacteria*, *Planctomycetes*, *Acidobacteria*, *Firmicutes*, *Chloroflexi*, *Verrucomicrobia* and *Bacteroidetes* in all samples. *Proteobacteria* and *Bacteroidetes* showed a higher abundance, while the other phyla decreased both in agricultural and forest soils in 2014. Results showed a different biodiversity in soils, with a great perturbation in 2014, suggesting a bioremediation process started to counteract the effect of oil pipeline installation.

MUTATIONAL SIGNATURES IN CANCER (MuSiC): A WEB APPLICATION TO IMPLEMENT MUTATIONAL SIGNATURES FRAMEWORK IN CANCER SAMPLES

Marcos Díaz-Gay[1], Maria Vila-Casadesús[2], Sebastià Franch-Expósito[1], Juan José Lozano[2], Sergi Castellví-Bel[1]

1. Gastroenterology Department, Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Rosselló 149, 08036 Barcelona, Catalonia, Spain.
2. Bioinformatics Platform, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Rosselló 149, 08036 Barcelona, Catalonia, Spain.
Presenter e-mail: diaz2@clinic.cat

Mutational processes in somatic cells are led by endogenous or exogenous mutagenic agents, as well as errors in DNA replication or repair machineries. Any type of agent or defect is responsible for a specific footprint in the form of a different burden and pattern of mutations.
To identify these profiles, mutational signatures have been proved as a valuable pattern in somatic genomics mainly regarding cancer, with a potential application as a biomarker in clinical practice. Up to now, several bioinformatic packages to address this topic have been developed in different languages/platforms (mostly in R). However, no web application is available to extract the underlying mutational signatures for single samples by comparing with the signatures currently reported in the COSMIC database.
In this work we present Mutational Signatures in Cancer (MuSiC), a new web application based on the Shiny framework and written in R language. By means of a user-friendly interface, it permits the visualization of the somatic mutational profile and the contribution of the reported mutational signatures in the analyzed samples.
MuSiC web application is accessible at http://bioinfo.ciberehd.org/GPtoCRC/en/tools.html and source code is freely available at https://github.com/marcos-diazg/music.

PREDICTION OF ALTERNATIVE SPLICING EVENTS AND SQTL IN NORMAL COLON TISSUE

Virginia Díez-Obrero, Ferrán Moratalla, Víctor Moreno

Unit of Biomarkers and Susceptibility, Catalan Institute of Oncology, IDIBELL and CIBERESP, Hospitalet de Llobregat, Catalonia, Spain.
Presenter e-mail: vdiezo@idibell.cat

Alternative splicing (AS) is a highly tissue-specific process by which multiple mRNA isoforms are produced, which could be markedly different while originating from the same locus. The AS events are the outcome of the different mechanisms by which genes can be alternatively spliced. Splicing QTL (sQTL) studies contribute to understand the mechanisms responsible for AS regulation. They have been conducted in the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2015). In this study, we aim to explore the AS landscape in normal colon and to identify sequence variants that contribute to the regulation of gene expression at the isoform level in this tissue. Specifically, we assess cis sQTL associated with changes in the relative abundances of the different isoforms. RNA-Seq and DNA genotyping data belonging to 178 individuals were used. The main steps and methods of the study were RNA-Seq read mapping with STAR, isoform quantification with RSEM and AS events and sQTL prediction with sQTLseekeR (Monlong et al., 2014). We used RefSeq mRNA annotation. From the starting 494,012 variants and 41,970 transcripts, we predicted 675 sQTL that affect 220 genes (called sGenes) with a 5% FDR threshold. Most of them were intronic. Besides, the most common AS event is exon skipping, followed by complex events involving the 5' and 3' ends of the isoforms. As a conclusion, the AS events associated with normal colon tissue isoforms have been defined, and it has been produced a catalog of sQTL for this tissue. These results are a useful resource for future studies regarding regulation of gene expression for both healthy and diseased tissue.

HUMAN EVOLUTION INFERENCE BY DEEP LEARNING IMPLEMENTATION

Olga Dolgova, Iago Maceda, Oscar Lao

Population Genomics Team, Centre Nacional d'Anàlisi Genòmica, Centre de Regulació Genòmica (CRG-CNAG), Parc Científic de Barcelona, Baldiri Reixac 4, 08028 Barcelona, Catalonia, Spain.
Presenter e-mail: olga.dolgova@cnag.crg.eu

Detecting archaic introgression using current genetic variation of *Homo sapiens* is an extremely active field within the community of human population genomics. Nevertheless, quantifying the amount of admixture and the relationship of (sometimes) unknown ancestral populations is complex. Proper identification of its signature is essential for interpreting the different evolutionary processes, demographic as well as selective, that shaped the genome of the human species. Several algorithms have been proposed for unraveling these evolutionary events. However, a number of limitations were suggested to the proposed methods, both in biological and technical terms. First of all, the algorithms do not model the tree topology of the ancestral populations. As a consequence, several demographic scenarios can produce the same output complicating the interpretation of the results. This situation is even more complex when considering both ancient and modern samples at the same time. The algorithms do not correct for the temporal difference among the samples, thus producing a systematic bias on the estimated proportions of ancestry in the ancient sample. Moreover, the sensitivity for estimating the ancestry proportions and divergence times considerably differ among various algorithms. Furthermore, so far proposed algorithms do not consider low frequency alleles, despite a considerable proportion these variants represent in the genetic variation of the species and taking into account that they can constitute a key information for detecting population substructure.
In the present ongoing project we have been developing a novel approach based on coupling of Deep Learning with Approximate Bayesian Computation for alleviating these reported problems with the aim to detect the demographic processes along human history and to infer the more reliable *Hominin* phylogeny, including multiple introgression and admixture events among ancient and even unknown archaic populations, using 10 Native American population as a target dataset for their ancestry inference.

RECONSTRUCTING A COMPLEX SNAKE VENOM PROTEIN GENE LOCUS WITH NANOPORE SEQUENCER

Vincent L. Viala[1], Líbia Sanz[2], Jordi Durban[2], Alícia Pérez[2], Diego Dantas Almeida[1], Pollyanna Campos[1], Ursula C. Oliveira[1], Milton Nishiyama-Junior[1], Juan J. Calvete[2], Inácio Junqueira-De-Azevedo[1]

1. Laboratório Especial de Toxinologia Aplicada, Unidade de Genômica da Biodiversidade CeTICS, Instituto Butantan, São Paulo, Brasil.
2. Lab. Venòmica Estructural y Funcional, Institut de Biomedicina de València, CSIC.
Presenter e-mail: jdurban@ibv.csic.es

In the absence of a **snake genome**, the structural organization of some genes expressed in the venom gland is unknown. In the present work, we aimed at solving one of these locus by sequencing BAC clones of genomic DNA using on the 3rd generation sequencing platform.
Seven clones from a snake BAC genomic library were identified as Snake Venom Metalloproteinases toxin proteins. The BAC DNA was extracted separately and samples were fragmented following the procedure. The sequencing libraries were prepared with Native barcoding kit EXP-NBD103 and Ligation sequencing kit SQK-LSK108 1D. Libraries were run on two flow cells FloMin106/R9 on a **MinION®** device (Oxford Nanopore Technologies), for up to 48 hours. Data analysis workflow used was: Albacore basecalling (+ FastaQ extraction and barcode separation), Porechop trimming, Canu *de novo* assembly and Pilon correction.
The data allowed us to assemble each BAC in a **single contig**, totalizing 7 contigs from 158.940 to 221.527 bp length. The sequencing depth was sufficient to correct the pBeloBAC vector sequence up to **99.9% accuracy**, ranging from 112 to 318X. The size of the longest read that aligned to each one of the contigs ranged from 66.073 bp to 98.951 bp. We identified 22 full genes and exon-containing residual gene elements indicating that recombination can be one mechanism of duplication and thus a source of genetic novelties. Some BACs showed high identity (>99%) within a long 125 kb terminal segment, indicating superposition and probably the same genomic origin, which could enable the assembly of a larger scaffold (~200 kb). Next steps will be to integrate genomic and venom gland transcriptomic data from Illumina® sequencing to correct contigs, to precisely annotate genes and to analyze gene expression levels. The comparison of gene sequences within his locus will allow the recognition of duplication events and a more solid reconstruction of the evolutionary history of these genes.

A SEMI-SUPERVISED SYSTEM FOR ONTOLOGY ENRICHMENT

Elizabeth T. Hobbs[1], Matthew Koert[1], Patrick K. O'Neill[2], Ivan Erill[1]

1. Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD (USA).
2. Tyche Analytics Company Inc., San Jose, CA (USA).
Presenter e-mail: erill@umbc.edu

Methods to analyze high-throughput datasets require curated knowledge-bases to perform inference and extract knowledge. These resources are currently compiled through slow-paced manual curation and addressing this bottleneck is of fundamental importance for biomedical research. This work focuses on the development of a semi-supervised system for the semantic annotation of ontological terms in unlabeled text, and the supervised expansion of a cross-domain ontology to incorporate identified textual citations as a resource for further development of text-mining systems across the biomedical landscape. Using the Gene Ontology and the Evidence Ontology as a reference framework for distant learning we put forward a semi-supervised system to mine biomedical articles. The semi-supervised approach is based on a Bayesian author-topic generative model operating on an expectation-maximization (EM) framework. We report preliminary results showing that this approach is feasible and can take into account the preferences of authors for specific topics, leading to adequate refining of word mappings for GO and ECO terms as a first step towards reliable mapping of ontological terms on unstructured text.

IDENTIFYING CHARACTERISTIC NETWORK SIGNATURES IN CANCER CELL LINES

Adrià Fernandez[1], Miquel Duran-Frigola[1], Patrick Aloy[1,2]

1. Join 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, Spain.
Presenter e-mail: adria.fernandez02@estudiant.upf.edu

During the last years, network-based approaches have emerged as powerful tools to shed light on drug activity, as well as for understanding biological relationships. These approaches are particularly interesting when dealing with complex diseases. In cancer, the mutational heterogeneity complicates the discovery of associations between individual mutations and a clinical phenotype. However, the amount of available experimental data for different cancer cell lines gives us the opportunity to explore the biological mechanisms that are behind this complex disease. Here we propose a methodology that takes advantage of the available gene expression data to characterize a cancer cell line at molecular level. For a given cell line, we are able to identify not only the genes that are differentially expressed but also which of those are close to each other in a biological network. The obtained genes form characteristic subnetworks that can be easily interpretable by identifying Gene Onthology terms associated to them. Furthermore, we demonstrate that these subnetworks show clinical relevance and can be used to predict drug response with a performance similar to the one obtained by using the raw gene expression. Overall, from hundreds of differentially expressed genes found in a single cell line, our methodology is able to delimit those that are biologically associated among them.

THE EUROPEAN RABBIT GUT MICROBIOME: INSIGHTS INTO THE INTESTINAL HEALTH

Gerard Funosas-Planas[1], Francisca Castro[2], Rafael Villafuerte[2], Emilio O. Casamayor[1], Xavier Triadó-Margarit[1]

1. Integrative Freshwater Ecology Group, Centre of Advanced Studies of Blanes-Spanish Council for Research (CEAB-CSIC), Carrer d'accés a la Cala St. Francesc 14, 17300 Blanes, Catalonia, Spain.
2. Institute of Advanced Social Studies (IESA-CSIC), Campo Santo de Los Martires 7, 14004 Córdoba, Andalusia, Spain.
Presenter e-mail: gfunosas@ceab.csic.es

The gut microbiome is a current hot topic of research because of the close relationships with physiological and immunological responses in animals, proving to be a key factor in the host health status. We analysed the gut microbiome of the wild European rabbit *Oryctolagus cuniculus*, a keystone species for the Iberian biodiversity. For comparisional purposes, two subspecies were studied: *algirus* and *cunniculus*, in both captive and wild conditions. We hypothesized that the proportion of different microbial groups can be used as a proxy for host intestinal health status.
None of the studied factors (diet, captivity status, subspecies, sex, age) neither an environmental gradient within the wild rabbits directly shaped the gut microbiome composition. The highest diversity was found in a group of translocated selected and apparently healthy wild rabbits used for restocking purposes, whereas in other animals poor communities with a consistent dominance by *Escherichia-Shigella* or *Enterobacter* were observed. A significant negative correlation between relative abundances of Ruminococcaceae-Lachnospiraceae-Rikenellaceae and Enterobacteriaceae was found. Our study suggests that the composition of the rabbit gut microbiome may change accordingly to the intestinal health status of the animals, unveiling the ratio Ruminococcaceae/Enterobacteriaceae as a key parameter to determine healthy individuals. Further, the analyses of predicted functional profiles and a meta-analysis approach reinforced this statement.

TRANSCRIPTOME ANALYSIS OF EN-TEX DATA USING PERSONALIZED GENOMES

Anna Vlasova[1], Alexander Dobin[2], Beatrice Borsari[1], Dinar Yunusov[2], Alessandra Breschi[1], Julien Lagarde[1], Carrie Davis[2], Fritz J. Sedlazeck[3], EN-TEx Consortium, Michael C. Schatz[2], Thomas R. Gingeras[2], Roderic Guigó[1]

1. Center for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Spain
2. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.
3. Johns Hopkins University, Baltimore, MD, USA.
Presenter e-mail: roderic.guigo@crg.eu

The genome of every individual carries thousands of short and long variants when compared to the reference human genome used for the functional analysis. Some of these variants may affects the expression and regulation of nearby genes, but our understanding of impact of these alterations is limited by using a reference genome and reference annotation without taking into account personal differences. A collaborative project between ENCODE and GTEx, the EN-TEx, aims to evaluate the impact of using personalized diploids genomes for functional analysis, in particular for analysis of gene and transcript expression across multiple conditions.
Within the scope of this project the genomes of 4 human samples were sequenced using a combination of technologies: Illumina short reads, PacBio long reads, and 10x Genomics Chromium linked reads. For each individual, a personal diploid genome and annotation was reconstructed. By analysing the resulting data, we have identified structural variations that affect genes and regulatory regions, as well as transcripts present in one allele only. At the same time, transcriptomic data was mapped to the reference genome GRCh38 and quantified with Gencode annotation v.24. We then compared expression values obtained with these two approaches and found genes changing their expression at least twice. These included the highly polymorphic HLA-DQ and HLA-DR genes. Overall, we have observed that using personal diploid genomes can lead to more accurate RNA-seq mappings and quantification.

ALLELE-SPECIFIC EXPRESSION ANALYSIS REVEALS A HIGH STABILITY OF GENE EXPRESSION PROFILES IN HYBRID YEASTS

H. Hovhannisyan[1,2], T. Gabaldón[1,2,3]

1. Bioinformatics and Genomics Programme, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain.
2. Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain.
3. ICREA, Pg. Lluís Companys 23, Barcelona, Spain.
Presenter e-mail: grant.hovhannisyan@crg.eu

Interspecific hybridization is one of the key evolutionary mechanisms of adaptation. Recently, numerous studies have demonstrated that many yeast species, representing a wide range fungal clades, are interspecific hybrids. However, the mechanisms of interplay between parental genomes of hybrid yeasts, underlying their adaptive phenotypes, are poorly investigated. To address this issue on gene expression level we generated and analysed RNAseq data of *S. cerevisiae x S. uvarum* artificial hybrid, its parentals and data of two strains of opportunistic human pathogen *Candida orthopsilosis* – *C. orthopsilosis MCO456* and *C. orthopsilosis 90-125,* with the latter being a likely parent of *MCO456* hybrid strain. Differential expression (DE) analysis between parents and homeologs in artificial hybrid has shown a little proportion of DE genes – 13 in *S. cerevisiae* and 17 in *S. uvarum* homeolog and their corresponding parentals ($|L2FC|>2$, $p<0.01$), demonstrating a high conservation of parental gene expression patterns after hybridisation. Allele-specific expression (ASE) analysis revealed 110 imbalanced homeologous genes, involved in different GO functional categories, e.g. oxidoreductase activity, transmembrane transporters activity, etc. To perform a similar analysis for *C. orthopsilosis MCO456*, we have phased 115 genes based on heterozygous mutations compared to parental strain. Parent-homeolog comparison revealed three DE genes, while ASE analysis showed imbalance for one gene being up-regulated in *90-125* homeolog.
Summarizing, our results show that despite genomic shock resulted from hybridization event, hybrid yeasts demonstrate relatively high stability of gene expression profiles of their parental genomes. Further analysis has to be carried out to reveal the molecular mechanisms of this stability.

EVALUATION OF MUTATIONAL IMPACT ON PRION-LIKE PROTEINS AGGREGATION PROPENSITY

Valentin Iglesias[1]*, Oscar Conchillo-Sole[1]†, Cristina Batlle[1], Xavier Daura[1,2] and Salvador Ventura[1]

1. Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain.
Presenter e-mail: valentin.iglesias@uab.cat
† These authors contributed equally.

Proteins bearing prion-like domains (PrLD) are widespread throughout all kingdoms of life. These domains resemble the intrinsically disordered, low complexity, Q/N-rich regions present in most yeast prions. Multiple studies have predicted around 1% of the human proteome corresponds to these prion-like proteins. Characterization of this human protein subset has stated its enrichment in DNA and RNA binding proteins and their involvement in the formation of biomolecular condensates. These transient membraneless compartments phase separate through highly dynamic liquid-liquid demixing and are related to several neurodegenerative diseases. This is particularly evident in cases of naturally occurring mutations that increase the aggregation propensity of PrLDs by converting these liquid compartments into solid aggregates, compromising their dynamic nature. Hence there is a need for in silico tools able to quantify the impact of mutations on the aggregation propensities of this kind of disease-associated proteins.
The debate of whether the self-assembling propensities of prion-like proteins depend only on a biased amino acid composition accounting for the whole PrLDs, or instead on specific sequential features facilitating their transition to amyloid-like states is an unsolved hot topic in our field. Nonetheless, we have recently shown that a function that takes into consideration both parameters predicts better the impact of a wide spectrum of punctual and multiple mutations or deletions on the aggregation of the model ALS-associated prion-like hnRNPA2 protein. Accordingly, we introduce the AMYCO (combined AMYloid and COmposition based prediction of prion-like aggregation propensity) webserver which implements this approach to allow fast and automated predictions.

A REFINED GENOME ATLAS OF MAMMALIAN LONG NON-CODING RNAs

Julien Lagarde[1,2], Barbara Uszczynska-Ratajczak[1,2,6], Silvia Carbonell[3], Silvia Pérez-Lluch[1,2], Amaya Abad[1,2], Carrie Davis[4] , Thomas R Gingeras[4] , Adam Frankish[5] , Jennifer Harrow[5,6] , Roderic Guigó[1,2] & Rory Johnson[1,2,6]

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.
[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[3] R&D Department, Quantitative Genomic Medicine Laboratories (qGenomics), Barcelona, Spain.
[4] Functional Genomics Group, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA.
[5] Wellcome Trust Sanger Institute, Hinxton, UK.
[6] Present addresses: Centre of New Technologies, Warsaw, Poland (B.U.-R.); Illumina, Cambridge, UK (J.H.); Department of Clinical Research, University of Bern, Bern, Switzerland (R.J.).
Presenter e-mail: julien.lagarde@crg.eu

Accurate annotation of genes and their transcripts is a foundation of genomics, but currently no annotation technique combines throughput and accuracy. As a result, reference gene collections remain incomplete—many gene models are fragmentary, and thousands more remain uncataloged, particularly for long noncoding RNAs (lncRNAs). To accelerate lncRNA annotation, the GENCODE consortium has developed RNA Capture Long Seq (CLS), which combines targeted RNA capture with third-generation long-read sequencing. Here we present an experimental reannotation of the GENCODE intergenic lncRNA populations in matched human and mouse tissues that resulted in novel transcript models for 3,574 and 561 gene loci, respectively. CLS approximately doubled the annotated complexity of targeted loci, outperforming existing short-read techniques. Full-length transcript models produced by CLS enabled us to definitively characterize the genomic features of lncRNAs, including promoter and gene structure, and protein-coding potential. Thus, CLS removes a long-standing bottleneck in transcriptome annotation and generates manual-quality full-length transcript models at high-throughput scales. (Lagarde *et al., Nat. Genet*. 2017)

COMPUTER-AIDED DRUG DESIGN APPLIED TO MARINE DRUG DISCOVERY: MERIDIANINS AS ALZHEIMER'S DISEASE THERAPEUTIC AGENTS

Laura Llorach-Pares[1,2], Alfons Nonell-Canals[1], Melchor Sanchez-Martinez[1] and Conxita Avila[2]

1. Mind the Byte S.L., Barcelona, Catalonia.
2. Dept. of Evolutionary Biology, Ecology and Environmental Sciences, Faculty of Biology and Biodiversity Research Institute (IRBio), Universitat de Barcelona, Catalonia.
Presenter e-mail: laura@mindthebyte.com

Drug discovery is the process of identifying new molecules with a certain therapeutic activity. This process is very expensive in terms of money and time, but in the last decades, the usage of computer-aided drug design (CADD) techniques at various stages of the drug discovery pipeline could effectively reduce that cost. CADD techniques can be applied to several steps and tasks of the discovery process such as biological profile prediction. For instance, virtual profiling (VP) methods, can predict mechanisms of action for a certain molecule as well as their targets; molecular modelling techniques can predict molecule stability and/or ligand-target interactions; virtual screening (VS) methods are able to find analogs (similar molecules) and/or build compounds libraries; hit to lead (H2L) optimization techniques are used to design new molecules and both, potency and ADME/Tox properties, can be tested before synthetizing it.
Historically, natural products have been a rich source of compounds for drug discovery. On 2013, an estimation of all FDA-approved new molecular entities (NME's) revealed that natural products and their derivatives represent over one-third of them. In this line, the use of marine molecules, could be a great strategy to discover novel structures since the oceans harbor most of the biodiversity of the world.
To exemplify and highlight the power of CADD techniques in marine drug discovery, as part of an ongoing study of bioactive marine molecules from benthic invertebrates, we present here a case study of CADD methods applied to Meridianins A-G, a group of marine indole alkaloids consisting of an indole scaffold connected to a pyrimidine ring, isolated from specimens of the tunicate genus *Aplidium*. We show how starting from the 2D chemical structure of a Meridianin we have been able to predict biological indications, as well as some targets, more precisely, various protein kinases involved in Alzheimer disease (AD).

MACHINE-LEARNING QSAR MODEL FOR PREDICTING ACTIVITY AGAINST MALARIA PARASITE'S ION PUMP PFATP4 AND IN SILICO BINDING ASSAY VALIDATION

Angela Lopez-del Rio[1,2], Laura Llorach-Pares[1,3], Alexandre Perera-Lluna[2,4], Conxita Avila[3,5], Alfons Nonell-Canals[1], Melchor Sanchez-Martinez[1]

1. Mind the Byte, S.L., Barcelona Science Park, Baldiri Reixac, 4-8, 08028 Barcelona, Catalonia, Spain.
2. Department of ESAII, Center for Biomedical Engineering Research, Universitat Politècnica de Catalunya, Pau Gargallo 14, 08028 Barcelona, Catalonia, Spain.
3. Department of Evolutionary Biology, Ecology and Environmental Sciences, Faculty of Biology, Universitat de Barcelona, Diagonal 643, 08028 Barcelona, Catalonia, Spain.
4. Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Monforte de Lemos 3-5, 28029 Madrid, Spain.
5 Biodiversity Research Institute (IRBio), Faculty of Biology, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Catalonia, Spain.
Presenter e-mail: angela@mindthebyte.com

Malaria is a mosquito-borne infectious disease caused by parasitic protozoans of the genus Plasmodium. Although different effective antimalarial medicines have been developed, there is serious concern that parasites are developing widespread resistance to these drugs. To avoid this, now the efforts are concentrated on treating the malaria inside the Anopheles mosquito.
Several academic groups and companies are working worldwide to develop new compounds that fight the disease without provoking drug resistance. Within this context and making use of the Open Source Malaria project (http://opensourcemalaria.org/) which provides a large collection of molecules and 3D models of the targets with which they interact, we developed a machine learning-based QSAR model that predicts which molecules will block the malaria parasite's ion pump, PfATP4. The model was then employed to screen and classify the DrugBank database molecules and compounds coming from a proprietary marine molecules library. Finally, by means of an in silico binding assay, the predicted behavior was validated in the positive cases.
Summarizing, we have created a new set of repositioned drugs and marine molecules against malaria, establishing a good starting point for further studies and highlighting the key role that computational methods can have in the rational design of new drugs against infectious diseases.

ILLUMINATING POTENTIALLY FUNCTIONALIZED ALU REPEATS IN GREAT APES

Izaskun Mallona, Berta Martín, Mireia Jordà, Miguel A. Peinado

Germans Trias i Pujol Health Science Research Institute (IGTP), PMPPC, Badalona 08916, Catalonia, Spain
Presenter e-mail: imallona@igtp.cat

About half of the human genome is composed by transposable elements. Although multiple hypotheses point out roles in genome structure and function, their contribution to genome regulation is still poorly understood. Alu repeats are restricted to the primate lineage and constitute the most abundant transposon in human. By application of the NSUMA (Next-generation Sequencing of UnMethylated Alu) technique we have shown that unmethylated Alu repeats display epigenetic features consistent with regulatory potential in normal and cancer cells (Mallona *et al*, J Biomed Inform. 60:77-83, 2016; Jordà *et al*. Genome Res 27:118-132, 2017).
Given the high phenotypic diversity across primates, we have explored the potential regulatory roles of unmethylated Alu elements among them. We applied the NSUMA technique to scrutinize more than 130,000 individual Alu elements in the whole blood from four species of great apes: gorilla, chimpanzee, orangutan and human.
We estimate the rates of Alu unmethylation across species and characterize the genomic and epigenomic features of the unmethylated Alus, including sequence conservation, selective pressure, proximity to coding genes, transcription factor binding sites and chromatin modifications.
This work was supported by grants from MINECO and FEDER (SAF2015-64521-R). CERCA Programme/Generalitat de Catalunya.

COMPREHENSIVE IDENTIFICATION OF DRIVER GENES AND DRIVER MUTATIONS ACROSS TUMORS WITH IntOGen2017

Francisco Martínez-Jiménez[1]*, Loris Mularoni[1,2]*, Ferran Muiños[1]*, Carlota Rubio-Perez[1,2], Jordi Deu-Pons[1,2], Inés Sentís[1,2], Iker Reyes-Salazar[1,2], David Tamborero[1,2], Abel Gonzalez-Perez[1,2], Núria López-Bigas [1,2,3]

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.
* These authors contributed equally to the work.
Presenter e-mail: francisco.martinez@irbbarcelona.org

The ever-increasing availability of sequenced cancer genomes has opened the possibility to comprehensively analyze the genomic events leading to tumorigenesis. Identifying those events is of paramount importance to understand tumor biology and to provide better cancer treatments. Finding signals of positive selection in the pattern of tumor mutations has proven to be an effective way to identify cancer genes. Methods to fit this purpose continue to be developed, each having limitations intrinsic to the specific signal it aims to identify. Herein, we present IntOGen2017 (https://www.intogen.org/), which includes a pipeline implementing an array of methods that exploits complementary signals of positive selection, a new combination strategy to reliably identify cancer genes and a novel approach to identify individual driver mutations. Using IntOGen2017 we have analyzed mutations from more than 25,000 tumor samples, including whole exomes, whole genomes and gene panels; creating a catalog of cancer genes and driver mutations across tumors. To our knowledge this is the largest analysis of genomic driver events to date, providing a comprehensive characterization of more than 120 cohorts of over 60 cancer types. IntOGen2017 fuels cancer research by helping the identification of drivers across tumor cohorts which ultimately aids the selection of the most suitable cancer treatment.

GEMBS - FAST AND EFFICIENT WGBS DATA PROCESSING

Angelika Merkel[1,2], Marcos Fernandez Callejo[1,2], Eloi Casals[1,2], Santiago Marco-Sola[3], Ronald Schuyler[1,2], Ivo Gut[1,2] and Simon Heath[1,2]

[1] Centro Nacional de Análisis Genómico (CNAG-CRG), Centre de Regulacio Genómico (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain
[2] Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain.
[3] Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain.
Presenter e-mail: angelika.merkel@cnag.crg.es

DNA methylation is essential for normal development and cell differentiation in mammals. Acting in concert with other epigenetic marks to alter chromatin conformation, it has been implicated in the regulation of gene expression and a multitude of biological processes such as genomic imprinting, silencing of transposable elements, and disease, particularly cancer. Whole genome bisulfite sequencing (WGBS) is the gold standard for studying genome-wide DNA methylation at base pair resolution. However, processing of NGS data from bisulfite converted DNA requires specific, yet efficient, processing to accommodate for the cytosine to thymine conversion that allows the distinction of methylated from unmethylated cytosines (methylated cytosines resist the conversion). We present GEMBS, a pipeline specifically designed for the high-throughput analysis of WGBS data, which has already been successfully applied to several small and large scale projects including BLUEPRINT, the European epigenome project (http://www.blueprint-epigenome.eu).
GEMBS is implemented with the 'JIP Pipeline system' (http://pyjip.readthedocs.io/en/latest/) and comprises two core modules: a BS-adapted version of 'gem mapper' (Marco-Sola et al., 2012) and the 'BS_call' genotype-methylation caller. QC matrices and mapping statistics are conveniently presented as html reports, and processing can be highly automated in connection with a Laboratory Information Management system. Benchmarking demonstrates high efficiency and speed of GEMBS over other commonly used pipelines, such that a standard 30X WGBS data set can be processed 5-8h on a computing cluster. We further demonstrate that given sufficient coverage GEMBS can be used for SNP calling from WGBS.

PREDICTING THE IMPACT OF MUTATIONS ON TRANSCRIPTION FACTOR BINDING

Alberto Meseguer[1], Oriol Fornes[2], Juan I. Fuxman Bass[3] and Baldo Oliva[1]

1. Structural Bioinformatics Lab (GRIB-IMIM), Department of Experimental and Health Science, University Pompeu Fabra, Barcelona 08005, Catalonia, Spain.
2. Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, BC V5Z 4H4, Canada.
3. Department of Biology, Boston University, Boston 02215, Massachusetts, United States of America.
Presenter e-mail: alberto.meseguer@upf.edu

Interactions between transcription factors (TFs) and their DNA binding sites (TFBSs) play a central role in gene regulation. Mutations in both TF genes or their associated TFBSs can lead to altered (increased or decreased) TF binding, dysregulation of genes, and ultimately impact phenotype[1][2]. For example, mutations in sonic hedgehog enhancers have been associated to limb malformation[2][3]. Here, we present a computational approach to predict the impact of mutations within TFBSs on TF binding. Briefly, for a TF of interest, we first model its interactions with DNA at heterozygous site binding events using a combination of comparative modelling (for modelling the TF) and threading (for modelling the TFBS). Each model is then scored using our previously described split-statistical potentials[4] that we adapted for TF-DNA interactions[5]. The resulting scores are finally used to train a machine learning classifier to predict the impact of mutations within the binding sites of the TF on the ability of the TF to recognize and bind to these mutated TFBSs. We benchmarked our approach on a dataset of altered TF binding events from yeast one-hybrid experiments[2]. These altered TF binding events comprised mutations resulting in both gain and loss of TF binding. Our method achieved similar precision values >0.7 for both 1) loss of TF binding, for members of the ETS, nuclear receptor and T-box families, and 2) gain of TF binding, for members of the bHLH, ETS, homeodomain and SMAD families. As a case study, we were able to predict the impact of mutations linked to developmental malformations and hormonal diseases with a success ratio >0.8. Our results demonstrate the potential use of TF-DNA interactions at the molecular-level for elucidating the functional role of regulatory variants across the human genome.

GERMLINE CNV DETECTION IN GENETIC DIAGNOSTICS: BENCHMARK OF ALGORITHMS FOR CNV CALLING FROM NGS DATA AT SINGLE EXON RESOLUTION

José Marcos Moreno-Cabrera[1, 2, 3], Eduard Serra[1,3], Conxi Lázaro[2,3], 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.
Presenter e-mail: jmoreno@igtp.cat

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 has been multiplex ligation-dependent probe amplification (MLPA). However, in a multi-gene testing scheme facilitated by the NGS capability, its use can be compromised, due to cost and time.
Most CNV calling algorithms perform the best when calling large CNVs (in the order of megabases) but are not able to reliably detect small CNVs affecting only one or few exons. In addition, most of them are designed to work with whole exome data and have problems with the more sparse data generated by NGS panels, currently widely used in routine genetic diagnostics.
New tools adapted to work in this context have been published recently, along with validated datasets. The aim of this work is to evaluate the performance of the existing tools to identify one suitable to be implement as part of the data analysis pipeline of our I2HCP genetic diagnostics strategy for hereditary cancer.
We have performed a benchmark of three CNV calling algorithms that have shown to perform well on NGS panel data at exon level: DECoN, CoNVaDING and panelcn.MOPS. These algorithms have been evaluated over two reference datasets of 96 and 170 samples with exon and multi-exon CNVs already analyzed. Each algorithm has been optimized programmatically under the same criteria, trying different parameters values to maximize sensitivity.

CHARACTERIZING THE PATHOGENIC LANDSCAPE OF COBALAMIN DEFECTS

Natàlia Padilla[1], Casandra Riera[1], Belén Pérez[2], Xavier de la Cruz[1,3]

1. Translational Bioinformatics, Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain.
[2] Centro de Diagnóstico de Enfermedades Moleculares, Centro de Biología Molecular, Universidad Autónoma de Madrid, CIBERER, Madrid, Spain.
[3] ICREA, Barcelona, Spain.
Presenter e-mail: natalia.padilla@vhir.org

Cobalamin defects are caused by nutritional deficiency or genetic disorders that affect either absorption and cellular uptake of the vitamin or the synthesis of active derivates. Such problems are associated with elevated methylmalonic acid (MMA), homocysteine (HCys) or both (MMAHC) in plasma and they have an estimated incidence of 1:37,000 births in Hispanic populations. In this study, we analysed 20 patients from *Centro de Diagnóstico de Enfermedades Moleculares* with a suspected cobalamin genetic disorder. A total of 26 different variants were identified in 12 genes, including 14 missense variants, 6 small deletions and 6 nonsense variants. The genotypes found were diverse including homozygosity, compound heterozygosity and suggestion of digenic inheritance. Here, we present an exhaustive analysis of these variants which includes the identification of functional relevant residues of the protein altered by the variants, the estimation of the functional impact of the variants by pathogenicity predictors, the evolutionary analysis of the region affected by the variants and the structural analysis of three dimensional models of the mutated proteins and their interaction partners when available. Our results provide the first global view of the damage caused by these variants: most of them either damaging the active site or the interaction surface of the protein.

GENOME LANDSCAPE OF CANCER MODULATORS

Luis Palomero, Alvaro Aytés, Miquel Angel Pujana

ProCURE, Catalan Institute of Oncology, Oncobell, IDIBELL, Hospital Duran i Reynals, Gran via 199, L'Hospitalet del Llobregat, Barcelona 08908, Catalonia.
Presenter e-mail: lpalomerol@gmail.com

Understanding of the genetic determinants (i.e., drivers) of cancer onset, progression, and therapeutic benefit has increased considerably in recent years. All too often, however, prediction of patient prognosis and therapeutic response is not accurate. The complex interplay of molecular interactions in cancer networks and the potential of reprogramming interactions in this context constitute key factors influencing the predictions. Here, by analyzing statistical significant modifications of patient survival and cancer relapse, we depict the modulators of all established cancer drivers across 16 cancer types. Critically, the number of modulators for a given cancer gene driver vary from few to hundreds. The modulators are divided in different functional categories that depend on cancer type and driver class. Mitochondrial activity emerges as a common linked modulator function across most cancer types. The relevance of the identified modulator sets is validated by integrating data from protein expression and synthetic lethal studies. Integration of known and predicted drug targets provides a framework to modulate the activity of known cancer drivers and, thus, potentially improve cancer therapy.

GAIA: INTEGRATED METAGENOMICS SUITE

Andreu Paytuví-Gallart, Ermanno Battista, Fabio Scippacercola, Riccardo Aiese Cigliano, Walter Sanseverino

Sequentia Biotech SL, Carrer Comte d'Urgell 240, 08036 Barcelona, Spain.
Presenter e-mail: ap@sequentiabiotech.com

Identifying the biological diversity of a microbial population is of fundamental importance due to its implications in industrial processes, environmental studies and clinical applications. Today, there is still a need to develop new, easy-to-use bioinformatics tools to analyze both shotgun and targeted metagenomics with the highest accuracy and the lowest running time. With the aim of overcoming this need, we introduce GAIA, an online Software as a Service (SaaS) solution that has been designed to provide the maximum information from whatever metagenomics sample: 16/18S, virome or shotgun analysis. GAIA is able to obtain a comprehensive and detailed overview at any taxonomic level of microbiomes of different origins: human (e.g. stomach or skin), agricultural and environmental (e.g. land, water or organic waste). Recent publications have benchmarked commonly-used 16/18S pipelines (Siegwald, *et al*. 2017) as well as shotgun metagenomics pipelines (McIntyre, *et al*. 2017), and we also benchmarked GAIA with the same datasets. GAIA is currently the best pipeline to analyze shotgun metagenomics data as it obtained the highest F-measures above all tested pipelines (CLARK, Kraken, LMAT, BlastMegan, DiamondMegan, NBC and OneCodex). In addition, GAIA also obtains excellent F-measures analyzing 16S data, yielding better F-measures than CLARK, kraken, BMP, mothur and QIIME. The overall objective of GAIA is to provide to academia and industries with an integrated metagenomics suite that will allow to perform metagenomics data analysis easily, quickly, and affordably with the best accuracy.

NEW INSIGHTS INTO THE ANALYSIS OF HI-C DATA

Andreu Paytuví-Gallart[1,2], Riccardo Aiese Cigliano[1], Walter Sanseverino[1], Covadonga Vara[2,3], Aurora Ruiz-Herrera[2,3]

1. Sequentia Biotech SL, Carrer Comte d'Urgell 240, 08036 Barcelona, Spain.
2. Departament de Biologia Cel·lular, Fisiologia i Immunologia, Universitat Autònoma de Barcelona, Campus UAB, Cerdanyola del Vallès, Spain.
3. Genome Integrity and Instability Group, Institut de Biotecnologia i Biomedicina (IBB), Universitat Autònoma de Barcelona, Campus UAB, Cerdanyola del Vallès, Spain.
Presenter e-mail: ap@sequentiabiotech.com

Hi-C is a genome-wide approach that allows to study the chromatin structure by identifying any contact between a pair of loci by coupling proximity-based ligation with deep sequencing (Erez Lieberman-Aiden, *et al*. 2009). Current bioinformatics approaches to analyze Hi-C data are resource-hungry besides of failing at the time to deliver the results within a reasonable timeframe and to perform visual or statistical analysis of two or more contact maps. Here we describe HiCloud, a bioinformatics pipeline integrated to a web framework that only needs the upload of the reads and few clicks to obtain: heatmaps, TADs, compartments, and differential interacting bins if more than two conditions are provided. In terms of speed, the pipeline has been benchmarked together with other pipelines, such as HiCUP (Wingett, *et al*. 2015) or TADbit (Serra, *et al*. 2017), using 10 CPUs and a subset of 5M Hi-C reads from real data generated in our lab coming from mouse cells. The time consumed by HiCloud to complete mapping, filtering and preparation of the results was ~14m. In contrast, HiCUP and TADbit required ~52m and ~10h, respectively. In terms of performance, HiCloud have 155.415 more valid reads than TADbit, 113.722 out of them (73.1%) due to higher mapping efficiency. The overall objective of HiCloud is to provide to academia and industries with an online integrated tool that will allow to perform Hi-C data analysis easily, quickly and affordably, without the need to have bioinformatics skills or powerful machines.

ASSESSING A BRCA1 VARIANT OF UNCERTAIN SIGNIFICANCE THROUGH SYSTEMS BIOLOGY: FUNCTIONAL, STRUCTURAL AND CONTEXTUAL ANALYSIS

Simón Perera[1], Laura Artigas[1], Teresa Sardón[1], Rafael Morales[2], José Manuel Mas[1]

1. Anaxomics Biotech, Barcelona (Spain), Balmes 89, 4º 2ª, 08008 Barcelona, Catalonia, Spain.
2. Genetic Counselling Unit, Medical Oncology Department, Hospital La Mancha Centro, Av. Constitución, 3, 13600 Alcázar de San Juan, Ciudad Real (Spain).
Presenter e-mail: simon.perera@anaxomics.com

Interpreting variants of uncertain significance (VUS) for their effect on protein function, and therefore for the risk of developing cancer, has become a challenge in clinical practice for genetic counselling services. Deleterious variants in the BRCA1 and BRCA2 genes account for approximately 20% of cases of hereditary breast and ovarian cancer. The BRCA1 gene plays a crucial role in DNA damage response and inactivating mutations lead to genetic instability, indirectly causing tumours as a result of an accumulation of mutations in cell cycle regulatory genes. In Spain, families that carry the deleterious mutation in the BRCA1 gene have a 52% cumulative risk of developing breast cancer and a 22% risk of developing ovarian cancer by the age of 70. The present work combines structural bioinformatics and Systems Biology-based mathematical modelling approaches with the aim of determining the pathogenicity of the mutation c.5434C>G (p.Pro1812Ala) in the BRCA1 gene (detected in a patient from a high risk family) and also to mechanistically understand the effect of this mutation in DNA damage response, a key process in cancer development. The results obtained showed that this mutation prevents the interaction of BRCA1 with key proteins of the cell cycle, subsequently impairing BRCA1-dependent induction of cell cycle arrest. The comparison of the molecular mechanisms associated with the native BRCA1 protein and the mutated variant function in DNA damage response showed that the latter undergoes a reduction in its ability to modulate pathways that are critical for DNA repair and cell cycle control. Therefore, this variant will not be able to exert its tumour suppressive action. Interestingly, these conclusions can be extrapolated to all mutations that, like c.5434C>G (p.Pro1812Ala) BRCA1, cause loss of BRCT domain activity.

A STATISTICAL MODEL FOR ADDRESSING TOPOLOGICAL BIAS IN NETWORK-BASED INFERENCE

Sergio Picart-Armada[1,2,3], Francesc Fernández-Albert[1,2], Maria Vinaixa[4,5,6], Oscar Yanes[4,5,6], Wesley K. Thompson[7], Alfonso Buil[8], Alexandre Perera[1,2,3]

1. Universitat Politècnica de Catalunya, Barcelona, Spain.
2. Centro de Investigación Biomédica en Red en el área temática de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Madrid, Spain.
3. Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Barcelona, Spain.
4. Centre for Omic Sciences, Reus, Spain.
5. Rovira i Virgili University, Tarragona, Spain.
6. Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas, Madrid, Spain.
7. University of California, San Diego, USA.
8. Mental Health Center Sct. Hans, Roskilde, Denmark.
Presenter e-mail: sergi.picart@upc.edu

The usage of biological networks has proven its usefulness in solving a variety of problems in bioinformatics, such as disease gene prioritisation, disease module finding, protein function prediction and pathway analysis. Examples of biological networks at different molecular levels include gene regulatory networks, protein-protein interaction networks or metabolic networks. The “guilt by association” principle is the main drive behind the usage of such networks.
Label propagation is a fundamental family of algorithms on networks. In short, labelled nodes are forced to propagate their labels through the edges of the network, in order to make inference on other nodes. The heat diffusion is an intuitive model for understanding the rationale of label propagation, but the direct interpretation of diffusion scores can lead to biased results. We present a statistical model for diffusion scores addressing topological bias and providing further insights on the behaviour of diffusion scores.
Our model has been applied to pathway analysis in metabolomics, being publicly available as an R package. The user provides a list of metabolites and obtains a relevant network consisting of not only the metabolites, but also reactions, enzymes, modules and pathways from the KEGG database. We have also released a general purpose R package that can deal with user-defined networks and labels. We show an example in yeast protein function prediction where normalisation outperforms unnormalised scores (FDR < 25%). Although we focus on the study of disease genes, the package is not limited to any specific omic science.

TRANSLATOME ANALYSIS OF GLUTAMATERGIC NEURONS IN A LEIGH SYNDROME MOUSE MODEL

Patricia Prada-Dacasa, Elisenda Sanz and Albert Quintana

Department of Cell Biology, Physiology and Immunology, Institut de Neurociències, Autonomous University of Barcelona, 08193-Bellaterra, Spain.
Presenter e-mail: patricia.prada@uab.cat

The most common pediatric presentation of mitochondrial disease is a fatal neuropathology known as Leigh Syndrome (LS). LS affects 1 out of 40 000 live births. Even though LS patients show a remarkable clinical heterogeneity, they are characterized by presenting severe breathing problems, motor issues, seizures and especially neurological damage mostly in basal ganglia and brainstem. Moreover, more than 75 disease genes have been described further complicating the understanding of the disorder. Currently, there is no cure and all treatments are palliative. [1]
Our group has characterized a mouse model of LS (Ndufs4KO) that lacks the NDUFS4 subunit of mitochondrial complex I, a gene frequently affected in LS. Complex I alterations lead to malfunction of mitochondrial oxidative phosphorylation system. Accordingly, the Ndufs4KO mouse develops a progressive encephalopathy that recapitulates the human pathology. [2] Previous results have identified that glutamatergic neurons of vestibular nuclei (VN) are more susceptible and that the deficiency of Ndufs4 in this cell type is sufficient to resemble the phenotype of Ndufs4KO. However, the molecular determinants of this susceptibility have not been characterized. Here, we present our approach to map the cell type-specific translatome of affected neuronal populations by combining the RiboTag technology and Next Generation Sequencing (NGS) with the overarching goal to elucidate the main pathways altered in LS. [3]

[1] Lake, N.J., Compton, A.G., Rahman, S., and Thorburn, D.R. "Leigh Syndrome: One disorder, more than 75 monogenic causes." Annals of Neurology 79.2 (2016): 190-203.

[2] Quintana, A., Kruse, S.E., Kapur, R.P., Sanz, E., and Palmiter, R.D. "Complex I deficiency due to loss of Ndufs4 in the brain results in progressive encephalophaty resembling Leigh Syndrome." PNAS 107.24 (2010): 10996-11001.

[3] Sanz, E., Yang, L., Su, T., Morris, D.R., McKnight, G.S, and Amieux, P.S. "Cell-type-specific isolation of ribosome-associated mRNA from complex tissues." PNAS 106.33 (2009): 13939-13944.

GENETIC LINKAGE ANALYSIS ON HERITABLE PULMONARY ARTERIAL HYPERTENSION (HPAH) TO DEAL WITH REDUCED PENETRANCE

Pau Puigdevall[1], Lucilla Piccari[2], Dan Geiger[3], Montserrat Milà[4], Celia Badenas[4], Irene Madrigal[4] and Robert Castelo[1]

1. Research Program on Biomedical Informatics, Functional genomics and Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain.
2. Institut d'Investigacions Biomomèdiques August Pi i Sunyer (IDIBAPS-FCRB), Rosselló 149-153, 08036 Barcelona, Spain.
3. Technion, Israel Institute of Technology, Computer Science Department, Taub 616, 36000 Haifa, Israel.
4. Centre de Diagnòstic Biomèdic, Bioquímica i Genètica Molecular, Hospital Clínic, Villarroel 170, 08036 Barcelona, Spain.
Presenter e-mail: pau.puigdevall@upf.edu

Large-scale genetic profiling and clinical sequencing are revealing an increasing number of carriers of disease-causing mutations who do not develop the disease phenotype. This characteristic is clinically reported as a genetic disorder of reduced or incomplete penetrance. Several mechanisms have been proposed to explain reduced penetrance, such as the molecular context of mutations, patient characteristics, such as age or sex, as well as specific environmental conditions that delay or trigger the disease onset (Cooper et al., 2013). The phenomenon of reduced penetrance constitutes a major challenge in the field of genetic diagnosis and counselling because phenotypes no longer unambiguously exhibit underlying genotypes. Nevertheless, its existence also provides new opportunities to learn how genotypes shape phenotypes. Here, we describe our efforts using linkage analysis to find a genetic modifier that explains the reduced penetrance in a particular genetic disorder: heritable pulmonary arterial hypertension. The results from linkage are further discussed regarding evidence on haplotype prediction, functional enrichment analysis as well as other functional genomics tools. These steps are required to narrow down the list of potential candidates to map the modifier and eventually to hypothesize about a particular genetic mechanism underlying reduced penetrance.

MANAGING THE ANALYSIS OF HIGH-THROUGHPUT SEQUENCING DATA

Javier Quilez[1,2*], Enrique Vidal[1,2], François Le Dily[1,2], François Serra[1,2,3], Yasmina Cuartero[1,2,3], Ralph Stadhouders[1,2], Thomas Graf[1,2], Marc A. Marti-Renom[1,2,3,4], Miguel Beato[1,2] and Guillaume Filion[1,2]

[1]Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain.
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[3]CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain.
[4]ICREA, Pg. Lluis Companys 23, 08010 Barcelona, Spain.
*Presenter e-mail: javier.quilez@crg.eu

In the last decade, there has been a great increase in sequencing throughput, decrease in the cost per base pair and rise in the number of sequenced-based applications. As a result, high-throughput sequencing (HTS) applications are pervasive in the life sciences, from small research groups to large-scale endeavors. The scientific community has put much effort to (i) develop new tools for the analysis of state-of-the-art HTS applications (e.g. single-cell and long-read sequencing) and (ii) improve existing algorithms (i.e. faster and with lower memory footprint).
However, we have paid less attention to the efficient management and analysis automation of the vast amount of HTS data generated. As a result, many bad practices are persistent in teams dealing with HTS data: (i) poor description of HTS experiments, (ii) unambiguous sample identification, (iii) anarchic data organization, (iv) non-scalable analysis pipelines, (v) scarce documentation of the computational procedures and (vi) heavy dependence on the data analyst to interpret the results. These practices have consequences for the cost, quality and reproducibility of research. We argue that these issues cannot be attributed to the need of better tools but to the human factor. We tend to plan in the short term, follow our own rules, change our mind and resist change.
As an insurance against fiasco, we advocate for embracing an attitude of documentation, automation, traceability and autonomy (DATA). Document HTS data by collecting metadata and detailing the computational procedures. Automate the analysis by writing code that is scalable, parallelizable, free of manual configuration and modular. Trace samples and the associated data using unique sample identifiers as well as a structured and hierarchical data organization. Empower the autonomy of researchers via the use of interactive web applications. In conclusion, scientific teams must develop a DATA culture to successfully managing and analysis of HTS data.

BIOCOR: A NEW TOOL FOR ASSESSING FUNCTIONAL SIMILARITIES BASED ON BIOLOGICAL CORRELATION

Lluís Revilla Sancho[1], Juan José Lozano[2], Pau Sancho-Bru[1]

[1] August Pi i Sunyer Biomedical Research Institute (IDIBAPS); Centre Esther Koplowitz, C/ Rosselló, 149-153, 08036. Barcelona, Catalonia, Spain.
[2] Enfermedades Hepaticas y Digestivas, Centro de Investigación Biomédica en Red (CIBEREHD), Instituto de Salud Carlos III, C/ Monforte de Lemos 3-5. 28029 Madrid, Spain.
Presenter e-mail: lrevilla@clinic.cat

Gene ontologies' similarity measures have been successfully applied to several purposes, including comparing functional similarity of two genes. However there is a lack of similarity measures based on pathways to calculate functional similarities. The aim of this study was to create a functional similarity measure based on pathways to assess biological similarities between genes. In addition, we evaluated how the similarity improved the biological relevance of the modules created with weighted gene correlation networks analysis (WGCNA). We implemented the measure in an R package (BioCor) using the Dice similarity coefficient between pathways. BioCor can also calculate similarities between clusters by combining the similarity of their pathways. Gene similarities are used to asses cluster similarities. BioCor can use pathways from Reactome, KEGG or any list of genes with pathway annotations from any database. Using BioCor with WGCNA increased the biological enrichment of modules created when using the similarity measures implemented in BioCor in pathway information from Reactome. BioCor is integrated with Bioconductor repository for further use in several different analyses like miRNA.

THE PAN-CANCER LANDSCAPE OF INTERACTIONS BETWEEN SOLID TUMORS AND INFILTRATING IMMUNE CELL POPULATIONS

David Tamborero*[1,2], Carlota Rubio-Perez*[1,2], Ferran Muiños[2], Sabarinathan Radhakrishnan[2], Aura Montasell[3], Rodrigo Dienstmann[4,5], Nuria Lopez-Bigas[1,2,6], Abel Gonzalez-Perez[2]

1 Department of Experimental and Health Sciences, University Pompeu Fabra, Barcelona, Spain.
2 Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, Spain.
3 Hospital del Mar Medical Research Institute, Barcelona, Spain.
4 Vall d'Hebron Institute of Oncology, Barcelona, Spain.
5 Sage Bionetworks, Seattle, USA.
6 Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.
Presenter e-mail: carlota.rubio@irbbarcelona.org

Tumors and immune cells are in a dynamic equilibrium that shapes the progression of cancer. While several mechanisms of tumor immune evasion have been identified through laborious experimental work in some malignancies, we still lack a comprehensive view of tumor evasion routes. In the present study, we used the bulk sample transcriptomic data of 9,403 tumors from 28 different solid cancers (TCGA data) and of 8,034 samples of 22 tissues from healthy donors (GTEx data) to estimate the infiltrate of 16 different immune populations. We observed that the immune infiltration of tumors is not driven only by the tissue of origin, nor determined by the type of cancer; it is also shaped by specific features of each individual tumor. We therefore grouped the solid tumors according to their immune infiltrate patterns and identified clinical, genomic and transcriptomic characteristics that are differently distributed across these immune-phenotype groups. Although some events associated to the immune-phenotypes are cancer type-specific, with the pan-cancer view we were able to define also cross-malignancy features. We observed that tumors tend to progress in three different scenarios of immune selective pressure, which are strongly associated to the clinical stage of the tumor at diagnosis and impact its prognosis. While some genomic drivers (mutations and copy number alterations) exhibiting strong association with certain immune-phenotypes are known to interfere with immune surveillance, some constitute novel candidates. Moreover, pathway enrichment analyses delineated very distinct signaling profiles associated to each of these immune-phenotypes, including replicative programs, metabolism and stromal invasiveness. Finally, we discuss which of these tumor-intrinsic features likely drive immune-edition, and which are probably bystanders of the development of each tumor in its micro-environment.

AN IMPROVED ASSEMBLY AND ANNOTATION OF THE MELON GENOME

Radhakrishnan Sabarinathan[1], Loris Mularoni[1], Jordi Deu-Pons[1], Abel Gonzalez-Perez[1], Núria López-Bigas[1,2]

1. IRTA, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, 08193 Bellaterra, Barcelona, Spain.
2. CRAG, Centre for Research in Agricultural Genomics, 08193 Bellaterra, Barcelona, Spain.
3. Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, 305-8572, Japan.
4. Institute of Plant Sciences Paris-Saclay (IPS2), INRA, CNRS, University of Paris-Sud, University of Evry, University Paris-Diderot, Sorbone Paris-Cit_e, University of Paris-Saclay, Orsay, France.
Presenter e-mail: valentino.ruggieri@irta.cat

In the last years, high-throughput genomic and transcriptomic data as well as new bioinformatic approaches were made available in melon. Taking profit from these genomic resources and technical advances, an updated version of the melon genome (v3.6.1) was obtained. The new genome assembly includes the correction of the orientation for more than 20% of the scaffolds and a better definition of the gaps extension. In addition, the possibility to rely on public and private RNA-seq collections as well as the development of a new ad hoc repeat annotation, allowed to proper define the gene models, updating the previous release with 8,500 new genes, correcting the structure of about 6,000 genes and removing 4,000 genes that lacked any supporting evidences. This updated annotation accounting for 29,980 protein-coding genes offers refined gene structures and improved functional description of many genes. Comprehensively, these results lead to an improved melon genome annotation (CM4.0). To make all the new resources easily exploitable and completely available for the scientific community, a new Melonomics genomic platform was designed that is available at http://melonomics.net. In particular, it hosts a customized instance of JBrowse. The browser currently incorporates, besides the genome reference, 15 tracks, including the v4.0 gene models, two repeat annotations, a cumulative RNA-Seq expression track, two methylation tracks (leaf and root) with the corresponding RNA-Seq tracks and a variome set related to the re-sequencing of seven melon accessions. The updates and the new resources will provide new insights for future studies concerning melon and related species of the Cucurbitaceae family.

COMBINED ANALYSIS OF GENOME SEQUENCING AND RNA-MOTIFS REVEALS NOVEL DAMAGING NON-CODING MUTATIONS IN HUMAN TUMORS

Babita Singh[1], Juan L. Trincado[1] , PJ Tatlow[2], Stephen R. Piccolo[2,3], Eduardo Eyras[1,4]

[1]Pompeu Fabra University (UPF), E08003 Barcelona, Spain.
[2]Department of Biology, Brigham Young University, Provo, Utah, USA.
[3]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA.
[4]Catalan Institution for Research and Advanced Studies (ICREA). E08010 Barcelona, Spain.
Presenter email: babita.singh@upf.edu

A major challenge in cancer research is to determine the biological and clinical significance of somatic mutations in non-coding regions. This has been studied in terms of recurrence, functional impact, and association to individual regulatory sites, but the combinatorial contribution of mutations at common RNA regulatory motifs has not been explored. We developed a new method, MIRA, to perform the first comprehensive study of significantly mutated regions (SMRs) with overrepresented binding sites for RNA-binding proteins (RBPs) in cancer. We found multiple RBP motifs, including SRSF10, PCBP1 and HNRPLL motifs, as well as a specific subset of 5' and 3' splice-site sequences, enriched in cancer mutations. Gene targets showed association to cancer-related functions, and analysis of RNA sequencing from the same samples identified alterations in RNA processing linked to these mutations. MIRA facilitates the integrative analysis of multiple genome sites that operate collectively through common RBPs and can aid in the interpretation of non-coding variants in cancer. MIRA is available at https://github.com/comprna/mira.

Keywords: cancer, non-coding mutations, RNA-processing, RNA binding proteins

INTEGRATED DE NOVO AND REFERENCE-GUIDED APPROACH FOR THE COMPLETE ANNOTATION OF THE FICUS CARICA GENOME

David Tomás[1], Annalisa Tarantino[2], Giuseppe Ferra[2], Andreu Paytuví-Gallart[1], Rosa Barcelona[1], Walter Sanseverino[1], Riccardo Aiese Cigliano[1], Agata Gadaleta[2]

1. Sequentia Biotech SL, Carrer Comte d'Urgell 240, 08036 Barcelona, Spain.
2. Università degli Studi di Bari Aldo Moro,Dipartimento di Scienze Agro Ambientali e Territoriali (DiSAAT). Piazza Umberto I. Bari, Italy.
Presenter e-mail: dtomas@sequentiabiotech.com

Ficus carica L. is a diploid species member of the Moraceae family, which is cultivated as a fruit tree also known as common fig. Edible figs can be produced either as part of sexual reproduction through pollination or by parthenocarpy in the female flowers. The latter are particularly appreciated by the consumers for the lack of seeds but are produced only by specific cultivars.
The University of Bari (Italy) has collected a large collection of *Ficus carica* cultivars, some of which can be an interesting resource of genetic variation for breeding and for understanding the parthenocarpic production of figs.
The genome of *Ficus carica* has been recently sequenced (Mori K, et al. Sci Rep 2017), however no genome annotation has been released. The aim of our work is to produce a representative genome annotation by using an integrated pipeline based on: 1) RNA-seq data obtained from fruits of parthenocarpic and non-parthenocarpic varieties; 2) public transcriptome sequences; 3) *ab initio* genome annotation.
RNA-seq reads were mapped against the reference genome sequence with STAR (version 2.5.0c) and then a reference-guided transcriptome assembly was performed with Trinity. After filtering, about 50,866 transcripts were obtained. Our assembled transcriptome was then merged with a set of transcripts produced by Liceth Solorzano Zambrano, et al. (2017) and used as input for the Maker pipeline. At the same time an *ab initio* annotation was performed with Augustus which was also fed to Maker. A raw annotation is now being analyzed to remove possible artifacts before proceeding to functional annotation.
The final step of the project will be the differential expression analysis of parthenocarpic vs non-partenocarpic varieties in order to identify candidate genes for the production of seedless fig fruits.

COMPUTATIONAL ASSESSMENT OF CLINICAL RELEVANCE IN PRE-CLINICAL CANCER MODELS

Vladimir Uzun, Ian Sudberry

Department of Molecular Biology and Biotechnology, University of Sheffield, S10 2TN Sheffield, United Kingdom.
Presenter e-mail: vuzun1@sheffield.ac.uk

Pre-clinical cancer models, such as tumour-derived cell-lines and animal models, are essential in cancer research. Consistently used as a platform to investigate mechanism of action, they can identify potential biomarkers prior to clinical trials where similar exploration is more complicated and expensive. However, whilst cell-lines are the most used pre-clinical model, their applicability in certain settings is questioned because of the difficulty of aligning the appropriate cell-lines with a clinically relevant disease segment.
We aim to develop computational tools which would determine, for some pre-clinical model, suitability for clinical experiments, and the most relevant disease segment. This increase the information researchers have at their disposal when choosing a pre-clinical model prior to the experiments and, thus, potentially reduce the usage of unsuitable models and increase the reliability of conducted experiments. Genomics profiling data from patient tumours (The Cancer Genome Atlas) and cell-lines (Cancer Cell Line Encyclopaedia) were used to train and test the methods. Machine learning techniques (random forests, principal component analysis, Gaussian processes) were applied to create predictive models based on patient training data. Their accuracy was evaluated on the patient test data and then applied to cell-line data.
Endometrial and breast cancer classification achieved good correspondence with established subtypes (around 0.90 AUC). With the appropriate classifiers (copy-number for endometrial, expression for breast), cell-lines mostly accurately differentiated into respective subtypes. Cancer-related genes were predominant in the most influential genes in the models' decision making.
Whilst most cell-lines associated with clinically relevant segments, a significant number were ambiguous. Furthermore, cell-line suitability scores across different subtypes were not complementary - inappropriate cell-line for one subtype isn't necessarily more likely to be appropriate for the other. We will refine the methodology and create an online scoring tool aiming to improve the usage of pre-clinical cancer models in therapeutic testing.

HYDROPHOCIC SIMILARITY BETWEEN MOLECULES: APPLICATION TO THREE-DIMENSIONAL MOLECULAR OVERLAYS WITH PHARMSCREEN

Javier Vazquez[1,2], Alessandro Deplano,[1] Albert Herrero[1], Enric Gibert[1], Tiziana Ginex[2], Obdulia Rabal[2], Julen Oyarzabal[3], Enric Herrero[1], F. Javier Luque[2]

1. Pharmacelera, Plaça Pau Vila, 1, Sector 1, Edificio Palau de Mar, Barcelona 08039, Spain.
2. 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.
3. Small Molecule Discovery Platform, Molecular Therapeutics Program, Center for Applied Medical Research (CIMA), University of Navarra, Avda. Pio XII 55, Pamplona E-31008, Spain.
Presenter e-mail: javier.vazquez@pharmacelera.com

Molecular alignment is a standard procedure for measurements of the 3D similarity between compounds and pharmacophore elucidation. This process 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 donor /acceptor groups in a given compound. A brief description of the method, as implemented in the software package PharmScreen, including discussion on the calculation of the fractional hydrophobic contributions within the quantum mechanical version of the MST continuum method, and the procedure utilized for searching of the optimal superposition between molecules, is presented.

DECIPHERING MICRORNA TARGETS IN PANCREATIC CANCER USING MIRCOMB R PACKAGE

Maria Vila-Casadesús[1,2], Elena Vila-Navarro[1], Giulia Raimondi[3], Cristina Fillat[3], Antoni Castells[1], Juan José Lozano[1,2], Meritxell Gironella[1]

1. Gastrointestinal & Pancreatic Oncology Group, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD) / Hospital Clínic of Barcelona/ Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Catalonia, Spain.
2. Bioinformatics Platform, CIBEREHD, Barcelona, Catalonia, Spain.
3. Gene Therapy and Cancer, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)/ Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona, Catalonia, Spain.
Presenter e-mail: maria.vila@ciberehd.org

MiRNAs are small non-coding RNAs that negatively regulate mRNA expression. They play important roles in cancer but little is known about the specific functions that each miRNA exerts in each type of cancer. More knowledge about their specific targets is needed to better understand the complexity of molecular networks taking part in cancer. In this study we report the miRNA-mRNA interactome occurring in pancreatic cancer by using a bioinformatic approach called miRComb, which combines tissue expression data with miRNA-target prediction databases (TargetScan, miRSVR and miRDB). MiRNome and transcriptome of 12 human pancreatic tissues (9 pancreatic ductal adenocarcinomas and 3 controls) were analyzed by next-generation sequencing and microarray, respectively. Analysis confirmed differential expression of both miRNAs and mRNAs in cancerous tissue versus control, and unveiled 17401 relevant miRNA-mRNA interactions likely to occur in pancreatic cancer. They can be sorted depending on the degree of negative correlation between miRNA and mRNA expression. Results highlighted the importance of miR-21 and miR-148a interactions among others. A CRISPR-Cas9 cellular model was generated to knock-out the expression of miR-21 in PANC-1 cells. As expected, the expression of two miRComb miR-21 predicted targets (PDCD4 and BTG2) was significantly upregulated in these cells in comparison to control PANC-1. Concerning miR-148a, MiaPaCa-2 cells overexpressing miR-148a showed significantly lower expression of two miRComb miR-148a targets (ADAM17 and EP300). These results highlight miRComb as a useful tool to filter the huge amount of data obtained when studying miRNA target candidates in a specific context.

INDEPENDENT MULTIFACTORIAL ASSOCIATION ANALYSIS APPLIED TO IMAGING GENETIC STUDIES

N. Vilor-Tejedor[1-3], A. Cáceres[1-3], S. Alemany[1-3], J. Sunyer[1-4], JR. Gonzalez[1-3]

[1] Research Institute for Global Health (ISGlobal), Barcelona, Spain
[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[3] CIBER Epidemiologa y Salud Pblica (CIBERESP), Barcelona, Spain.
[4] IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain.
Presenter e-mail: natalia.vilor@isglobal.org

The joint analysis of multiple sources of data still represents a challenge to adequately evaluate complex behaviors due to the high dimensionality and the specific nature of the data. From a statistical point of view, data involving multiple sources can be formulated as a multiblock framework problem. For instance, Imaging Genetics studies (IG), that are focused on the joint analysis of genomic and neuroimaging phenotypes. In this proposal, we present a novel multifactorial algorithm, referred as Independent Multifactor Association Analysis (IMFA-ICR), which uses Independent Component decomposition to derive relevant features from block data and so improve the amount of variability explained. In addition, this approach improves Multiple Factorial analysis by allowing a prediction step based on a meaningful independent component regression and allows obtaining those features significantly correlated with these components. We evaluated the performance of IMFA-ICR with Multifactorial Analysis and univariate anal- ysis in a simulation study. In addition, a real data analysis was performed to illustrate the importance of improving multivariate assessment in the context of Imaging Genetic studies. Specifically, we used IMFA-ICR to detect genetic features related to structural brain regions, which are known to play an important role in the mechanisms of executive function. We showed how IMFA-ICR outperforms common multifactorial analysis and univariate regressions in terms of variability explained and goodness of fit, demonstrating the stability of the algorithm.

Keywords: ICA, Imaging Genetics, Independent Multifactorial Analysis, data integration.

Acknowledgments
Natalia Vilor-Tejedor is funded by a pre-doctoral grant from the Agència de Gestió d'Ajuts Universitaris i de Recerca (2017 FI B 00636), Generalitat de Catalunya - Fons Social Europeu. This research was also supported by the MTM2015-68140-R grant from the Ministerio de Economia e Innovacion (Spain).

References

Abdi, H., Williams, L. J. and Valentin, D. (2013) Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *WIREs Comp Stat*, 5: 149-179. doi:10.1002/wics.1246
Lee, T.W. (1998) *Independent component analysis: Theory and applications*, Boston, Mass: Kluwer Academic Publishers, ISBN 0-7923-8261-7.

THE REGULOME OF *DROSOPHILA* REGENERATION

Elena Vizcaya[1], Cecilia C. Klein[2], Florenci Serras[1], Rakesh Mishra[3], Roderic Guigó[2] and Montserrat Corominas[1]

1. Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia and Institut de Biomedicina (IBUB), Universitat de Barcelona. Barcelona, Catalonia, Spain.
2. Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.
3. The Centre for Cellular and Molecular Biology (CCMB), Hyderabad, India.
Presenter e-mail: elenavizcayamolina@gmail.com

The ability to regenerate varies greatly not only between species but also between tissues and organs or developmental stages of the same species. Differential activation of the genome, determined by a complex interplay of regulatory elements functioning at the level of chromatin, must be the initial mechanism behind these different regenerative capabilities. Resetting gene expression patterns during injury responses is, thus, shaped by the coordinated action of genomic regions that integrate the activity of multiple sequence specific DNA binding proteins. *Drosophila* imaginal discs, which show a high regenerative capacity after genetically induced cell death, are a great model to interrogate chromatin function trough the regeneration process. Using genome- wide approaches (RNA-seq and ATAC-seq) at different tissue time points after injury we have identified the regulatory elements and the expression profile dynamics governing the process. Our findings point to a global co-regulation of gene expression and provide evidence for a regeneration program driven by different types of Damage Responsive Regulatory Elements (DRRE). Among them, novel-DRRE are found acting exclusively in the damaged tissue, and cooperating with DRRE co-opted from other tissues and developmental stages. Altogether, our results decipher the regulome of regeneration and suggest the existence of a specific toolkit to drive the regenerative capacity.

GENOMIC SELECTION IN POLYPLOID SPECIES: AN APPROACH USING REAL AND SIMULATED OCTOPLOID- STRAWBERRY DATA.

M L Zingaretti[1,2], A. Monfort[3], M. Pérez Enciso[1,4,5]

1. Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Barcelona, Spain.
2. IAPCBA-IAPCH, Universidad Nacional de Villa María, Córdoba, Argentina.
3. IRTA - Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Barcelona, Spain.
4. Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.
5. ICREA, Passeig de Lluís Companys 23, 08010 Barcelona, Spain.
Presenter e-mail: m.lau.zingaretti@gmail.com

Genomic Selection (GS) is the procedure whereby molecular information is used to predict complex phenotype, and is becoming standard in many animal and plant breeding schemes. It relies on linkage disequilibrium (LD) between markers and the causal mutations, without the need to identify them. GS has a long tradition in animal breeding and it has a wide use in agricultural crops, too. However, only a small number of studies have been reported in horticultural crops and in polyploids species. In this study, we have developed a forward simulation tool adapted to polyploids species. We have used real Genotyping by Sequencing (GBS) strawberry dataset as input to simulate a new progeny genomes and phenotypes in a very efficient and flexible way. We have evaluated different genetics architectures, using information on sugar degradation pathways. The prediction capability was assessed using 'GBLUP' and mixed model methodology. A comparison of GS prediction with traditional BLUP, which ignores genomic data and relies only on information from ancestors using a pedigree based relationship matrix, show an average increase in predictive ability of ~15% of GS. This was rather robust to alternative genetic models and suggests that GS can significantly accelerate genetic progress in strawberry.