XI Jornada de Bioinformàtica i Genòmica

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

Patrocinada per:

PROGRAMA

Museu de la Ciència COSMOCAIXA

C/ d'Isaac Newton, 26, Barcelona

15 i 16 de desembre de 2023

COMITÈ ORGANITZADOR:

Marta Melé (BSC)
Santiago Marco-Sola (UPC-BSC)
Mario Cáceres (ICREA, IMIM)
Roderic Guigó (CRG-UPF)

SUPORT:

Mariàngels Gallego (SCB)
Paqui Lorite (SCB)
Rosa Bover (BIB)

PROGRAM (15th December)

8:45 – 9:15 Registration

9:15 – 9:30 Welcome and opening of the symposium
Valentí Farràs (Director del Museu de la Ciència, CosmoCaixa)
Marc Martí-Renom (president SCB)
David Torrents (vicepresident BIB)

SESSION I.

9:30 - 10:15 **Invited Lecture Alejandra Medina (UNAM)** Lupus RGMX: demographic, clinical and genomic characterization of systemic lupus erythematosus in a mexican population cohort

10:15 - 10:30 **Valentin Iglesias (IBB-UAB)** A3D Model Organism Database (A3D-MODB): a database for proteome aggregation predictions in model organisms

10:30 - 10:45 **Arnau Comajuncosa (IRB)** Comprehensive characterization of human druggable pockets through novel binding site descriptors built upon inverse docking

10:45 - 11:00 **Ivan Erill (UAB)** Discovery of flexible and interpretable DNA motifs incorporating structural features

11:00 - 11:30 Coffee Break

SESSION II.

11:30 - 11:45 **Ricardo Moreira (UAB)** Using long read data for a complete characterization of human polymorphic inversions

11:45 - 12:00 **Sara Azidane (STALICLA)** Identification of novel driver risk genes in CNV loci associated with neurodevelopmental disorders

12:00 - 12:15 **Rodrigo Martín (BSC)** A comprehensive benchmarking solution for monitoring, improving and harmonizing somatic variant calling across genomic oncology centers

12:15 - 12:30 **Noemia Morales-Díaz (CRAG)** Computational detection of a transposable element insertion associated with longer rice grain

12:30 – 12:45 **Audald Lloret (ETH Zürich)** Size and composition of haplotype reference panels impact the accuracy of imputation from low-pass sequencing in cattle

12:45 – 13:05 **Sponsored talk: Julien Lagarde (Flomics)** Liquid biopsies and next-generation sequencing for research and clinical applications

13:05 - 14:00 Lunch

14:00 - 15:00 **Poster viewing with authors 1 (posters 1-26)**

SESSION III.

15:00 - 15:45 **Invited Lecture Erik Garrison**

15:45 - 16:00 **Tamara Perteghella (CRG)** The role of GENCODE in unveiling the uncharted Non-Coding Layer of Human and Mouse Transcriptomes

16:00 - 16:15 **José Miguel Ramírez (BSC)** Transcriptional and epigenetic impact of cigarette smoking across human tissues

16:15 - 16:30 **Nadezhda Makarova (UB)** Uncovering the determinants of stop codon readthrough in insects

16:30 - 17:00 Coffee Break

SESSION IV.

17:00 - 17:15 **Miquel Anglada-Girotto (CRG)** Unraveling the splicing factor programs driving cancer

17:15 - 17:30 **Joan Pau Cebrià Costa (CRG)** Unveiling the role of histone post-translational modifications during cell differentiation beyond transcription

17:30 - 17:45 **Oleksandra Soldatkina (BSC)** Transcriptional and histological changes across tissues during human aging (and menopause)

17:45 - 18:00 **Max Ticó-Miñarro (UB)** Selenoproteins challenge genomic annotation in the era of massive sequencing

18:00 - 18:15 **Sílvia Pérez-Lluch (CRG)** The epigenetic logic of gene activation

18:15 – 18:35 **Sponsored talk: Anais González (AWS)** Unlashing the power of genomics: How AWS can help to improve your research and your patients

18:35 - 19:30 **Poster viewing with authors 2 (posters 27-52) and meet the companies**

19:30 - 20:30 Cocktail dinner at CosmoCaixa

PROGRAM (16th December)

SESSION V.

10:00 - 10:45 **Invited Lecture Laura Cantini (CNRS)** Multi-view learning for multi-omics single-cell data integration

10:45 - 11:00 **Franz Arnold Ake (IDIBELL)** Characterization of alternative polyadenylation at single-cell resolution

11:00 - 11:15 **Maria Sopena (BSC)** Transcriptional landscape of the aging immune system at single-cell resolution

11:15 - 11:30 **Marcel Schilling (IDIBELL)** Post-transcriptional regulation in iPSC derived neural and glial cells from Alzheimer disease patients

11:30 - 12:00 Brunch

SESSION VI.

12:00 - 12:15 **Elena Pareja (IRB)** Cell states exploration based on location and surrounding environment

12:15 - 12:30 **Murat Tugrul (FREIE Universität in Berlin)** Evolution of Nuclear Receptor Protein Sequences and DNA Binding Features (Motif and Sites) in Crustacean Genomes: A Case Study of the NR2B Family

12:30 - 12:45 **Marta Olivé-Muñiz (UB)** Genome size variation in Dysdera spiders: an evolutionary comparative analysis

12:45 - 13:00 **Miquel Àngel Schikora (IRB)** Recent gene selection and drug resistance underscore clinical adaptation across Candida species

13:00 - 13:15 **Aina Colomer (IBB-UAB)** Evaluating allele frequency trajectory and selection coefficient estimates from genealogies including ancient DNA

13:15 - 13:30 Best oral communication and poster award.

13:30 - 15:00 Free visit to CosmoCaixa

ABSTRACTS

Oral presentations/ 15th December

SESSION I.

LUPUS RGMX: DEMOGRAPHIC, CLINICAL AND GENOMIC CHARACTERIZATION OF SYSTEMIC LUPUS ERYTHEMATOSUS IN A MEXICAN POPULATION COHORT

Alejandra Medina, Universidad Nacional Autónoma de México

Although higher prevalence, disease activity, damage accumulation and mortality of systemic lupus erythematosus (SLE) are observed among Latin American, North American admixed population, African descendants and Native Americans, the information about SLE in Latin American countries, such as Mexico, is scarce.
Lupus RGMX, ia a multidisciplinary effort to generate a national digital patient registry to enrich the understanding of Mexican people with SLE.

Mexican patients with SLE registered between May 2021 and January 2023 in Lupus RGMX were recruited. Sociodemographic, socioeconomic and clinical characteristics, along with quality of life perception (QoL) were assessed using self-reported data. With more than 2000 registered individuals, these rich database can allow for functional genomic studies of a well characterized population.

As data from healthy individual is of most importance for understanding disease, we launched JAGUAR Project, aiming at characterizing the immune cells of healthy individuals in LATAM.

A3D MODEL ORGANISM DATABASE (A3D-MODB): A DATABASE FOR PROTEOME AGGREGATION PREDICTIONS IN MODEL ORGANISMS

Aleksandra E. Badaczewska-Dawid[1], Aleksander Kuriata[2], Carlos Pintado-Grima[3], Javier Garcia-Pardo[3], Michał Burdukiewicz[3,4], Salvador Ventura[3,*], Sebastian Kmiecik[2,*] and Valentín Iglesias[3,*]
[1] Genome Informatics Facility, Office of Biotechnology, Iowa State University, Ames, 50011 IA, USA
[2] Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland
[3] Institut de Biotecnologia i de Biomedicina (IBB) and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
[4] Clinical Research Centre, Medical University of Białystok, Kilińskiego 1, 15-369, Białystok, Poland
* To whom correspondence should be addressed. Email: valentin.iglesias@uab.cat

Protein aggregation has been associated with aging and different pathologies and represents a bottleneck in the industrial production of biotherapeutics. Numerous past studies performed in *Escherichia coli* and other model organisms have allowed to dissect the biophysical principles underlying this process. This knowledge fuelled the development of computational tools, such as Aggrescan 3D (A3D) to forecast and re-design protein aggregation. Here, I present the A3D Model Organism Database (A3D-MODB) http://biocomp.chem.uw.edu.pl/A3D2/MODB, a comprehensive resource for the study of

structural protein aggregation in the proteomes of 12 key model species spanning distant biological clades. In addition to A3D predictions, this resource incorporates information useful for contextualizing protein aggregation, including membrane protein topology and structural model confidence, as an indirect reporter of protein disorder. A3D-MODB is openly accessible without need for registration. We anticipate the A3D Model Organism database becoming an important platform for conducting in-depth multi-species aggregation analyses.

COMPREHENSIVE CHARACTERIZATION OF HUMAN DRUGGABLE POCKETS THROUGH NOVEL BINDING SITE DESCRIPTORS BUILT UPON INVERSE DOCKING

Arnau Comajuncosa-Creus 1, Miquel Duran-Frigola 1, 2 , Xavier Barril 3,4 and Patrick Aloy 1,4,* 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. Ersilia Open Source Initiative, Cambridge, UK. 3. Facultat de Farmàcia and Institut de Biomedicina, Universitat de Barcelona, Barcelona, Catalonia, Spain 4. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

Pocket descriptors characterize protein binding sites in the shape of numerical vectors. Unlike small molecule fingerprints, strategies to derive such descriptors are scarce and usually exhibit limited applicability. We herein present PocketVec, a novel approach to generate a pocket descriptor for any protein binding site of interest based on the chemogenomics principle stating that proteins having similar pockets do bind similar ligands. We first benchmark PocketVec descriptors in several pocket similarity exercises and compare its performance with state of the art methods. Indeed, PocketVec ranks as the 2nd best pocket descriptor to assess pocket similarity while overcoming all limitations of existing strategies. We then gather structural information from the PDB and AlphaFold DB to identify all pockets in the Human Proteome and to further characterize them through PocketVec descriptors. In fact, such characterization simplifies the exploration of the pocket space in a high-throughput manner and the evaluation of pocket similarity from a proteome-wide perspective. We finally demonstrate that the use of PocketVec descriptors enables the identification of similarities and dissimilarities between protein domains that cannot be uncovered through sequential or structural comparisons alone, which may be key to study multiple-target drug binding events

DISCOVERY OF FLEXIBLE AND INTERPRETABLE DNA MOTIFS INCORPORATING STRUCTURAL FEATURES

Elia Mascolo1, Quinn Mood1, Álex Velasco Cañete de Cardenas2, Raül Gómez Buisan3, Ivan Erill2,1 1 Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, MD, USA 2 Departament d'Enginyeria de la Informació i de les Comunicacions, Universitat Autònoma de Barcelona, Bellaterra, Spain 3 Universitat Oberta de Catalunya, Barcelona, Spain

Cells must coordinate gene expression in response to changes in their environment. Transcription factors (TF) regulate gene expression by binding to specific sequence patterns, known as motifs, in the vicinity of target genes. The computational study of transcriptional regulatory networks (TRN) is predicated on the availability of models describing the DNA binding affinity of transcription factors. Motif discovery algorithms have been developed to

efficiently locate overrepresented sequence patterns in sets of sequences targeted by a transcription factor. These methods assume that binding affinity is adequately modeled by considering short contiguous segments of DNA represented by a rigid position frequency matrix. Many transcription factors, however, act as oligomers and target multipartite sequence patterns located at variable distance. Furthermore, transcription factors are also known to rely on structural properties of DNA, such as minor groove width, to identify its targets. As a consequence, the transcriptional regulatory networks defined by these transcription factors cannot be effectively studied with bioinformatics approaches. Here we describe a new model of DNA binding affinity that extends conventional matrix models through the incorporation of flexible connections among sub-motifs and direct recognition of DNA shape features into a log-likelihood ratio framework. We show how this model can be incorporated into a genetic programming framework to perform discriminative motif discovery and recover complex, DNA sequence- and structure-based motifs from sets of unaligned sequences. The formal derivation of this new model using a log-likelihood framework guarantees that the inferred affinity models are intuitively modular, comparable across transcription factors and directly interpretable by users.

SESSION II.

USING LONG READ DATA FOR A COMPLETE CHARACTERIZATION OF HUMAN POLYMORPHIC INVERSIONS

Ricardo Moreira[1,2], Illya Yakymenko[1,2], Konstantinos Karakostis[1,2], Andrés Santos[3], Jaime Martinez-Urtaza[3], Marta Puig[1,3], Mario Cáceres[1,2,4]
1. Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
2. Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute (IMIM), Barcelona, Spain.
3. Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
4. ICREA, Barcelona, Spain.

Inversions are a special type of structural variants (SVs) defined by changes in the orientation of a segment without significant gain or loss of DNA. A big fraction of inversions have large inverted repeats (IRs) at their breakpoints, which complicate considerably the detection of the two alleles. The last years, new techniques able to cross these breakpoints are allowing us to get a full picture of human polymorphic inversions. However, in most cases, just a limited number of individuals has been studied, which precludes the analysis of the effects of the detected variants. Here, we merged inversion predictions from different studies and methods to generate an exhaustive catalogue of IR-mediated inversions. In addition, we developed a computational pipeline that uses Oxford Nanopore Technologies (ONT) long read data to genotype inversions. This method is based on determining the orientation of the reads spanning inversion breakpoints by mapping probe sequences located both outside and inside the inverted regions. We applied this approach to interrogate 631 candidate inversions, ranging from 235 bp to 7.9 Mb and flanked by up to 190 kb long IRs, in a set of 68 individuals. In our dataset, we detected the two orientations in 221 inversions, validating 167 novel inversions detected with different genome-wide techniques. ONT genotypes matched perfectly previous experimental genotypes of 54 inversions, showing that the methodology is

highly accurate. Moreover, 131 additional SVs were identified during the analysis, revealing the complexity of repeat-rich regions. Finally, by comparing the orientation of the inversions in the Pangenome samples, we have also have been able to determine the accuracy of these new reference genomes. Thus, this work demonstrates that long reads present great potential for the characterization of currently missed inversions and other complex genomic regions in multiple individuals, contributing to the understanding of their exact nature and functional implications.

IDENTIFICATION OF NOVEL DRIVER RISK GENES IN CNV LOCI ASSOCIATED WITH NEURODEVELOPMENTAL DISORDERS

Sara Azidane Chenlo[1], Xavier Gallego[1], Lynn Durham[1], Mario Cáceres [2,3], Emre Guney[1], Laura Pérez-Cano[1]

[1] STALICLA Discovery and Data Science Unit, World Trade Center, Moll de Barcelona, Edif Este, Barcelona, 08039, Spain

[2] Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain.

[3] ICREA, 08010 Barcelona, Spain.

Copy number variants (CNVs) are genome-wide structural variations involving the duplication or deletion of large nucleotide sequences, which can range from a few kilobases (kb) to several megabases (Mb). While these types of variations can be commonly found in humans, large and rare CNVs including coding sequence gains or losses are known to contribute substantially to the development of various neurodevelopmental disorders (NDDs), and particularly to autism spectrum disorder (ASD). Nevertheless, given that these NDD-risk CNVs cover broad regions of the genome, it is particularly challenging to pinpoint the critical gene(s) responsible for the manifestation of the phenotype. In this study, we performed a meta-analysis of CNV data from 11,570 NDD patients and 4,114 controls from the SFARI gene database to identify NDD-risk regions and to later determine the deletion and/or duplication of which of the genes entailed within these broad regions were driving the expression of the phenotype. We identified 38 NDD-risk CNV loci surpassing Bonferroni correction, including 23 novel regions, and provided evidence for dosage-sensitive genes within these regions being significantly enriched for known NDD-risk genes and pathways. In addition, a significant proportion of these genes was found to i) converge in protein-protein interaction networks; ii) be among most expressed genes in the brain across all developmental stages; and iii) carry deletions that are significantly over-transmitted to individuals with ASD within multiplex ASD families from the iHART cohort. Finally, we conducted a burden analysis using 3,708 NDD cases from Decipher and iHART and 2,504 neurotypical controls from iHART and the 1000 Genomes database that resulted in the validation of the association of 154 dosage sensitive genes driving risk for NDDs, including 22 novel NDD-risk genes. Importantly, most NDD-risk CNV loci entail multiple NDD-risk genes in agreement with a polygenic model associated with the majority of NDD cases.

A COMPREHENSIVE BENCHMARKING SOLUTION FOR MONITORING, IMPROVING AND HARMONIZING SOMATIC VARIANT CALLING ACROSS GENOMIC ONCOLOGY CENTERS

Rodrigo Martín[1], Nicolás Gaitán[1], Frédéric Jarlier[2,3,4,5], Matias Mendeville[6], Lars Feuerbach[7], Henri de Soyres[2,3,4,5], Tom Gutman[2,3,4,5], Montserrat Puiggròs[1], Alvaro Ferriz[1], Daphne van Beek[6], EUCANCan Consortium, Asier Gonzalez[1], Lucía Estelles[8], Ivo Gut[8], Salvador Capella-Gutierrez[1], Lincoln D. Stein[9,10], Benedikt Brors[7,11], Edwin Cuppen[6,12], Romina Royo[1], Philippe Huppé[2,3,4,5,13], David Torrents[1,14]

1. Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, 08034, Spain.
2. Institut Curie, Paris, F-75005, France. 3U900, Inserm, Paris, F-75005, France. 4PSL Research University, Paris, France.
5. Mines Paris Tech, Fontainebleau, F-77305, France.
6. Hartwig Medical Foundation, Amsterdam, The Netherlands.
7. Applied Bioinformatics Division, German Cancer Research Center (DKFZ), Heidelberg, 69120, Germany.
8. Centro Nacional de Análisis Genómico, Barcelona, 08028, Spain. 9Department of Molecular Genetics, University of Toronto, Toronto, Canada.
10. Ontario Institute for Cancer Research, Toronto, Canada.
11. German Cancer Consortium (DKTK), Heidelberg, Germany.
12. Center for Molecular Medicine and Oncode Institute, University Medical Center Utrecht, Utrecht, The Netherlands.
13. UMR144, CNRS, Paris, F-75005, France.
14. Instituci´o Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, 08010, Spain.

Presenter e-mail: rodrigo.martin@bsc.es

The identification and characterization of the somatic genomic variation associated with the biology of tumors is one of the central pillars of Cancer Research and Personalized Medicine. The quality and the scope of the somatic variant calling determine the reliability and the impact of Cancer Genomic studies, as well as their potential downstream clinical applicability. But the overall quality, scope and consistency of the analysis of the somatic genome across different centers and studies remain significantly limited, affecting not only the overall outcome and the reach of cancer studies, but also the possibilities of improving discovery through the sharing and integration of datasets and results across centers. With the aim of providing users with actionable recommendations for the overall improvement and standardization of somatic variant identification strategies across research environments, we have developed ONCOLINER, an integrated platform with benchmarking data and tools for the detailed assessment, improvement and quality-based harmonization of analysis pipelines across centers. This will not only improve the overall efficiency of somatic variant identification globally, but it will also enable and guide emerging multi-center environments to share and integrate cancer datasets and results.

COMPUTATIONAL DETECTION OF A TRANSPOSABLE ELEMENT INSERTION ASSOCIATED WITH LONGER RICE GRAIN

Noemia Morales-Díaz1 , Raúl Castanera1 , Josep Casacuberta1 1 Centre for Research in Agricultural Genomics, CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Bellaterra, Barcelona, 08193, Spain

Transposable elements (TEs) are DNA sequences that can be mobilized to a different genome locaEon due to the transposiEon process. This change can alter the regulaEon of genes and potenEally contribute to geneEc variability. Genome-wide associaEon studies (GWAS) are a used methodology to examine the relaEonship between geneEc variants and agronomically important traits. Here we used short-read paired-end informaEon to detect 152K transposon inserEon polymorphisms (TIPs) from 738 rice varieEes. We used TIPs as geneEc markers for GWAS, idenEfying putaEve causaEve inserEons accountable for trait variability in grain length. Our results show that the presence of an LTRretrotransposon inserEon is associated with longer indica-3 rice grains. This inserEon was idenEfied 961 base pairs upstream of a gene encoding a calmodulin-binding domain containing protein, as is the case in other well-known seed regulators like GW5. Gene expression analyses indicate that the expression of the candidate gene is significantly reduced when the LTR-retrotransposon is present. BioinformaEc analyses using the rice pangenome suggest that the inserEon could comprise a tandemly duplicated retrotransposon inserEon. Studying this scenario may be useful for the idenEficaEon of a new gene controlling grain structure and to shed light on the dynamics of the repeated sequences in the evoluEonary process of rice.

SIZE AND COMPOSITION OF HAPLOTYPE REFERENCE PANELS IMPACT THE ACCURACY OF IMPUTATION FROM LOW-PASS SEQUENCING IN CATTLE

Audald Lloret (ETH Zürich)

Millions of cattle are genotyped every year for the purpose of genomic prediction. Low-coverage whole-genome sequencing (lcWGS) followed by genotype imputation is a cheap alternative to routine microarray-based genotyping. We assessed the impact of haplotype reference panel composition and sequencing coverage on the accuracy of lcWGS imputation in a target population consisting of cattle from the Brown Swiss (BSW) breed.
We showed that GLIMPSE can accurately impute sequence variant genotypes into cattle genomes sequenced at low coverages. For instance, a same-breed haplotype panel consisting of 75 sequenced samples enabled us to genotype more than 13 million sequence variants in animals sequenced at 0.5-fold sequencing coverage with F1 scores greater than 0.9. Overall, same-breed haplotype reference panels with n = 150 sequenced samples outperformed multibreed panels for sequencing coverages lower than 1-fold, including low allele frequencies. In absence of an adequately sized breed-specific panel (e.g., when less than 30 animals with sequence data are available), F1 scores of 0.9 could also be accomplished either by increasing the sequencing coverage of the target samples or by enlarging the reference panel with distantly related samples from other breeds. Nevertheless, since suboptimal haplotype reference panels lack variants private to the target breed, the resulting imputed lcWGS data are depleted for this type of variation.
Keywords: cattle, lcWGS, imputation, genotyping, variant calling

LIQUID BIOPSIES AND NEXT-GENERATION SEQUENCING FOR RESEARCH AND CLINICAL APPLICATIONS

Julien Lagarde, Flomics Biotech

Flomics is a biotechnology company from Barcelona operating in the fields of omics and bioinformatics. Our focus centers on pioneering a cutting-edge, minimally invasive liquid biopsy solution designed for the early and accurate detection of cancer. This is achieved through the integration of plasma cell-free RNA profiling with advanced bioinformatics and machine learning techniques. In addition, we extend our expertise to offer a diverse array of genomics and bioinformatics services tailored for both researchers and clinicians. Our comprehensive support spans from patient screening to large-scale biomarker discovery initiatives, providing end-to-end assistance for basic and translational research, as well as clinical projects. This talk will delve into some of Flomics' latest scientific and technical advancements, particularly focusing on our efforts to develop a gold-standard workflow for the profiling of cell-free nucleic acids in blood plasma.

SESSION III.

THE ROLE OF GENCODE IN UNVEILING THE UNCHARTED NON-CODING LAYER OF HUMAN AND MOUSE TRANSCRIPTOMES

Silvia Carbonell-Sala1 , Gazaldeep Kaur1 , Tamara Perteghella1,2 , Carme Arnan1 , Emilio Palumbo1 , Barbara Uszczynska-Ratajczak3 , Rory Johnson4 , Adam Frankish5 , The GENCODE Consortium, Roderic Guigó1,2 1. Computational Biology of RNA Processing, Bioinformatics and Genomics programme, CRG-Centre de Regulació Genòmica, Dr. Aiguader, 88, 08003, Barcelona, Catalonia, Spain 2. Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain 3. Computational Biology of Noncoding RNA, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland 4. School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4 5. EMBL-EBI, United Kingdom Presenter's email: tamara.perteghella@crg.es

A comprehensive understanding of genomic biological functioning requires a detailed transcriptome-scale view. As part of the GENCODE consortium, we aim to develop experimental approaches that capture and sequence genomic regions overlooked by standard transcriptomic methods. Our goal is to create full-length transcript datasets with minimal manual intervention. Our newly designed comprehensive catalog includes poorly annotated, yet predicted to transcribe, regulatory elements for lncRNA, putative secondary elements, enhancers, and small RNA precursors. This catalog provides an accurate perspective of the long non-coding transcriptome. Herein, we describe our improved CapTrap-CLS method, capturing full-length transcripts, with a focus on the designed lncRNA panel in human and mouse tissues at various developmental stages. To minimize manual intervention and ensure high-confidence full-length transcripts supported by experimental orthogonal data, we utilize our in-house developed pipeline, LyRic. The combination of CapTrap-CLS and

LyRic identifies entirely novel full-length transcripts, marking the largest-ever effort to explore the uncharted fraction of human and mouse transcriptomes. In comparison to the latest human GENCODE reference, we discover over 55,000 novel intergenic transcripts. Orthogonal data have been integrated moving towards the functional aspect of these novel transcripts. As an illustration of their relevance, we present the results obtained by reanalyzing RNA-Seq samples from various studies, spanning a variety of psychiatric and neurodegenerative disorders as well as different brain areas, using our CLS3 "enhanced" annotation. CapTrap-CLS proves its capability in unveiling the functional aspects of the lncRNA transcriptome, thereby improving genome interpretation

TRANSCRIPTIONAL AND EPIGENETIC IMPACT OF CIGARETTE SMOKING ACROSS HUMAN TISSUES

Jose Miguel Ramirez* 1 , Rogério Ribeiro* 2,3 , Oleksandra Soldatkina 1 , Raquel García-Pérez 1 , Pedro G. Ferreira* 2,3 & Marta Melé* 1 . 1. Department of Life Sciences, Barcelona Supercomputing Center (BSC), C/ Jordi Girona 29, 08034 Barcelona, Spain 2. Department of Computer Science, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal 3. Laboratory of Artificial Intelligence and Decision Support, INESC TEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

Tobacco smoke causes 8 million deaths annually, representing the main cause of preventable death worldwide. Yet, our understanding of the molecular mechanisms driving smoking-related health decline and tissue degeneration remains extremely limited as previous molecular studies mainly focused on lung and blood. Here, we characterize the effects of cigarette smoking on gene expression, alternative splicing, DNA methylation and histology across 46 human tissues from the Genotype Tissue Expression Dataset. Differential gene expression analysis shows a widespread impact of smoking in the human body, with some immune and metabolic genes systematically upregulated across tissues. At the DNA methylation level, we observe hypermethylation in chromatin regions targeted by the Polycomb-repressive complex, which are regions associated with developmental functions. Notably, we observe concordant additive effects of smoking and ageing in both gene expression and DNA methylation, suggesting that smoking has similar consequences to those of biological ageing. Using ex-smoker annotation, we study reversibility and find gene- and tissue-specific reversibility rates. We also report different reversibility rates between omics, as changes in DNA methylation are more enduring than in gene expression. In general, expression and methylation effects show very few correlations, as only very high DNA methylation differences result in perceptible changes in gene expression.

UNCOVERING THE DETERMINANTS OF STOP CODON READTHROUGH IN INSECTS

Nadezhda Makarova (UB)

Stop codon readthrough (SCR) occurs when the ribosome does not terminate and instead decodes the stop as a sense codon, promoting amino acid insertion. Recently, SCR was discovered through evolutionary conservation to be abundant in insects, occurring in hundreds of genes, and it may serve regulatory roles. It is the focus of the current study to investigate

the distribution, mechanism, and function of SCR in insects. We developed a computational pipeline to detect SCR through the analysis of ribosome profiling data. We applied it to public Drosophila data, comprising 80 Ribo-seq samples and spanning diverse biological conditions. We quantified ribosome density in the regions between annotated stop codons and the next in-frame stop codon. We applied multiple filtering criteria to identify actively translated extension regions and exclude potential artifacts, resulting in a set of 698 genes with evidence of SCR. Readthrough is activated in different biological conditions, supporting its regulatory role. Functional analysis revealed the involvement of SCR genes in synaptic transmission, sensory organ development, transcriptional regulation, and post-embryonic development. Notably, SCR genes exhibited stronger conservation of stop codons and SCR extensions showed lower protein hydrophobicity than non-SCR extensions, implying higher stability against degradation. Secondary structures in mRNA sequence can promote the readthrough process. We utilized the RNAz program to identify conserved RNA structures by analyzing whole-genome alignments of 23 Drosophila species. Our findings demonstrated that SCR genes are enriched with conserved secondary structures compared to other genes, particularly in proximity of the readthrough stops. Next, we aim to experimentally assay the readthrough activity of these sequences using a dual fluorescent reporter system where a stop codon is located between two fluorescent proteins. As a proof of principle, we tested the sequence surrounding the stop codon of the Tj gene, previously shown to promote SCR in S2 cells, and we verified its SCR activity in our system through flow cytometry. This system holds promise as a tool for detecting SCR events. With further optimization, it can be adapted for pooled screening, potentially leading to the discovery of novel SCR stimulators.

SESSION IV.

UNRAVELING THE SPLICING FACTOR PROGRAMS DRIVING CANCER

Miquel Anglada-Girotto 1 , Samuel Miravet-Verde 2 , Andrea Califano 3,4,5,6,7 , Luis Serrano 1,8,9 1Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. 2Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Zurich, Switzerland. 3 Department of Systems Biology, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, USA 10032 4Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, USA 10032 5Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, USA 10032 6Department of Biochemistry & Molecular Biophysics, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, USA 10032 7Department of Biomedical Informatics, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, USA 10032 8Universitat Pompeu Fabra (UPF), Barcelona, Spain. 9 ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain.

The regulation of exon inclusion through alternative splicing tunes the behavior of the cell by increasing the functional diversity of the transcriptome and the proteome. In cancer cells, splicing factors are co-regulated to generate gene isoform pools that favor tumor progression. However, identifying the disease-driving splicing factor programs remains challenging using single-omic measurements as aberrant splicing factor activities can originate from any type of regulatory alteration. Here, we inferred splicing factor activities solely from their target exon inclusion signatures by repurposing Virtual Inference of Protein activity by Enriched

Regulon analysis (VIPER) with splicing factor-exon networks. Our approach accurately generalized experimental validations involving protein-level modulation, combinatorial perturbations, and protein-protein interactions among splicing factors. Comparing splicing factor activities across 14 types of primary tumors to their healthy counterparts revealed two novel programs of recurrently activated -oncogenic- and recurrently inactivated -tumor suppressor- splicing factors. We demonstrate the coordinated activation and inactivation of these splicing factor programs not only mediate tumorigenesis but also predict patient prognosis as well as sample proliferative status. Altogether, our splicing factor activity analysis unveiled two novel splicing factor programs with a pivotal role in disease initiation and progression. Keywords cancer; alternative splicing; splicing factor; protein activity; VIPER; tumorigenesis

UNVEILING THE ROLE OF HISTONE POST-TRANSLATIONAL MODIFICATIONS DURING CELL DIFFERENTIATION BEYOND TRANSCRIPTION

Joan Pau Cebrià-Costa 1 , Sílvia Pérez-Lluch 1 , Marina Ruiz-Romero 1 , Roderic Guigó 1,2 1. Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona (BIST), Catalonia, Spain. 2. Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain Presenter e-mail: joan.cebria@crg.eu

Histone post-translational modifications (PTMs) are considered a key player in gene expression regulation. After many years of investigation where a correlation has been reported between histone PTMs and gene expression, recently, much evidence in temporal data started to question this assumption, hence remaining the role of these modifications controversial. To address the role of chromatin modifications along development and differentiation, in the lab we have profiled nine histone modifications related to gene activation –H3K4me3, H3K9ac, H3K36me3 and H4K20me1–, enhancer activity –H3K4me1, H3K4me2 and H3K27ac–, and heterochromatin –H3K27me3 and H3K9me3– in four different tissues –antenna, eye, leg and wing– from three developmental stages along the fruit fly development –third instar larva stage, early pupa and late pupa–. The complexity of the system requires new tools to integrate the temporal and spatial components of histone PTMs. Here we present a classification approach for ChIP-Seq temporal and spatial data independent of transcription or genomic features. Our approach is based on considering each histone modification peak as a consequence of two attributes: space and time. Therefore, each peak will be characterized by changes between tissues at the same stage of development (space) and changes between stages of development in the same tissue (time). Finally, taking advantage of Voronoi's diagrams and Euclidean distances we have been able to uncover 11 specific groups of peaks showing different dynamics patterns through time and space. With our approach, we have seen that histone marks are largely stable between tissues and stages of development. H3K4me3 and H3K36me3 are the epigenetic marks showing a higher stability between tissues and stages of development. Even so, our approach allows us to identify specific peak distribution for each histone PTM along cellular differentiation. Finally, the gene-agnostic classification of histone PTMs method used here makes this approach a perfect tool for the discovery of remarkable specificities for cellular differentiation, such as new DNA motifs associated with particular patterns of histone PTMs controlling cellular fate. In conclusion, our classification approach unlocks histone PTMs as a key characteristic of chromatin organization beyond and further the role as a code to regulate gene transcription.

TRANSCRIPTIONAL AND HISTOLOGICAL CHANGES ACROSS TISSUES DURING HUMAN AGING (AND MENOPAUSE)

Oleksandra Soldatkina, Clara Suarez, Marta Mele

Age is a major risk factor for many important diseases, such as cardiovascular disease and osteoporosis. For women aging comes with menopause — the process also associated with several health issues, although largely understudied in its systemic impact on the body. With differential expression analysis, we have shown that aging affects many human tissues to varying extents. Here we look deeper into these results and argue that tissue-specific aging occurs at different paces and time points. To investigate the per-tissue transformation trajectories, we use a deep learning approach on paired samples of histological images and gene expression profiles of tissues we found most affected by aging — arteries and female reproductive organs. We find that a significant part of the age effect on the arteries can be attributed to the gradual development of atherosclerosis, and describe both the gene expression change with the disease and the difference with what occurs in healthy aging. By training a classifier on uterus images of different ages, we identify the point of transition to menopause and characterize the change that comes with it in other body tissues. We find a steep shift in the morphology of the uterus following a gradual change in the ovary, and a systemic cross-tissue change in the expression of genes that have been described in association with anemia, heart disease, and collagen deficiency along with other terms. We believe that these findings contribute to our understanding of the complex transformations the tissues undergo while aging and that this knowledge can help develop personalized medicine to account for age- and menopause-stage-related factors

SELENOPROTEINS CHALLENGE GENOMIC ANNOTATION IN THE ERA OF MASSIVE SEQUENCING

Max Ticó, Marco Mariotti. Department of Genetics, Microbiology and Statistics, Universitat de Barcelona, Barcelona, Spain.

Selenium is a vital trace element that plays a crucial role in various biological processes. It is primarily found in the form of selenocysteine (Sec), which is incorporated into selenoproteins. Sec insertion is performed in response to an in-frame UGA codon, which normally is interpreted as a stop codon but which is "recoded" in selenoproteins. Selenoproteins have many essential roles in vertebrates and other organisms, including redox homeostasis, protein quality control, and metabolism. Selenoproteins are a prominent example of translational recoding, wherein programmed exceptions are made to the genetic code for specific genes. Numerous other forms of translational recoding exist. High-throughput sequencing has led to an abundance of public genome sequences, but accurate annotation remains challenging. Selenoprotein genes pose a particular difficulty due to recoding, often resulting in misannotation. We quantified this phenomenon in Ensembl, considered the gold standard of genomic annotation. Our analysis across 315 genomes showed that only around 11% of selenoproteins were well annotated. ~9% of selenoproteins have no annotation at all. The majority (~80%) have flawed annotations which lack the Sec-encoding UGA; most often, the annotation skips at this codon. Thus, there is an urgent need to incorporate selenoprotein annotation tools to standard gene annotation programs. To facilitate this, we are improving

Selenoprofiles, a computational pipeline able to correctly annotate selenoprotein genes, so that it can automatically produce gene predictions with quality comparable to manual curation. Our approach identifies selenoprotein genes by homology, classifies orthologs, then filters the predictions based on the expected selenoprotein content of taxonomic groups, defined in our prior studies. We advocate for this and analogous approaches to include translational recoding in public gene annotations

THE EPIGENETIC LOGIC OF GENE ACTIVATION

Beatrice Borsari[1], Amaya Abad[1], Vasilis F. Ntasis[1], Silvia González-López[1], Cecilia C. Klein[1,2], Ramil Nurtdinov[1], Diego Garrido-Martín[1,2], Carme Arnan[1], Alexandre Esteban[1], Emilio Palumbo[1], Marina Ruiz-Romero[1], Raül G. Veiga[1], Maria Sanz[1], Bruna R. Correa[1], Rory Johnson[1], Sílvia Pérez-Lluch[1] and Roderic Guigó[1,3]

1. Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona (BIST), Catalonia, Spain.
2. Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia and Institut de Biomedicina (IBUB), Universitat de Barcelona, Barcelona 08028, Catalonia, Spain.
3. Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain.

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

Histone modifications are considered to play a causal role in the regulation of gene expression. This role has been challenged by reports showing that gene expression may occur in the absence of histone modifications. To address this controversy, we generated densely-spaced transcriptomic and epigenomic maps in a time-course cell homogeneous system that occurs with massive transcriptional changes. Our analyses suggest a model that reconciles these seemingly contradictory observations. While chromatin dynamics do not fully recapitulate the dynamics of gene expression, histone modifications are strongly associated with expression specifically at the time of initial gene activation, when they are deposited in a dominant order at promoter regions, generally following gene activation and preceding deposition at enhancers. Here, we provide a model for the epigenetic logic underlying gene activation where modifications are involved in stabilizing rather than driving gene expression, collectively acting as recordings of the recent transcriptional trajectory of activated genes.

16th December

SESSION V.

MULTI-VIEW LEARNING FOR MULTI-OMICS SINGLE-CELL DATA INTEGRATION

Laura Cantini, CNRS, París

Single-cell RNA sequencing (scRNAseq) is revolutionizing biology and medicine. The possibility to assess cellular heterogeneity at a previously inaccessible resolution, has profoundly impacted our understanding of development, of the immune system functioning and of many diseases. While scRNAseq is now mature, the single-cell technological development has shifted to other large-scale quantitative measurements, a.k.a. 'omics', and even spatial positioning. In addition, combined omics measurements profiled from the same single cell are becoming available.

Each single-cell omics presents intrinsic limitations and provides a different and complementary information on the same cell. The current main challenge in computational biology is to design appropriate methods to integrate this wealth of information and translate it into actionable biological knowledge.

In this talk, I will discuss two main computational directions for multi-omics integration, currently explored in my team: (i) joint dimensionality reduction to study cellular heterogeneity simultaneously from multiple omics and (ii) multilayer networks to integrate a large range of interactions between the features of various omics and isolate the regulators underlying cellular heterogeneity.

CHARACTERISATION OF ALTERNATIVE POLYADENYLATION AT SINGLE CELL RESOLUTION IN ALZHEIMER DISEASE

Franz Ake1,2, Ana Gutierrez-Franco1,2, Sandra Maria Fernandez-Moya1,2, Mireya Plass1,2,3
1 Gene Regulation of Cell Identity, Regenerative Medicine Program, Bellvitge Institute for Biomedical Research (IDIBELL), 08908, L'Hospitalet del Llobregat, Barcelona, Spain
2 Program for Advancing Clinical Translation of Regenerative Medicine of Catalonia, P-CMR[C], 08908, L'Hospitalet del Llobregat, Barcelona, Spain
3 Center for Networked Biomedical Research on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), 28029, Madrid, Spain
Presenter e-mail: fake@idibell.cat

Alternative polyadenylation (APA) is a widespread mechanism of gene regulation that generates mRNA isoforms with distinct 3'ends. APA is well known to be regulated during cell differentiation and is a major source of gene regulation in the brain. Proliferating cells tend to have shorter 3' UTRs while differentiated cells have longer 3'UTRs. Changes in APA patterns are not only characteristic of cellular differentiation but also have been associated with pathological processes such as cancer or neurodegenerative diseases like Alzheimer's disease. The rapid development of 3'tag-based single-cell RNA sequencing (scRNAseq) has enabled the study of gene expression and the implementation of methods for describing isoform usage at single cell resolution. Here we present SCALPEL, a tool for quantifying isoforms expression at single cell resolution using 10X Genomics or Dropseq scRNA-seq dataset. SCALPEL isoform

quantification enables to identify an alternative isoform usage between cell populations, states and defined conditions. We used SCALPEL to study the changes in APA during the differentiation of human induced pluripotent stem cells (iPSCs) to neuroprogenitor cells (NPCs). The results from our analysis show clear changes in 3'end usage between iPSCs and NPCs. We aim to use SCALPEL to investigate the role of APA during neural differentiation and how these changes are altered in neurodegenerative disease context.

TRANSCRIPTIONAL LANDSCAPE OF THE AGING IMMUNE SYSTEM AT SINGLE-CELL RESOLUTION

Maria Sopena-Rios 1 , Aida Ripoll-Cladellas 1 , Marta Melé 1 1. Life Sciences Department, Barcelona Supercomputing Center (BSC), Pl. Eusebi Güell 1-3, 08034 Barcelona, Spain

In an era of extended lifespans, the escalating incidence of age-related disorders presents a significant public health challenge. Therefore, developing strategies for promoting healthy aging is crucial. Despite being a universal biological phenomenon, aging is not easily defined. Among the primary hallmarks of aging, a gradual decline of immune functions has been distinguished. This progressive immune-related deterioration, known as immunosenescence, leads to a higher risk of diseases including infections, cancer, or autoimmune disorders. Yet, our understanding of the transcriptional and cellular alterations of the immune system with aging remains elusive. Single-cell RNA sequencing (scRNA-seq) measures gene expression at an unprecedented resolution, enabling the study of both the cell type specific gene expression changes to the cellular composition variation. However, until very recently, scRNA-seq datasets included a limited number of donors, impeding our ability to identify cell-type-specific transcriptional changes driven by individual demographic traits, such as age. To fill this gap, we use the OneK1K cohort, the largest available scRNA-seq dataset, to characterize the age-induced transcriptional variation across immune cells. This comprehensive dataset encompasses nearly one million human peripheral blood mononuclear cells (PBMCs) from 982 individuals, spanning a wide range of ages. We explore age-related changes in cellular abundances using a cell-type-agnostic method. Among the significant shifts in cell type abundances with age, we distinguish a substantial decrease in CD8T naïve and B memory cells, whereas CD8T effector memory and natural killer cells were notably increased. Beyond cellular alterations, our differential expression analysis reveals pervasive aging effects on immune cell types, with CD8T naive cells exhibiting the most substantial impact. We identify a significant bias in the directionality of the differentially expressed genes. Cell types with predominantly downregulated genes exhibit reduced translation and ribosomal function, whereas cell types with mostly upregulated genes showed heightened immune functions and inflammation. Notably, we identify a strong decrease in the expression of the naive marker LRRN3 and a strong upregulation of cytotoxic genes such as GZMH. Lastly, by employing a variance partitioning approach, we quantify the relative contribution of age to gene expression variation. Consistent with our previous analysis, CD8T naive emerges as the most influenced cell type by age. Genes with the largest age contribution were enriched in ribosomal proteins, supporting the observed alterations in translation with age. Overall, our study offers a comprehensive characterization of the impact of age across immune cell types, providing valuable insights into the molecular and cellular alterations underlying aging in the immune system.

POST-TRANSCRIPTIONAL REGULATION IN IPSC DERIVED NEURAL AND GLIAL CELLS FROM ALZHEIMER DISEASE PATIENTS

Marcel Schilling, Ana Gutiérrez-Franco, Franz Ake, Natalie Chaves Cayuela, Nadia Jamshaid, Loris Mularoni, and Mireya Plass

Alzheimer's disease (AD) is the most common cause of dementia, but its pathogenesis still remains poorly understood. Post-transcriptional regulation, e.g. via RNA-binding proteins (RBPs) or alternative cleavage and poly-adenylation (APA) has been shown to be implicated in AD. However, whether the corresponding alterations play a causative role or rather represent consequential symptoms has not yet been shown. To address this question, and ideally identify potential new therapeutic targets, we investigate the transcriptional landscape of single cells progressing from iPSCs derived from AD patients and healthy controls throughout their differentiation to neural and glial cells. We detect up to 8000 genes per cell across more than 80000 cells over five developmental time points. Clustering and transcriptome profile analyses identified all major cell types (iPSCs, NECs, NPCs, neurons and astroctyes) including various precursor and mature cell populations in both conditions (AD and control). Differential gene expression analyses identified several differentially expressed genes that have been previously linked to AD. Amongst the differentially expressed genes were several RBPs, implicating the corresponding regulatory networks as potential targets for further investigation. Differential 3' UTR usage analyses using SCALPEL revealed that several genes previously implicated in AD pathogenesis display changes in APA between AD and control cells. Here we present those results and our efforts to validate individual findings, and integrate these lines of evidence towards a better understanding of post-transcriptional regulation in pre-neurodegeneration AD.

SESSION VI.

CELL STATES EXPLORATION BASED ON LOCATION AND SURROUNDING ENVIRONMENT

Elena Pareja-Lorente1 and Patrick Aloy1,2,∗ 1 Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain 2 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

To whom correspondence should be addressed. Spatial localization of cells in tissues plays a pivotal role in defining biological function. This organization is crucial for understanding the cellular microenvironment, cell-cell interactions, and organ function. Single-cell RNA sequencing (scRNA-seq) has advanced our comprehension of gene expression profiles across distinct cell types; however, this technique inherently loses the spatial dimension. While recent advances in spatial sequencing techniques facilitate gene expression profiling across tissue sections, they are still lacking single-cell resolution. This limitation hampers our ability to discern variations in cellular states based on location. In this context, we have developed a systemic approach that utilizes existing techniques to align individual cells to their spatial positions. With this detailed localization and gene expression profiles, we can then conduct diverse downstream analysis, enabling a deeper exploration of cell states influenced by location and the surrounding microenvironment. We apply this strategy to 3 different datasets

representing three distinct tissues: Breast Cancer (ER+ and TNBC), mouse cortex, and human heart. First, we mapped single cells to their spatial locations and evaluated how different localizations impact global gene expression. We hypothesize that the resolution might be insufficient to discern differences between the same cell types. Subsequently, our objective was to pinpoint genes with niche-dependent expression however, our findings were ambiguous. Finally, we explored variations in intercellular communication, taking into account the regional variations in cell composition, yet clear patterns of concordance between co-location and cell interaction were absent. The lack of agreement between spatial location and interacting cells is consistent with our earlier findings of small differences between gene expression profiles, as the majority of cell-cell communication methods utilize the complete gene expression vector rather than just focusing on the differentially expressed genes.

EVOLUTION OF NUCLEAR RECEPTOR PROTEIN SEQUENCES AND DNA BINDING FEATURES (MOTIF AND SITES) IN CRUSTACEAN GENOMES: A CASE STUDY OF THE NR2B FAMILY

Murat Tugrul[1] and Ferran Palero[2]

[1]Institute of Biology, Freie Universitat Berlin, Berlin, Germany

[2]Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia, Paterna, Spain.
Presenter email: murat.tugrul@fu-berlin.de

Nuclear receptors (NRs), a special class of transcription factors (TFs), control gene expression in response to signalling molecules and environmental stimuli, influencing physiology and fitness. NR evolution is crucial for metazoan diversification and adaptation, yet the coevolution of NR protein sequences, DNA binding preferences and sites remains poorly understood. We focus on a key family NR2B (usp/rxr) and on crustaceans, a highly diverse group of invertebrates occupying various ecological niches, with available whole genome sequences and transcriptomic data. Our findings reveal that the evolution of NR2B protein sequences is closely aligned with the phylogeny of Crustacea, in agreement with a common ancestor shared between Copepoda and Cirripedia, while Euphausiacea and Decapoda diverged later within Malacostraca. Moreover, using a machine learning approach, we predict binding motifs for various taxa, revealing mostly conserved motifs with limited evolutionary change. However, interestingly, motif similarity between species shows unpredictability, loosely following main phylogenetic clades. Furthermore, we explore the binding site evolution within crustacean genomes. We find slight motif changes trigger numerous transcription factor binding site (TFBS) turnovers, potentially rewiring gene regulatory networks and leading to distinct phenotypes. This study provides new insights into the coevolution of nuclear receptors and their DNA binding features in crustaceans, enhancing our understanding of the molecular mechanisms governing their development and adaptation to diverse ecological conditions.

GENOME SIZE VARIATION IN DYSDERA SPIDERS: AN EVOLUTIONARY COMPARATIVE ANALYSIS

Marta Olivé-Muñiz(1,2) , Vadim A. Pisarenco(1,2), Adrià Boada-Figueras(1,2), Paula Escuer(1,2) , Miquel A. Arnedo(2,3), Pablo Librado(4), Alejandro Sánchez-Gracia(1,2), Sara Guirao-Rico(1,2) and Julio Rozas(1,2) . (1) Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain. (2) Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain. (3) Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Universitat de Barcelona, Barcelona, Spain. (4) Institut de Biologia Evolutiva, (CSIC- Universitat Pompeu Fabra), Barcelona, Spain.

The red devil spider genus Dysdera colonised and underwent a striking species diversification in the Canary Islands, where up to 60 endemic species have been described. This radiation was associated with repeated events of dietary shifts and specialisations to feed on woodlice, endorsed by morphological, behavioural, metabolic and transcriptomic evidences. In addition, despite belonging to the same genus, Canary Islands endemic species have a much smaller genome (1.7 Gb) compared to closely related continental species (3.3 Gb), likely as a result of a genome reduction in the first group. This scenario posits an ideal framework to explore the genomic basis of adaptive radiations, island diversification and speciation, and genome size evolution. To gain insights into the molecular basis of this candidate adaptive radiation, we performed a comparative genomic analysis across high-quality genome assemblies (three of them at the chromosome-level) of six Dysdera species, comprising five Canary Islands endemic species plus one continental relative. Here, we investigate how changes in different genomic features (number of protein-coding genes, intron sizes, intergenic distances, gene family sizes, TE content and the abundance of other repetitive elements) can account for the different genome size observed between island and continental species, while discussing their putative adaptative role.

RECENT GENE SELECTION AND DRUG RESISTANCE UNDERSCORE CLINICAL ADAPTATION ACROSS CANDIDA SPECIES (manuscript accepted at Nature Microbiology)

Miquel Àngel Schikora-Tamarit 1,2 , Toni Gabaldón 1,2,3,4,* 1) Barcelona Supercomputing Centre (BSC-CNS). Plaça Eusebi Güell 1-3, 08034 Barcelona, Spain. 2) Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain. 3) Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. 4) Centro Investigación Biomédica En Red de Enfermedades Infecciosas, Barcelona, Spain. *Author for correspondence: toni.gabaldon@bsc.es

Understanding how microbial pathogens adapt to treatments, humans and clinical environments is key to infer mechanisms of virulence, transmission and drug resistance. This may help improve therapies and diagnostics for infections with poor prognosis, such as those caused by fungal pathogens, including Candida. Here, we analyzed genomic variants across ~2,000 isolates from six Candida species (C. glabrata, C. auris, C. albicans, C. tropicalis, C. parapsilosis and C. orthopsilosis) and identified genes under recent selection, suggesting highly complex clinical adaptation. These involve species-specific and convergently-affected adaptive mechanisms, such as adhesion. Using convergence-based genome-wide association studies we identified known drivers of drug resistance alongside potentially novel players. Finally, our analyses reveal an important role of structural variants, and suggest an

unexpected involvement of (para)sexual recombination in the spread of resistance. Our results provide insights on how opportunistic pathogens adapt to human-related environments and unearth candidate genes that deserve future attention.

EVALUATING ALLELE FREQUENCY TRAJECTORY AND SELECTION COEFFICIENT ESTIMATES FROM GENEALOGIES INCLUDING ANCIENT DNA

Aina Colomer-Vilaplana[1,2], Sònia Casillas[1,2], Antonio Barbadilla[1,2], Leo Speidel[3,4]

1. Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Parc de Recerca, Mòdul B, 08193 Cerdanyola del Vallès, Catalonia, Spain
2. Department of Genetics and Microbiology, Facultat de Biociències, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Catalonia, Spain
3. Genetics Institute, University College London, 99-105 Gower St, London WC1E 6AA, United Kingdom
4. Francis Crick Institute, 1 Midland Road, London NW1 1AT, United Kingdom

Presenter e-mail: aina.colomer@uab.cat

Humans have successfully adapted to different environments during their migration across the continents. Yet, a question still unsolved is the extent to which selection has played a role in shaping our genomes throughout these migrations. Recent advances in genomic methodologies, including the use of ancient DNA, have provided new opportunities to study our genetic past. The availability of large cohorts of ancient DNA samples from single populations has enabled the inference of allele frequency trajectories and the associated selection coefficients. Recently, new methods for inferring genealogies -such as Relate and tsinfer- as well as new methods that feed from these genealogies -like Clues- have made it possible to extrapolate allele frequency trajectories from sequencing data of modern-day samples. Here we evaluate the effectiveness of Relate and Clues benchmarking these methods against known and inferred genealogies from simulated data using SLiM. Moreover, we develop our own strategy to infer a selection coefficient from a pre-estimated genealogy incorporating ancient DNA. We test this method under different selective regimes ranging from neutral to strong selection, showing that aDNA substantially improves selection estimates. With our proposed method we aim for a better understanding of the genomic marks left by selection over the past tens of thousands of years.

Posters

1

TARGETING THE WNT SIGNALING PATHWAY: A NOVEL PREDICTIVE SIGNATURE FOR NEOADJUVANT CHEMOTHERAPY RESPONSE IN MUSCLE-INVASIVE BLADDER CANCER

Ariadna Acedo-Terrades[1], Júlia Perera-Bel[1], Marta Bódalo-Torruella[1], Maria Gabarós[1], Nuria Juanpere[1], Marta Lorenzo[1], Alejo Rodriguez-Vida[1], Oscar Buisan[2], Eulàlia Puigdecanet[3], Eduardo Eyras[1–4], Tamara Sanhueza[2], Lara Nonell[5], Joaquim Bellmunt[1–6]. [1]Hospital del Mar Medical Research Institute; Barcelona, Spain, [2]Germans Trias i Pujol Research Institute (IGTP); Badalona, Spain, [3]UVic-UCC; Barcelona, Spain, [4]EMBL Australia Partner Laboratory Network at the Australian National University; Canberra, Australia, [5]Vall d'Hebron Institute of Oncology (VHIO); Barcelona, Spain, [6]Division of Hematology and Oncology, Beth Israel Deaconess Medical Center; Boston, USA

In muscle-invasive bladder cancer (MIBC), neoadjuvant cisplatin-based chemotherapy (NAC) has become a standard of care prior to cystectomy for eligible patients based on the improved disease-specific and overall survival. Downstaging to non-MIBC at cystectomy leads to an enhanced outcome with 5-year overall survival of 80–90%. High-throughput DNA and RNA profiling technologies might help to overcome the inability to predict responders. Since most MIBC patients undergo NAC followed by cystectomy, pre-treatment tumor biopsy and post-chemotherapy cystectomy specimens are clinically available, creating an ideal setting to study the genomic and transcriptomic effects of NAC. Here we present RNA sequencing of a cohort of 113 MIBC patients treated with NAC from different hospitals. For each patient, FFPE pre (n=71) and post-treatment (n=29) samples were obtained from biopsy and cystectomy respectively. Response (n=58) was defined as downstaging to non-MIBC (< 0.05) upregulated in NR before treatment, associated with cancer growth and worse prognosis. On the other hand, R showed upregulated pathways related to the cell cycle. Interestingly, no differences were observed in immune cell proportions between the two groups. However, in the WGCNA, we identified a gene group negatively correlated with response, linked to crucial signaling pathways such as Wnt signaling and cell proliferation. WNT signature was obtained through performing the intersection between DE genes and genes related to several WNT pathways. This group of genes shows a significant correlation between low expression of those genes and overall survival, as well as response to NAC, in MIBC patients.

2

SINGLE-CELL RNA SEQUENCING ANALYSIS REVEALS CHALLENGES IN IDENTIFYING SENESCENT AND TUMOR CELLS IN A BREAST CANCER MOUSE MODEL

Irene Agustí-Barea[1], Marta Lalinde-Gutiérrez[2], Joaquín Arribas[2], Lara Nonell[1]

1. VHIOinformatics Unit, VHIO Vall d'Hebron Institute of Oncology, Carrer Natzaret 115-117, 08035 Barcelona, Catalonia, Spain
2. Growth Factors Group, VHIO Vall d'Hebron Institute of Oncology, Carrer Natzaret 115-117, 08035 Barcelona, Catalonia, Spain
Presenter e-mail: ireneagusti@vhio.net

Single-cell transcriptomics analysis (scRNA-seq) is widely used to characterize normal cell types in the tumor microenvironment (TME) as well as to understand the expression of tumor cells in many cancer types. Cellular senescence has a biological interest for its involvement in inhibiting tumor progression, but its characterization in a tumor environment remains a challenge. In this study, a breast cancer mouse model presenting oncogene-induced senescence (OIS) was treated to induce apoptosis of senescent cells at different stages of

tumor evolution, in order to study the tumorigenic effect of this type of cells. With scRNA-seq, we explore the difference in the cell populations across time and treatment status, predict tumor cells from their copy number aberration (CNA) profile using CopyKAT, and aim to identify the senescent cells through the expression of senescence-associated secretory phenotype (SASP) markers. We successfully detected tumor cells in the epithelial cell compartment for samples with enough amount of immune and putative tumor cells. However, difficulties raised when samples had a lower number of immune cells. In addition, the identification of senescent cells was also challenging due to the complexity of senescent cells, which are characterized by high stress levels, increased expression of mitochondrial markers, physical fragility, and an overall high heterogeneity. We suggest that the complex senescent phenotype could drive their loss during the single-cell isolation and confuse them with poor quality cells, hindering their detection and the prediction of tumor cells.

3

POPLIFE: A POPULATION GENOMICS BROWSER ACROSS THE TREE OF LIFE

Cristina Amor-Jiménez1,2 , Alejandro Arangua1,2, Adrià Mompart1,2, Sònia Casillas1,2, Antonio Barbadilla1,2 1. Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona 2. Bioinformatics of Genomic Diversity (BGD). Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona Presenter e-mail: cristina.amor@uab.cat

Next-generation sequencing technologies have led to a growing number of high-quality genomes, not only for humans and model organisms but also for natural species. In the face of the current biodiversity loss crisis, population genomics plays a crucial role in characterising genetic diversity among natural populations, offering valuable insights into their evolutionary history and providing powerful tools for their conservation. Although numerous genome browsers offering comprehensive data on a vast number of species are currently available, they rarely address population genomics information. To fill this gap and leverage the vast amount of data being generated, we are developing PopLife, a new genomic tool that aims to facilitate the analysis of genetic variation within and between populations across any species in the Tree of Life. PopLife provides a highly performant pipeline specially designed to compute population genetics statistics for large-scale genomic datasets, such as variation and divergence metrics, linkage disequilibrium parameters, and neutrality tests. Its novel user-friendly genome browser interface allows the interactive visualisation and retrieval of the data. The PopLife pipeline is built using the Python programming language and the scikit-allel package, while the genome browser for data visualisation is based on the open-source platform JBrowse2, and will be soon available at http://poplife.uab.cat

4

A BENCHMARK OF TISSUE DECONVOLUTION SOFTWARE FOR PLASMA CELL-FREE RNA MIXTURES

Giovanni Asole[1], João Curado[1] and Julien Lagarde[1]

1. Flomics Biotech SL, Carrer Pujades 94-96, 08005, Barcelona, Spain.
Presenter e-mail: giovanni.asole@flomics.com

Blood plasma carries cell-free RNA (cfRNA) released by cells from all the organs it irrigates, making the circulating transcriptome an important source of molecular biomarkers. The deconvolution of plasma cfRNA into its constituent tissues of origin (TOO) can offer valuable insights into an organism's health, making it a powerful method for early disease detection and classification, including cancer. In recent years, numerous efforts have been directed toward the development of increasingly precise deconvolution software. In addition, studies

have been carried out to test those tools through the use of simulated pseudo-mixtures. However, published benchmarking studies on deconvolution software are based on relatively simple tissue or cell mixtures, which therefore do not reflect the complexity present in plasma cfRNA. In this study, we introduce a novel computational simulation framework that enables better control of the complexity of pseudo-mixtures and its application for a benchmark study of deconvolution tools. Our results quantify how the number of tissues taken into consideration and their proportions affect the complexity of the pseudo-mixture, influencing the performance of deconvolution and testing their application on a real-world scenario such as plasma cfRNA-Seq. Our benchmark overall confirms results obtained in previous studies that used less complex pseudo-mixtures, while pointing out important differences. In summary, our study contributes valuable insights and sets the stage for further refinement of methodologies, ultimately advancing our understanding of cfRNA's diagnostic potential.

5

NOVEL INVERSION IDENTIFIED BY OPTICAL GENE MAPPING IN DIZYGOTIC TWINS WITH SYNDROMIC IMMUNODEFICIENCY

Laura Batlle-Masó1,2,3,4, Andrea Martín-Nalda1,2,3, Clara Franco-Jarava3,5,6 , Jacques G. Rivière1,2,3, Marina Garcia-Prat1,2,3, Alba Parra-Martínez1,2,3, Aina AguilóCucurull3,5,6, Mónica Martinez-Gallo3,5,6, Kornelia Neveling7 , Alexander Hoischen7,8, Pere Soler-Palacín1,2,3, Roger Colobran3,5,6,9,10 1 Infection and Immunity Research Group, Vall d'Hebron Research Institute (VHIR), Vall d'Hebron Barcelona Hospital Campus, Barcelona, Catalonia, Spain. 2Pediatric Infectious Diseases and Immunodeficiencies Unit, Vall d'Hebron Children's Hospital (HUVH), Vall d'Hebron Barcelona Hospital Campus, Barcelona, Catalonia, Spain. 3Jeffrey Modell Diagnostic and Research Center for Primary Immunodeficiencies, Barcelona, Catalonia, Spain. 4Pompeu Fabra University (UPF), Barcelona, Catalonia, Spain. 5Translational Immunology Research Group, Vall d'Hebron Research Institute (VHIR), Vall d'Hebron Barcelona Hospital Campus, Barcelona, Catalonia Spain. 6 Immunology Division, Vall d'Hebron University Hospital (HUVH), Vall d'Hebron Barcelona Hospital Campus, Barcelona, Catalonia, Spain. 7Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands. 8Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands; Radboud Institute of Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands; Department of Internal Medicine and Radboud Center for Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, The Netherlands. 9Department of Clinical and Molecular Genetics, Vall d'Hebron University Hospital (HUVH), Vall d'Hebron Barcelona Hospital Campus, Barcelona, Catalonia, Spain. 10Department of Cell Biology, Physiology and Immunology, Autonomous University of Barcelona (UAB), Bellaterra, Catalonia, Spain.

We present a case of two dizygotic twins born from first-degree consanguineous parents in 2004. They debuted at 6 months of age with severe atopic dermatitis. During the first year of life, they presented with repeating bronchitis and pneumonia and severe HSV-1 infections. At 21 months they started monthly immunoglobulin replacement therapy (still ongoing). They do not have any malignancies or Fanconi anaemia. Although sharing some characteristics they also have distinct features, T1 has short stature, severe migraine attacks, self-injury behaviour and vomits, intermittent dysphagia and recurrent lower airway infections. The other (T2) has DM type 1, transmission deafness due to recurrent otitis and also recurrent lower airway infections. At a molecular level, they have high IgE, high eosinophils, low B lymphocytes, low NK cells, low native CD4+ cells and increased T cells. Over the years, many different genetic studies have been performed: targeted sequencing, WES (SolveRD reanalysis), WGS (ongoing), karyotyping, aCGH, and RNAseq (ongoing). All results were not conclusive but recently, optical gene mapping was performed in the context of SolveRD collaboration. This technique allowed us to identify a 14Mb de novo inversion

(chr13:32389924-46105368), previously undetected by karyotyping and array CGH. The variant includes 115 genes of which 4 are associated with inborn errors of immunity (BRCA2, RFXAP, SNORA31, TNFSF11). Using WGS data and targeted Sanger sequencing we identified the breakpoints which are located at intron 24 of BRCA2 and the 5' region of CPB2. The mother (deceased due to oesophagal cancer) also carries the inversion but did not show any immune or developmental alteration. The father is healthy and does not carry the inversion. Given the size and breakpoint location of the inversion, we believe that it may have an impact on the phenotype of the patients. However, the fact that the mother was healthy (except for cancer, which may be related to the disruption of BRCA2) adds another layer of complexity to the case. RNAseq is ongoing to elucidate the effect of the genetic event at an expression level. An extended analysis of the family is being performed. Although more research needs to be done, we discovered an inversion that may be related to the clinical phenotype of these two previously undiagnosed twins. This case illustrates the benefits of collaborative efforts and data-sharing consortia in the way to solve challenging cases.

6

GENOMICS MADE EESSI: FROM HPC TO CLOUD, THE BENEFITS OF A SHARED APPSTACK

Erica Bianco, PhD - HPCNow!

The heterogeneity of the scientific IT infrastructure is a fact. The same data can be analyzed on a local workstation or in different HPC facilities. Nowadays, running analysis in the cloud is becoming more common for testing new tools, getting access to different IT technologies, or for empowering your cluster on demand with cloud bursting techniques. How to be sure the analyses are reproduced under the same conditions if the used tools are not exactly the same? How to save time and get the results sooner, increasing the overall performance and reducing the IT costs? Using containers can solve the portability of the tools, but performance will suffer, and so does the execution time. The solution to get both portability and performance is EESSI. EESSI (pronounced easy) is the European Environment for Scientific Software Installations, a collaboration between different European HPC practitioners, both academic and industry partners, with the common goal of setting up a shared repository of ready-to-use and system agnostic scientific applications. In a few minutes, it is possible to have the same tools in any linux server, from a workstation to a cloud instance passing through the newly installed HPC cluster. Moreover, EESSI takes care to adapt the performance to the CPU architecture of the server, reducing the overall running time, cost and carbon footprint of the analyses. For example, in molecular modeling, using Gromacs optimized version, via EESSI, showed to be 57% faster in simulating the same amount of nanoseconds/day. In summary, EESSI gives any researcher the possibility to analyze genomic (and not only genomic) data without the need to know how to install the necessary tools. For more information about the EESSI project: ● Droge et al., 2023, EESSI: A cross-platform ready-to-use optimised scientific software stack, https://doi.org/10.1002/spe.3075 ● Website: https://www.eessi-hpc.org ● GitHub: https://github.com/EESSI ● Documentation: https://eessi.github.io/docs ● Twitter: https://twitter.com/eessi_hpc

7

EXPLORING THE IMPACT OF PREVIOUS DISEASE TRAJECTORIES ON LONG COVID INCIDENCE

Natalia Blay[1,2], Xavier Farré[1,2], Judith Garcia-Aymerich[3,4,5], Gemma Castaño-Vinyals[3,4,5,6], Manolis Kogevinas[3,4,5,6], Rafael de Cid[1,2]

1. Genomes for Life-GCAT Lab, Germans Trias i Pujol Research Institute (IGTP), Badalona, Spain.
2. Grup de REcerca en Impacte de les Malalties Cròniques i les seves Trajectòries. (GRIMTra), (2021 SGR 01537).
3. ISGlobal, Barcelona, Spain, Barcelona, Spain.
4. Universitat Pompeu Fabra (UPF), Barcelona, Spain.
5. CIBER Epidemiología Y Salud Pública (CIBERESP), Madrid, Spain.
6. IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain.
Presenter e-mail: nblaym@igtp.cat

Background: Long COVID or post COVID-19 condition refers to the persistence of symptoms or sequela after the recovery of SARS-CoV-2 infection, but its cause remains still largely unknown. Identifying its possible causes and establishing an accurate risk profile could be important for prevention, rapid detection and treatment, providing also a better understanding of this new condition. In this study, we determine how previous comorbidities can lead to the development of long COVID, and especially disease trajectories, that are a set of conditions appearing in the same order in a large number of individuals.

Methods: The analysis was performed in 4886 individuals (2929 women, 59.9%) from the GCAT cohort. Disease trajectories were derived using chronic conditions obtained from electronic health records, and long COVID phenotype was determined through a questionnaire (COVICAT) administered in 2023, including individuals having COVID-19 symptoms or sequela for more than 3 months. We used logistic regression models to determine the effect of previous disease trajectories in long COVID incidence, adjusting by age and sex. Stratified analysis by sex was carried out to explore sex-specific trajectories.

Results: We found 9 disease trajectories associated with an increased risk of long COVID diagnosis. These included either mental/neurological (Depression, Anxiety, Migraine, Mononeuritis of upper limb, and Reaction to severe stress), respiratory (Vasomotor and allergic rhinitis, and Asthma) or cardiovascular (Hypertension, Disorders of lipoprotein metabolism and other lipidaemia, and Sleep disorders) diagnosis.

Conclusions: Long COVID incidence is affected by previous comorbidities, especially by disease trajectories included either in the mental/neurological, respiratory or cardiovascular system.

8

NOVEL IL1RN VARIANTS CAUSING THE DEFICIENCY OF INTERLEUKIN-1 RECEPTOR ANTAGONIST

Núria Bonet1 , Elena Urbaneja2 , Manuel Solis-Moruno1 , Anna Mensa-Vilaro3 , Iñaki Ortiz de Landazuri3 , Rocio Lara3 , Susana Plaza3 , Virginia Fabregat3 , Jordi Yagüe3 , Ferran Casals4 , Juan I. Arostegui3 * 1. Genomics Core Facility, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain 2. Department of Immunology and Pediatric Rheumatology, Hospital Clinico Universitario, Valladolid, Spain 3. Immunology Department, Hospital Clínic, Barcelona, Spain 4. Facultat de Biologia, University of Barcelona, Barcelona, Spain Presenter e-mail: nuria.bonet@upf.edu

The heterogenous group of undiagnosed autoinflammatory diseases (AIDs) includes patients and families with suspected, but not genetically confirmed, conditions mainly characterized

by recurrent episodes of sterile inflammation. The aim of this project was to identify the cause of an undiagnosed AID observed in two siblings born prematurely in the 80's from a non-consanguineous healthy couple. We identified two brothers who were afflicted by a chronic disease since the neonatal period. The disease was mainly characterized by fever, generalized vesiculo-pustular lesions, marked pain with movement, and painful swelling at clavicles, ribs, femur, and mandibula. Both patients died during infancy by a complicated varicella infection and a cerebral thrombosis, respectively. We hypothesized they suffered from a monogenic disease inherited as a recessive or X-linked trait. Different genetic studies were performed in the patients' healthy parents due to the absence of samples from patients. By sanger and next-generation sequencing (NGS) methods we identified two novel heterozygous variants at IL1RN, the c.318+2T>G variant in the father, and a ≈2600bp intragenic deletion in the mother. Relative mRNA expression analysis was performed to characterize the functional consequences of the detected variants at IL1RN. IL1RN mRNA production was markedly decreased in both progenitors when compared with healthy subjects, strongly supporting that both variants were loss-of-function. Intrafamilial IL1RN genotypes were compatible with those expected for a recessivelyinherited disease. The disease manifestations and the identification of two novel IL1RN variants predicted to generate truncated proteins in their progenitors strongly suggest that both patients suffered from a lethal form of the deficiency of interleukin-1 receptor antagonist. The main limitation of this study was the non-availability of the patients' samples, which prevented us from establishing unequivocally their definitive diagnosis

9

COMPUTATIONAL IDENTIFICATION AND VALIDATION OF DIFFERENTIAL DRUG SENSITIVITY IN METASTATIC AND PRIMARY CANCER CELL LINES

Maria Butjosa-Espín[1,2], Jose A. Seoane[1]

1. Cancer Computational Biology Group, Vall d'Hebron Institute of Oncology (VHIO), Centro Cellex, Carrer de Natzaret, 115-117, 08035 Barcelona, Catalonia, Spain
2. Universitat Autònoma de Barcelona (UAB), Barcelona, Plaça Cívica, 08193 Bellaterra, Catalonia, Spain
Presenter e-mail: mariabutjosa@vhio.net

Metastasis causes 90% of cancer-related deaths, urging innovative therapy research due to ineffective treatments and therapy resistance. New treatments, often first tested in metastatic patients, lack success in primary tumors. Our aim is to tackle this challenge by taking advantage of distinct genetic patterns in metastases and using databases containing drug response of thousands of drug - cell line pairs.

The PRISM and GDSC2 databases were used to identify differential drug response between metastatic and primary cell lines by employing a logistic regression model. Moreover, a 'drug set enrichment analysis' was conducted to identify enriched mechanisms of actions.

The analysis revealed drugs and drug families with differential effects in metastatic versus primary cell lines, particularly at a pan-cancer level. Specifically, the Akt inhibitors group was enriched in metastatic cell lines in both the PRISM (FDR < 0.01) and GDSC2 (FDR < 0.25) databases. However, when stratifying by cancer types, no significant results were found in GDSC2, limited by its lower statistical power. Further exploration in PRISM did unveil the enrichment of EGF receptor inhibitors in colon adenocarcinoma (COAD) metastatic cell lines (FDR < 0.05), and the significance of epigenetic regulation in lung adenocarcinoma (LUAD).

Our approach offers a promising strategy for identifying optimal drug candidates for specific cancer types in metastatic versus primary diseases. Further exploration and application of our

methodology in cancer subtypes cell line models may uncover additional avenues improving, patient prognosis in a more targeted way.

10

DIAGNOSTIC AND FILTERING OF GENOMIC DNA CONTAMINATION IN RNA-SEQ DATA WITH GDNAX

Beatriz Calvo-Serra 1 (ORCID: 0000-0002-7614-396X), Robert Castelo 1,* (ORCID: 0000-0003-2229-4508) 1Dept. of Medicine and Life Sciences, Universitat Pompeu Fabra *Corresponding author: robert.castelo@upf.edu

Total RNA-sequencing (RNA-seq) is the most unbiased approach to characterize the whole transcriptome, and often the only available choice with degraded samples of clinical or biological interest. Unfortunately, it is also prone to genomic DNA (gDNA) contamination due to the fluctuating efficiency of the gDNA digestion step (i.e. DNase treatment), or the complete lack thereof, specially with low input samples. gDNA contamination introduces several biases in the analysis of gene expression, such as the overestimation of expression levels for transcribed genes, especially those with low expression. Moreover, it can lead to the misattribution of expression to unannotated regions of the genome. Given the significant influence of gDNA contamination in the reliability of RNA-seq results, it is crucial to check contamination levels in the quality control step before performing further analyses. Here we present gDNAx, a Bioconductor R package to quickly diagnose and quantify the presence of gDNA in RNA-seq data. Moreover, it provides functionality to filter out reads of potential gDNA origin, thereby mitigating the impact of gDNA contamination on subsequent analyses. In summary, the gDNAx package facilitates the diagnosis and adjustment of the influence of gDNA contamination in RNA-seq experiments

11

UNRAVELING ALTERNA.VE SPLICING AND CELL COMPARTMENT-SPECIFIC GENE REGULA.ON IN HUMAN LYMPHOBLAS.C LEUKEMIA CELLS USING LONG-READ TRANSCRIPTOMICS

Sílvia Carbonell-Sala, Emilio Palumbo, Tamara Perteghella, Mark Diekhans, Diego Garrido and Roderic Guigó.

Alterna.ve splicing, a crucial post-transcriptional process, enhances proteomic diversity without increasing gene numbers, playing pivotal roles in gene regulation, specialization, and disease. Limitations in short-read sequencing have led to the exploration of long-read sequencing technologies, such as Oxford Nanopore Technologies. This study utilizes second and third-generation sequencing to investigate alterna.ve splicing in the nucleus and cytosol of the human lymphoblastic leukemia B (BLaER1) cell line. The analyzed long-read Oxford Nanopore data elucidates RNA localization and splicing regulation. Shortread Illumina RNA-Seq is employed in building long-read transcript models and complements this analysis for precise junction validation. Filters, including transcripts that are part of annotated genes, were applied to ensure data quality. The ggsashimi splicing junction visualization tool was adapted for long-read isoform and splicing event assessment. Specific genes (ZDHHC6, MTG1, LAP3) reveal compartmental variations, emphasizing the importance of gene regulation in specific cellular compartments. Enriched Gene Ontology terms indicate cytosol-nucleus differences in RNA processing, translation, and ribosome biogenesis. These findings contribute

to a thorough understanding of alterna.ve splicing, RNA processing, and gene regulation, with significant implications for the development of therapeutic strategies targe.ng splicing-related diseases

12
PROTEOPARC: A PIPELINE TO GENERATE DATABASES FOR ANALYZING ANCIENT PROTEINS

Guillermo Carrillo-Martín1, Johanna Krueger1, Tomas Marques-Bonet1,2,3,4, Esther Lizano2 1. Institute of Evolutionary Biology (UPF-CSIC), Dr. Aiguader 88, Barcelona, Spain 2. Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Edifici ICTA-ICP, Barcelona, Spain 3. Catalan Institution of Research and Advanced Studies (ICREA), Passeig de Lluís Companys, Barcelona, Spain 4. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain Presenter e-mail: guillermo.carrillo@upf.edu

Over the past few years, there has been an increasing interest in paleoproteomics, a discipline focused on studying ancient proteins found in biological remains. Therefore, there is an absence of bioinformatic tools to automatize essential processes that are currently manually performed. Here, we present ProteoParc, a pipeline to generate protein databases to be included in a mass spectrometry peptide sequencing workflow. To test this software, we generated a proboscidean enamel database in order to sequence ancient protein fragments from 7 tooth Deinotherium samples, an extinct Proboscidea species that inhabited Catalonia 10 million years ago. We were able to identify enamel-associated peptides in all the analyzed samples, also finding some single amino acid polymorphism compared to nowadays elephants. Thus, ProteoParc has been proven to be useful in paleoproteomics analysis, creating a new bioinformatic tool for future studies

13
ALTERNATIVE SPLICING VARIABILITY ACROSS INDIVIDUALS AT SINGLE-CELL RESOLUTION

Rubén Chazarra Gil[1], Marta Melé Messeguer[1], Martin Hemberg[2]

1. Barcelona Supercomputing Center, Plaça d'Eusebi Güell, 1-3, 08034 Barcelona, Spain
2. Harvard Medical School, 25 Shattuck St, Boston, MA 02115, Estats Units d'Amèrica

Presenter e-mail: ruben.chazarra@bsc.es

Alternative splicing expands transcriptome and protein diversity by generating multiple mRNA transcripts from a single gene. Alternative splicing has been extensively studied in bulk RNA-seq data, revealing its critical role in development, differentiation, and disease. Moreover, it exhibits significant inter-individual variability. Nevertheless, at the bulk RNA-seq level, the cellular heterogeneity of splicing can be confounded by differential cell type composition, potentially masking cell type specific patterns.

With the advent of large-scale single-cell transcriptomic datasets it is now possible to assess variation in alternative splicing patterns among individuals at single-cell resolution. We query changes in the relative abundance of gene transcripts between individuals of distinct genetic ancestries with a dataset of peripheral blood mononuclear cells under influenza infection. We observe that differences in transcript usage between populations are pervasive across blood

cell types and are consistent across stimulation contexts, highlighting the genetic basis of these population variances. Furthermore we highlight the limitations associated with using droplet-based scRNA-seq data for studying alternative splicing.

In summary, our study provides insights into immune-related splicing differences between different human populations, shedding light on both the advantages and limitations of characterizing alternative splicing in scRNAseq data

14

A DISTINCT CLASS OF PAN-CANCER SUSCEPTIBILITY GENES REVEALED BY ALTERNATIVE POLYADENYLATION TRANSCRITOME-WIDE ASSOCIATION STUDY

Hui Chen[1]#, Zeyang Wang[1]#, Qixuan Wang[1]#, Wenyan Chen[1], Jia Wang[2], Xuelian Ma[1], RuoFan Ding[1], Lihai Gong[1], Xing Li[1], Xudong Zou[1], Mireya Plass[3,4], Cheng Lian[5], Ting Ni[6], Gong-Hong Wei[5,7], Wei Li[8]*, Lin Deng[2]*, Lei Li[1]*

1. Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518055, China
2. Institute of Molecular Physiology, Shenzhen Bay Laboratory, Shenzhen 518055, China
3. Gene Regulation of Cell Identity Group, Regenerative Medicine Program, Bellvitge Institute for Biomedical Research (IDIBELL), and Program for Advancing Clinical Translation of Regenerative Medicine of Catalonia, P-CMR[C], L'Hospitalet de Llobregat, Barcelona, 08908, Spain
4. Center for Networked Biomedical Research on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, 28029, Spain
5. Fudan University Shanghai Cancer Center & MOE Key Laboratory of Metabolism and Molecular Medicine and Department of Biochemistry and Molecular Biology of School of Basic Medical Sciences, Shanghai Medical College of Fudan University, Shanghai, 200032, China
6. State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, Human Phenome Institute, School of Life Sciences and Huashan Hospital, Fudan University, Shanghai 200438, China
7. Disease Networks Research Unit, Faculty of Biochemistry and Molecular Medicine & Biocenter Oulu, University of Oulu, Oulu, 90410, Finland
8. Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, The University of California, Irvine, CA 92697, USA
Presenter email: hchen@idibell.cat

Alternative polyadenylation (APA) plays an important role in cancer initiation and progression; however, current transcriptome-wide association studies (TWAS) mostly ignore APA when identifying putative cancer susceptibility genes. Here, we performed a pan-cancer 3′ untranslated region (UTR) APA TWAS (3′aTWAS) analysis by integrating 55 well-powered (n>50,000) GWAS datasets across 22 major cancer types with APA quantification from 23,955 RNA sequencing samples across 7,574 individuals. We found that genetic variants associated with APA are co-localized with 28.57% of cancer GWAS loci and contribute a significant portion of cancer heritability. We further identified 642 significant cancer susceptibility genes predicted to modulate cancer risk via APA, from which 62.46% of which have been overlooked by traditional expression- and splicing- studies. As proof of principle validation, we show that alternative alleles of 3′aQTL facilitate 3′UTR lengthening of *CRLS1* gene leading to increased protein abundance and promoted proliferation of breast cancer cells. Together, our study highlights the significant role of APA in discovering new cancer susceptibility genes and provides a strong foundational framework for enhancing our understanding of the etiology underlying human cancers.

15

GENETIC-DRIVEN SPLICING OF RIBOSOMAL PROTEINS BETWEEN AFRICAN AND EUROPEAN ANCESTRIES

Pau Clavell-Revelles 1 , Marta Huertas 1 , Winona Oliveros 1 , Paula Iborra 1 , Raquel García-Pérez 1 , Marta Melé 1* 1. Life Sciences Department, Barcelona Supercomputing Center (BSC), Pl. Eusebi Güell 1-3, 08034 Barcelona, Spain First author: pau.clavell@bsc.es *Corresponding author: marta.mele@bsc.es

Alternative splicing (AS) is a fundamental source of transcriptome diversity by generating different mRNA isoforms from a single gene, being putatively relevant in diseases such as cancer and neurodegenerative disorders. Analysis of the Genotype-Tissue Expression (GTEx) dataset has shown that splicing is more variable between individuals than between tissues (Melé et al., 2015), with the largest inter-individual variation occurring at ribosomal proteins. This inter-individual variability is mainly driven by genetic differences between human ancestries, with most AS differences being shared across tissues (García-Pérez et al., 2023). Firstly, to further characterize differential splicing in ribosomal proteins between human populations, we perform differential gene expression (DGE), differential transcript expression (DTE) and differential transcript usage (DTU) analyses between individuals of African-American and European-American ancestries from the GTEx dataset. We find that more than 20 tissues have genes showing significant differential transcript usage enriched in ribosomal proteins. However, only four and one tissues have differentially expressed genes and differentially expressed transcripts, respectively. In addition, we find that the genes with DTU in at least 5 tissues are enriched in ribosomal-related pathways. Secondly, we leverage the 1000 Genomes Projects genotype data to compute fixation index (FST) and cross-population extended haplotype homozygosity (XP-EHH) scores at the single nucleotide polymorphism level. Then we develop a permutation approach to test if GTEx splicing quantitative trait loci (sQTL) associated to ribosomal proteins have been positively selected either in European or African (Yoruba) populations. In summary, we investigate the differences at the expression level of ribosomal proteins between human populations and we establish a framework to assess their potential evolutionary origin in local adaptation

16

GENOMIC ANALYSIS OF WILD BONOBO POPULATIONS USING NON-INVASIVE FECAL SAMPLES

Mar Crego, PhD student at the Comparative Genomics Lab, Universitat Pompeu Fabra. Presenter correu electrònic:mar.crego@upf.edu

Bonobos, sp. Pan paniscus, are one of our closest relatives and an endangered species (Estrada et al. 2019; Prüfer et al. 2012).

This primate species is found only in the Democratic Republic of Congo (Hickey et al. 2013). They are mostly frugivorous, however they are known to consume occasionally insects, fish and small mammals and they usually live in large groups (Furuichi 2009). The most striking feature of bonobos is their behavioural patterns which have often been pointed out as vital for understanding behavioural changes from a genetic perspective as they are closely related to humans and chimpanzees but have extremely different behaviours to the latter (Gruber & Clay 2016). Their social structure is female dominated, with less territorial and aggressive

behaviours than chimpanzees (Gruber & Clay 2016). Female alliances dominate mating strategies and food allocated, and are maintained by genital rubbing which reduces social tensions (Gruber & Clay 2016, Lacambra et al. 2005; Fruth et al. 2013; Reinartz et al. 2013). However, their location has made it historically difficult to study them as a species as they are found only in an area which has been in the past years spewed in civil unrest, which in turn has accelerated practices such as poaching and deforestation, thus increasing pressure on the species itself (Ilunga Kalenga et al. 2019; Hoffmann et al. 2020; Molinario et al. 2020; Ogunnoiki 2019). As such, previous studies that have aimed to research the genetic population structure of bonobos have had small and fragmented datasets (Medkour et al. 2021). The use of non-invasive fecal samples could therefore be a useful tool in obtaining DNA samples that are easier to collect and less damaging to individuals (Fontsere 2020). This project aims to improve understanding of the genetic evolution of Bonobos, new species conservation perspectives and unearth clues to our own evolutionary path. Our methods include use of target capture analysis of faecal samples and bioinformatic analysis in collaboration with the Evolutionary Genetics Group at the Universität Zurich, the Laboratoire d'Éco-anthropologie de l'Université de Paris, and the Hahn Lab Group at the University of Pennsylvania

17

SYSTEMS BIOLOGY PREDICTION OF SMALL MOLECULE PHARMACOLOGICAL PROPERTIES USING BIOACTIVITY DESCRIPTORS

Dylan Dalton1,* ,Patrick Aloy1,2 Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain 2 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain.

Compound bioactivity signatures allow the description of small molecules according to the biological effects that they exert, providing complementary opportunities to current drug discovery strategies. The Chemical Checker (CC) resource provides a rich collection of bioactivity signatures and Signaturizers allow signature inference for any given compounds. We here leverage both resources to build shallow and deep machine learning (ML) models for pharmacological property prediction. Specifically models have been built for mechanism of action, target, cell line, therapeutic areas, indications and side effects. For mechanism of action, metabolic gene, protein binding, cancer cell line, therapeutic areas, indications and side effect bioassay data. Results:. We present a novel resource, which feeds from CC and Signaturizers, with a large number of pre-trained stored signature-activity relationship (SigAR) models for a range of descriptors and ML algorithms. CC bioactivity descriptors are shown to outperform classical chemistry based descriptors in model training for most pharmacological predictions. The potential of this approach is shown by annotating the IRB Drug Screening Platform Library. Availability: All code publicly available at https://github.com/00dylan00. Contact: dylan.dalton@irbbarcelona.org Supplementary information: Supplementary Figures attached at the end of this document

18
EXPLORING THE INTER-SPECIES NON-HUMAN PRIMATE GUT MICROBIOME

Carles Domingo-Costa1 , Samuel Piquer-Esteban1,2, Wladimiro Diaz-Villanueva1,3 , Vicente Arnau1,3, Andrés Moya1,2,3 1. Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain 2. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region (FISABIO), Valencia, Spain 3. Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública(CIBEResp), Madrid, Spain Presenter e-mail: cardocos@alumni.uv.es

The human gut occupies a special place in studying different microbial environments, as the gut microbiota is linked to host health. In previous work, we suggested the existence of a universal microbial core in humans that could be a candidate for a hypothetical phylogenetic core of mutualistic microorganisms that coevolve with the human species (Piquer-Esteban et al., 2021). However, a clear relationship between the type of life and the microbiome composition at the genus level was also observed, highlighting the importance of the ecological environment. In this context, some authors have utilized the non-human primate (NHP) microbiome as a valuable method to better understand the functional role of human microbiome. That is applied not only to identify similarities in the coevolution of humans with their microbial communities (Amato et al., 2019), but also to trace environmental factors associated with adapting the human microbiome to diverse dietary niches (Sharma et al., 2020). This study presents a first meta-study results that compiles various NGS datasets of the gut microbiome from several NHPs. References: 1. Piquer-Esteban, S., Ruiz-Ruiz, S., Arnau, V., Diaz, W., & Moya, A. Exploring the universal healthy human gut microbiota around the World. Computational and structural biotechnology journal, 20, 421– 433 (2021). https://doi.org/10.1016/j.csbj.2021.12.035 2. Amato, K.R., Mallott, E.K., McDonald, D. et al. Convergence of human and Old World monkey gut microbiomes demonstrates the importance of human ecology over phylogeny. Genome Biol 20, 201 (2019). https://doi.org/10.1186/s13059-019-1807-z 3. Sharma, Ashok K et al. Traditional Human Populations and Nonhuman Primates Show Parallel Gut Microbiome Adaptations to Analogous Ecological Conditions. mSystems vol. 5, 6 e00815-20. (2020). https://doi.org/10.1128/mSystems.00815-2

19

ENVIRONMENT AND GENETICS ON DEPRESSION SYMPTOMS: AN EXPOSOME APPROACH DURING THE LOCKDOWN OF THE COVID-19 OUTBREAK

X. Farre[1,2], N. Blay[1,2], A. Espinosa[3,4,5,6], G. Castaño-Vinyals[3,4,5,6], A. Carreras[1], J. Garcia-Aymerich[3,5,6], E. Cardis[3,4,5,6], M. Kogevinas[3,4,5,6] , and X. Goldberg[3,4,7](*), R. de Cid[1,2] (*)

1. Genomes for Life-GCAT lab., Germans Trias i Pujol Research Institute (IGTP), Camí de les Escoles, s/n, 08916, Badalona, Spain
2. Research Group on the Impact of Chronic Diseases and their Trajectories (GRIMTra), Germans Trias i Pujol Research Institute (IGTP), Camí de les Escoles, s/n, 08916, Badalona, Spain
3. ISGlobal, Dr. Aiguader, 88, 08003 Barcelona, Spain
4. IMIM (Hospital del Mar Medical Research Institute), Dr. Aiguader, 88, 08003 Barcelona, Spain
5. Universitat Pompeu Fabra (UPF), C. de Ramon Trias Fargas, 25, 27, 08005, Barcelona, Spain
6. CIBER Epidemiología y Salud Pública (CIBERESP), Av. Monforte de lemos, 3-5, 28028, Madrid, Spain
7. CIBER Salud Mental (CIBERSAM), Madrid, Spain

(*) Last senior authors.
Presenter e-mail: xfarrer@igtp.cat

Risk of depression increased in the general population after the COVID-19 pandemic outbreak. By examining the interplay between genetics and the exposome during COVID-19 lockdown, we can gain an insight as to why some individuals are more vulnerable to depression while others are more resilient. This study, conducted on a cohort of 9,218 individuals (COVICAT), includes a comprehensive non-genetic risks analysis for depression, complemented by genomics analysis in a subset of 2,442 participants. Depression levels were evaluated using the Hospital Anxiety and Depression Scale (HADS). Together with Polygenic Risk Scores (PRS), we introduced a novel score, Poly-Environmental Risk Scores (PERS) for non-genetic risks to estimate the effect of each cumulative score and gene-environment interaction. The findings revealed a robust association between environmental scores and moderate depression symptoms and a modest genetic effect (PRS) without evidence of interaction. Social factors (PERSSoc) and life factors (PERSLife) demonstrated the most significant impact, alongside broader environmental and health-related factors (PERSEnv). In summary, actionable elements within the social, behavioral, and environmental domains emerged as the primary drivers of depression risk in this population, independent of genetic and clinical factors.

20

16S rRNA GENE METABARCODING TO UNCOVER MICROBIOTA FROM THE PORK PRODUCTION CHAIN

Núria Ferrer-Bustins[1], Belén Martín[1], Sara Bover-Cid[1], Anna Jofré[1]

1. IRTA, Food Safety and Functionality Program, Finca Camps i Armet s/n, 17121 Monells, Spain

Presenter e-mail: nuria.ferrer@irta.cat

Food-processing surfaces and animal microbiota contribute to the complex microbiota of meat ecosystems. The evaluation of microbial diversity, identification, and correlation of microbial populations of pork and cutting plant surfaces, in different companies (A, B and C), through culture-dependent and independent methodologies was the aim of the present study. Samples (N=108) were analysed by enumeration through viable plate count and 16S rRNA gene metabarcoding (MiSeq, Illumina). Facility surfaces (carcass partition saws, carcass cutting saw, conveyor belts, trays, and drains) and two type of pork products (ribs, bellies and/or loin) for each company were collected in two occasions, per duplicate. Microbiological analysis of facility surfaces was quickly performed after sample collection whereas meat microbiota samples were monitored during product shelf-life. Results showed that microbial diversity of pork was specific from each company ($p<0.05$). Major genera in pork products were *Carnobacterium*, *Serratia*, *Lactobacillus* and *Aeromonas* and in surfaces were *Acinetobacter* and *Moraxella*. 16S rRNA gene metabarcoding allowed a more exhaustive microbiota characterization than the classic plate count and showed that the company factor is the key element when determining the microbial community's composition. Evidence for cross-contamination between surfaces and pork is shown together with a progressive decrease in the microbiota diversity along storage. Strict cleaning and disinfection practices are necessary to minimize spoilage microorganisms' contamination in pork and extend its shelf-life.

21

ADAPTIVE ARCHAIC INTROGRESSION RELATED TO CELLULAR ZINC HOMEOSTASIS IN HUMANS

Ana Roca-Umbert1 , Jorge Garcia-Calleja1, Marina Vogel-González2 , Alejandro FierroVillegas2 , Anja Bosnjak2 , Gerard Ill-Raga2 , Víctor Herrera-Fernández2 , Gerard Muntané1,3,4, Felix Campelo5 , Rubén Vicente2&* , Elena Bosch1,4&* 1 Institut de Biologia Evolutiva (UPF-CSIC), Departament de Medicina i Ciències de la Vida, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Doctor Aiguader 88, 08003 Barcelona, Spain 2 Laboratory of Molecular Physiology, Department of Medicine and Life Sciences (MELIS), Universitat Pompeu Fabra, 08003 Barcelona, Spain 3 Hospital Universitari Institut Pere Mata, IISPV, Universitat Rovira i Virgili, 43206 Reus, Spain. 4 Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), 43206 Reus, Spain 5 ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860 Barcelona, Spain

The SLC30A9 human gene encodes a ubiquitously intracellular zinc transporter (ZnT9). Several genomewide scans of selection have consistently identified strong adaptive signals on SLC30A9 in East Asian populations and suggested rs1047626 as the candidate variant under selection. This SNP causes a methionine to valine substitution at ZnT9 codon 50 and has been associated with susceptibility to neuropsychiatric disorders. To understand its putative adaptive molecular phenotype, we overexpressed ZnT9 in HEK293 cells and explored whether the ZnT9 50Met and ZnT9 50Val variants induced differential zinc transport between the cytosol,

mitochondria, and the endoplasmic reticulum. Our results demonstrate intracellular differences in zinc handling with an impact on mitochondrial metabolism. We also found evidence for directional selection operating in two major complementary haplotypes carrying the ancestral and derived variant of this substitution, respectively, in Africa and East Asia. Furthermore, inspection of the Denisovan and Neandertal genomes revealed an unusually high sharing of derived alleles between the major haplotype outside Africa and archaic humans, and the presence of the ZnT9 50Val variant in the Denisovan genome. Considering the recombination rate of the SLC30A9 region, the persistence of such allelic configuration along a 70.6 kb segment cannot be explained by incomplete lineage sorting. Given the role of the mitochondria in skeletal muscle and brain, and that of zinc in glutamatergic neurotransmission, we propose that archaic adaptation to cold may have driven this selection event of adaptive introgression, while also impacting susceptibility to neuropsychiatric disorders in modern humans.

22

DISCOVERING THE GREY MATTER OF OUR GENOME EXPRESSION DIFFERENCES IN rRNA ALLELES ACROSS TISSUES OF THE COVID-19 OUTBREAK

Raquel García Pérez

More than a decade ago, RNA-Seq revolutionized the field of transcriptomics allowing the catalog and quantification of all transcripts in a cell. Since then, many studies have profiled patterns of gene expression across tissues, species and developmental stages and have associated expression changes with different phenotypic outcomes for most species of RNAs, but not ribosomal RNAs (rRNA). rRNAs are structural and functional components of the ribosomes, essential for protein synthesis. Although rRNAs make up about 80% of cellular RNA, they have been routinely excluded from transcriptomic studies. Their neglect stems from the long-standing conception that ribosomes are homogeneous cellular machines. However, in recent years, numerous studies in non-human species have evidenced the existence of ribosome heterogeneity and a few have shown functional differences between distinct subpopulations of ribosomes.

A major roadblock in the study of human rRNA expression variation has been the lack of a reference sequence for the hundreds of copies of ribosomal genes (rDNA) each individual harbors. Ribosomal genes lie in one of the most complex regions of the human genome, the centromeric regions of the acrocentric chromosomes, which were only recently assembled upon the completion of the first full human genome sequence. Importantly, the different rDNA copies carry sequence variants and thus, the incorporation into the ribosomes of different rRNA variants could generate different types of ribosomes that may vary across tissues, individuals and physiological conditions. Yet patterns of rRNA expression variation in humans remain unknown.

This project aims to comprehensively characterize human rRNA expression variation. I hypothesize that differences in rRNA expression may result in heterogeneous populations of ribosomes.

I will first develop computational pipelines to identify rDNA alleles and quantify their relative expressions from short-read sequence data. I will then reuse WGS and RNA-Seq data from the Genotype Tissue Expression (GTEx) project to 1) create an atlas of human rDNA alleles and 2) study how the expression of rRNA alleles changes between tissues, individuals and different

human phenotypes such as age. I will further investigate the functional consequences of differential rRNA allele expression.
This project will reveal today's unexplored variation in the expression of ribosomal genes and pave the way for future studies analyzing rRNA expression using short sequencing data. The insights derived from this study have the potential to completely transform our understanding of human ribosome composition by revealing the contribution of rRNA expression differences to inter-tissue and intra- and inter-individual variation. Importantly, the computational tools developed could later be applied in any large-scale transcriptomic datasets allowing the study of rRNA expression variation in any disease context

23

DEEPLEARNING FOR THE PREDICTION OF SNP EFFECTS ON TRANSCRIPTION FACTOR BINDING

Patrick Gohl[1], Baldomero Oliva[1]

1. Department of Medicine and Life Sciences, (SBI-GRIB), Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain.
Presenter e-mail: patrick.gohl@upf.edu

We trained a Deep learning model to predict the effects of Single Nucleotide Polymorphisms (SNP) on transcription factor binding. Allele specific binding (ASB) data from ChIP-seq experiments were paired with high sequence-identity motifs assessed in Protein Binding Microarray experiments. For each transcription factor a paired motif was selected from which we derived E-score profiles for reference and alternate DNA sequences of ASB events. A Convolutional Neural Network was trained to predict whether these profiles were indicative of ASB gain/loss or no change in binding. More than 23,000 E-score profiles from various transcription factors were split into train, validation and test data. To further test the generalizability of the Model, we classified data from 4 previously unseen transcription factors with both the trained model and state of the art tools. We compare the performance of the trained model to other available platforms for predicting the effect of SNP on transcription factor binding. To dissociate potential off-site events from ASB data we trained a second model on a subset of ASB data where changes in Position Weight Matrix scoring along the SNP supported observed changes in transcription factor binding (Concordance) and compared performance between our model and other available tools. As a result we saw an improvement in our model and it outperformed state of the art tools in all metrics. Introduction of a post-hoc filter further increased accuracy while reducing coverage of predictions.

24
SEX DISPARITY IN WEIGHT RESPONSE TO COCA-COLA CONSUMPTION

Carlos Golbano[1], Dafne Porcel Sachis[1], Paulina Belvončíková[2], Katarína Kmeťová[2], Jozef Čonka[2], Vicente Arnau[1,3], Peter Celec[2], Roman Gardlík[2], Mária Džunková[1]

1. Institute for Integrative Systems Biology, University of Valencia and Consejo Superior de Investigaciones Científicas (CSIC), 46980 Valencia, Spain.
2. Institute of Molecular BioMedicine, Faculty of Medicine, Comenius University, Bratislava, Slovakia
3. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region(FISABIO), Valencia, Spain.
   Presenter e-mail: cgolbano2000@gmail.com

Excessive consumption of sugary drinks, such as Coca-Cola, has been linked to weight gain and an increased risk of obesity. Nevertheless, it was also shown that individual responses to cola beverages consumption can vary, depending on individual metabolism. To investigate the factors contributing to weight gain, our focus centered on sex differences. Sex hormones, including estrogen and testosterone, play crucial roles in regulating body composition and weight. Furthermore, numerous previous studies have revealed sex-dependent variations in gut microbiome composition, a well-known factor in the regulation of body weight. In accordance with this, our experiments with mice drinking Coca-Cola and controls drinking water showed that Coca-Cola-consuming males surprisingly decreased their weight compared to controls after 19 weeks ($p<0.001$), while the females did not. In order to investigate whether the sex disparity in weight gain is reflected in gut microbiome composition, we sequenced 16S rDNA amplicons from 49 experimental mice (13 cola females, 14 water females, 11 cola males, 11 water males) on Illumina platform. The quality of paired-ends sequences was checked using FastQC and MultiQC. Primer and adapter sequences were removed by Fastp and Cutadapt respectively. Pipeline DADA2 was used for sequence ends trimming to remove low quality ends, as well as, for amplicon sequence variants (ASVs) generation and chimera removal. Afterwards, 16S gene sequences were classified by the IDTAXA classifier of the package with a 50% confidence using Silva and GTDB. ASV count data underwent a filtering process to eliminate sequencing noise and normalization was achieved through the Hellinger transformation. Multivariate analysis techniques, specifically Principal Components Analysis (PCA) and Redundancy Analysis (RDA), were employed using the vegan R package. The results indicated that the microbiome of males and females differed significantly ($r2=0.44$, $p<0.001$). The microbiome composition of males that consumed Coca-Cola differed significantly from the male controls, while females showed only small differences between the Coca-Cola and water groups (four groups comparison $r2=0.87$, $p<0.001$). Further research is needed to explore whether the differences in microbiome composition observed in males are a consequence of different responses of male hormones to caffeine in cola beverages.

Funding: Roman Gardlik: APVV-21-0370 and VEGA 1/0649/21 projects, Mária Džunková: CDEIGENT/2021/008

25

EPIGENETICS OF HUMAN B-CELL PRECURSORS TRANSDIFFERENTIATION INTO MACROPHAGES

Silvia González-López 1,2, Marina Ruiz-Romero 1, Sílvia Pérez-Lluch 1, Roderic Guigó 1,2 1. Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona (BIST), Catalonia, Spain 2. Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain Presenter e-mail: silvia.gonzalez@crg.eu

The role of histone post-translational modifications (hPTMs) has been historically associated with transcription regulation, where the latter derives from the former. However, recent studies have challenged this view, pointing towards a relationship not merely causal, but more complex. Hence, whether chromatin modifications have a truly deterministic role in gene expression regulation or not remains ambiguous. Here, we studied the dynamics of various hPTMs at twelve time points during the transdifferentiation process from human B-cell precursors to macrophages, spanning seven days. For this analysis, we developed two complementary approaches. First, we selected ENCODE's candidate cis-regulatory elements (cCREs) that overlapped ChIP-seq peaks for H3K4me3 and/or H3K27ac in at least one time point of our process, and classified them over time in few chromatin states. We observed high state stability on elements proximal to transcription start sites compared to distal ones; furthermore, we studied extensive genomic loci displaying coordinated dynamic chromatin patterns over transdifferentiation, which showed a likely biological relevance. Additionally, we defined regulatory regions putatively relevant to this transdifferentiation model by selecting regions with H3K4me2, H3K27ac, H3K4me3 and/or H3K27me3 peaks in at least one time point of the process. ChIP-seq signal profiles on these regions were used to identify and classify dynamic loci into different categories according to the combination of hPTMs exhibiting changes along the process. Active marks tend to correlate with chromatin accessibility, whereas the opposite is observed for H3K27me3. Regions showing stable active marks show higher conservation scores than dynamic loci, indicating that loci with lower selective pressure may have distinct cell type-specific regulatory roles. In conclusion, we have characterized regions with reported regulatory roles along human pre-B cell transdifferentiation and evaluated the characteristics of those manifesting dynamic behaviors to venture keys to their identification and relevance.

26

UNDERSTANDING THE ROLE OF THE RNA-BINDING PROTEIN STAUFEN 2 DURING NEUROGENESIS USING SINGLE CELL TRANSCRIPTOMICS

Akshay J. Ganesh[1,2], Sandra María Fernández Moya[1,2], Ana Gutiérrez-Franco[1,2], Natalie C. Cayuela[1,2], Damià Romero[1,2], Alessandra Giorgetti[1,2,3], Mireya Plass[1,2,4]

1. Gene Regulation of Cell Identity, Regenerative Medicine Programme, Bellvitge Institute for Biomedical Research(IDIBELL), L'Hospitalet del Llobregat, Barcelona, Spain

2. Programme for Advancing Clinical Translation of Regenerative Medicine of Catalonia, P-CMR[C], L'Hospitalet del Llobregat, Barcelona, Spain

3. Department of Pathology and Experimental Therapeutics, Faculty of Medicine and Health Sciences, Barcelona University, Barcelona, Spain

4. CIBER-BBN, Madrid, Spain

Neurogenesis is a crucial process involving the formation of new neurons in the developing cortex of embryos, which is regulated by multiple factors, including RNA-binding proteins (RBPs). Staufen 2 (STAU2) is an RBP implicated in the asymmetric distribution of mRNAs in radial glial cells (RGCs), thereby dysregulating the balance between neural stem cell maintenance and differentiation. However, the molecular mechanisms of STAU2-mediated regulation in human neurogenesis remain largely unknown. To characterize these regulatory mechanisms, we performed single-cell RNA-seq (scRNA-seq) on different neurogenic populations derived from STAU2 KO and control human induced pluripotent stem cells (hiPSCs). The samples were sequenced at multiple timepoints, reflecting the transition from hiPSCs (D0) to neuroepithelial cells (D11), neural progenitor cells (D25) and to mature neuronal and glial cell types (D55 and D70). Clustering analysis revealed that the expected cell types such as iPSCs, committed progenitor cells, and mature neurons were recapitulated in the integrated scRNA-seq dataset. Differential Expression analysis suggests that neuronal differentiation of STAU2 KO hiPSCs may occur at earlier stages of differentiation (D11) due to an upregulation of *NEUROD6*, a transcription factor for neuronal differentiation previously described as STAU2 target. Immunohistochemical and qPCR analyses of differentiated neuroepithelial cell cultures show increased RNA expression of neuronal markers such as *MAP2* in STAU2 KO compared to control cultures. Based on these results, we propose a model where STAU2 regulates *NEUROD6* at the neuroepithelial stage during human neurogenesis. To understand how such regulation affects differentiation trajectories, we are performing pseudotime and RNA Velocity analyses. Gene Regulatory Network analysis could provide further insights into the active regulatory modules that may be contributing to the accelerated differentiation observed in STAU2 KO.

27
BENCHMARKING READ MAPPING ON PANGENOMIC VARIATION GRAPHS

Hajar BOUAMOUT 1 , Benjamin LINARD 1 and Matthias ZYTNICKI 1 1 Unité de Mathématiques et Informatique Appliquées, INRAE Occitanie-Toulouse, Auzeville-totosane, France benjamin.linard, matthias.zytnicki}@inrae.fr

References [1] Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017 May;27(5):665-676. [2] Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018 Oct;36(9):875-879. [3] Shuo Wang, Yong-Qing Qian, Ru-Peng Zhao, Ling-Ling Chen, Jia-Ming Song, Graph-based pan-genomes: increased opportunities in plant genomics, Journal of Experimental Botany, Volume 74, Issue 1, 1 January 2023, Pages 24–39 [4] Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol. 2020 Sep 24;21(1):253. [5] Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 2020 Feb 12;21(1):35. [6] Ghanshyam Chandra, Chirag Jain. Sequence to graph alignment using gap-sensitive co-linear chaining. bioRxiv 2022.08.29.505691;

1 Pangenomes and variation graphs A pangenome represents the total genetic diversity of a species or a species complex. One can describe pangenomes in terms of gene presence / absence variations (PAVs) but a more recent alternative aims to integrate full length genomes in a sequence graph [1]. In particular, pairwise alignments of a genome set can be used to build a “Variation Graph” (VG) in which nodes represent words of genome fragments and edges represent the contiguity of the words in one to several genomes (e.g., edges are associated to genome subsets). Each input genome consequently corresponds to a particular path in the graph. It has been showed that VGs can improve variant calling and genotyping processes [2]. Indeed, variant calling is often based on the mapping of linear query sequences to linear reference genomes, thus biasing the prediction to regions present in the reference and limiting the identification of structural variations that are specific to other genomes. In particular, biases are reduced when large structural variations (>50bp) are targeted.
2 Sequence to graph mapping Identifying new variants via a pangenome graph requires a compulsory preliminary step of querying sequence to graph mapping. Several approaches have been proposed (see [3] for a review), with algorithms dedicated to either long or short sequence reads. Most of these tools use a 2 steps method with: a) the identification of candidate sub-graphs showing similarity to the query read via different techniques of indexation followed by b) a refined alignment in the selected sub-graphs (generally via a technique of partial order alignment). In practice, it remains unclear how this preliminary will impact further variants predictions. In particular, it has yet to be shown how resilient will be the different approaches to diverse mutation and indel rates.
3 Tools evaluation The poster will present results produced by Hajar Bouamout during her Master internship dedicated to the evaluation of sequence to graph read mapping tools. It will briefly describe the main ideas behind the algorithms proposed by 4 tools: GraphAligner [4], vg map [5], vg giraffe [5] and Minichain [6], and report benchmarks made with these tools.

28

TRANSCRIPTOMIC ANALYSIS OF CHEMOSENSORY PROTEINS THROUGH THE LARVAL DEVELOPMENT OF HYLAMORPHA ELEGANS (COLEOPTERA: SCARABAEIDAE)

Paula Lizana Ramírez 1, 2 , Julio Rozas 3 , Ana Mutis 2 , Herbert Venthur 2 (1) Programa de Doctorado en Ciencias de Recursos Naturales, Universidad de La Frontera, Temuco, Chile. (2) Laboratorio de Química Ecológica, Departamento de Ciencias Químicas y Recursos Naturales, Facultad de Ingeniería y Ciencias, Universidad de La Frontera, Temuco, Chile. (3) Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), Barcelona, España

Hylamorpha elegans (Coleoptera: Scarabaeidae) is a beetle endemic to Chile, considered an important pest for red clover and cereal crops during its larval stage as white grubs. During this stage, it affects productive areas, such as agriculture and livestock. Considering the limited control of this pest, due to its subterranean behavior, the study of the olfactory system could be an important focus of interest for the knowledge and control of this beetle, where odorant binding proteins (OBP), chemosensory proteins (CSP), odorant receptors (OR), ionotropic receptors (IR) and gustatory receptors (GR) are fundamental for the transport and recognition of organic compounds in the environment. In this sense, the identification of the proteins responsible for the phytophagous condition of the white grub H. elegans would lay the foundation for the development of new effective control mechanisms against this pest. Therefore, the objective of this study was to evaluate presence, evolutionary relationships, and abundance of chemosensory proteins from white grubs of H. elegans. In this study, we performed experiments related to RNAseq, annotation and phylogenetics. Transcriptomic data revealed 2 ORs, 1 GRs, 11 IRs, 24 OBPs and 9 CSPs from two larval stages, L2 and L3 of H. elegans. Subsequently, transcript abundances were determined, where GR1, OBP6 and CSP2 were found to be more abundant in L3 (i.e., the last stage of larval development) than L2, while IR4 is more abundant in L2. Also, the analysis of H. elegans larvae has revealed a repertoire of chemosensory proteins, which showed a differential expression according to the tissues of this beetle. Finally, some proteins have been identified that are of great interest due to the role they may play in larval chemosensation. Moreover, we are working on a comparative genomics of chemosensory proteins of some Coleoptera species of the most important families of this order.

29

EPIGENETIC FINGERPRINTS LINK EARLY-ONSET COLORECTAL CANCER TO LIFESTYLE AND ENVIRONMENTAL EXPOSURES

Silvana C.E. Maas[1], Odei Blanco Irazuegui[1], Iosune Baraibar[2], Jose A. Seoane[1]
1.Cancer Computational Biology Group, Vall d'Hebron Institute of Oncology (VHIO), Centro Cellex, Carrer de Natzaret, 115-117, 08035 Barcelona, Catalonia, Spain
2. Gastrointestinal and Endocrine Tumors Group, Vall d'Hebron Institute of Oncology (VHIO), Centro Cellex, Carrer de Natzaret, 115-117, 08035 Barcelona, Catalonia, Spain
Presenter e-mail: silvanamaas@vhio.net

Recent trends show an increase in early-onset colorectal cancer (eoCRC) in individuals under 50. This study hypothesizes that shifts in the exposome—encompassing behavioral and environmental exposures—may contribute to this rise. Given the impracticality of measuring all exposome traits directly, we use DNA methylation changes as a proxy of the exposome effects. We analyze blood and tumor samples, using methylation risk scores (MRS) and

Pathway Level Analysis of Gene Expression (PLAGE) comparing eoCRC patients to later-onset (loCRC) patients, those diagnosed at 70 or older. Our data comprises samples from the Ontario Familial Colon Cancer Registry and The Cancer Genome Atlas, with further validation through meta-analysis across additional datasets. Logistic regressions—adjusted for family history and sex—contrasting eoCRC with loCRC, using MRS and PLAGE scores as exposure variables. Controls include smoking influence on eoCRC, validating our methodology.
Notable findings indicate that picloram—a pesticide previously understated in literature—affects eoCRC. Significant differences were observed in both blood and tumor samples across several analyses. In conclusion, our results demonstrate exposome-related disparities in methylation between eoCRC and loCRC cases, implicating lifestyle traits and pesticides. These findings suggest the potential for reducing eoCRC incidence through modifications at personal and policy levels.

30

REFGENDETECTOR: A TOOL TO INFER THE REFERENCE GENOME ASSEMBLY FROM HUMAN GENOMIC READ ALIGNMENT FILES

Mireia Marin-Ginestar1,*, Mauricio Moldes Quaresma1, Lauren Fromont1, Arcadi Navarro1 and Jordi Rambla1 1European Genome-phenome Archive (EGA) in the Centre for Genomic Regulation (CRG), the Barcelona Institute of Science and Technology Dr. Aiguader 88, Barcelona, 08003 Spain. *To whom correspondence should be addressed. Presenter e-mail: mireia.marin@crg.eu

Since the completion of The Human Genome in 2004 (International Human Genome Sequencing Consortium 2004), until the most recent release by the Human Pangenome Reference Consortium in 2023 (Liao et al. 2023), significant advancements have been made in genomic research regarding the human reference genome. However, with each successive genome release, the integration of results generated using different assemblies has become increasingly challenging, due to alterations in genome coordinates and annotations (Zhao et al. 2014). To ensure precise and reproducible downstream analysis, it is vital to consider the specific version of the reference genome employed for read alignment. By doing so, accurate interpretation of data and consistent research outcomes can be ensured. To address the challenge of identifying the reference genome assembly used in alignment files, where metadata is unavailable or incomplete, we have developed RefgenDetector, a tool that infers the human reference genome used to create the alignment file using thier header's mandatory information (Samtools 2022). Specifically, this tool (a) compares the lengths of the chromosomes in the different assemblies, as they change in every major release (NCBI 2004v1, NCBI 2004v2, NCBI 2006, NCBI 2009, NCBI 2013, NCBI 2022) with the lengths in the header, and (b) checks the presence of unique decoy contigs, since some of them are only found in one of the derived reference assemblies (Google Genomics 2015, Caetano-Anolles 2015). The RefgenDetector tool allows researchers to easily infer the human reference genome necessary for result interpretation and reproducibility. It can infer among hg16, hg17, hg18, GRCh37, hg19, b36, hs37d5, GRCh38, Verily's GRCh38, hs38DH_extra and T2T.

References Caetano-Anolles, D. (2022). GRCh37 hg19 b37 humanG1Kv37 - Human Reference Discrepancies. Retrieved from ttps://gatk.broadinstitute.org/hc/enus/articles/360035890711-GRCh37-hg19-b37-humanG1Kv37-Human-ReferenceDiscrepancies (Accessed September 4, 2023).

31
PANGENOME GRAPH ANNOTATION TRANSFER

Nina Marthe

Annotation transfer on "linear" genomes usually consists of aligning the annotated regions' sequences on the genome we want to transfer on. In order to annotate pangenome graphs, so to get the positions of genes (and other features) in the graph, we can use the same strategy and simply align the gene sequences on the graph using a graph alignment tool. However this approach can fail to align prefectly some sequences as alignment is a heavy operation, and requires heuristic methods to run in reasonable time.
Pangenome graphs allow a new approach for annotation transfer that doesnt rely on alignement. On the conditions that a graph contains all the sequences of the genomes used to build it, and the paths of these genomes through the graph, simple operations of coordinate conversion can give the precise positions of the regions from a gff/gtf annotation file on the graph. We then obtain the path in the graph of each annotated region, resulting in an annotated graph, without ever having to look at a DNA sequence.

Additionnaly, once the annotated regions are placed on the graph, we can transfer them on any genome contained in the graph with the same precision. We can then compare the annotation transfer between two genomes through the graph and a classical transfert with gene sequence alignement. A comparison with Liftoff showed that the vast majority of the genomic features are placed at the exact same location, showing that this new approach correlates well with a widly used method. A closer look at the features transfered at different locations reveals that these Lifotff transfers have a weaker coverage and identity, suggesting that Liftoff struggles to precisely align the more divergent features. Since the proposed approach only focuses on transfering the shared parts of the annotated features (represented by shared nodes in the graph), its transfer is not affected by the divergences.

32
DECIPHERING THE TRANSCRIPTOMIC LANDSCAPE OF TREATMENT-RESISTANT DEPRESSION

Anna M. Sirés[1], Jorge Domínguez[1], Alessandra Minelli[2], Johannes Zang[3], Britta Kelch[3], María Martínez de Lagrán[4], Bernardo Carpiniello[5], Massimo Gennarelli[2], Filip Rybakowski[6], Marie-Claude Potier[7], Ferran Sanz[1], Bernhard T Baune[3*], Mara Dierssen[4*], Júlia Perera[1*]

1. Research Program on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra (IMIM), Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain
2. University of Brescia; IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli
3. Department of Psychiatry, Albert-Schweitzer-Campus, University of Münster, Münster, Germany
4. Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain
5. University of Cagliari, Italy
6. University of Medical Sciences, Poznan, Poland
7. Paris Brain Institute ICM, Salpetriere Hospital, Paris, France.

Presenter e-mail: amartinez3@researchmar.net

Major Depressive Disorder (MDD) inflicts a substantial global health burden, but current treatments have limited effectiveness and adherence. While pharmacotherapy is typically the

first line of treatment, around one third of MDD patients do not respond to initial therapeutic attempts, leading to the development of Treatment-Resistant Depression (TRD). This resistance not only causes prolonged suffering in MDD patients, but also exposes them to unnecessary treatments and potential secondary effects. Acknowledging this challenge, this study employs integrative clinical and multi-omics data analysis to elucidate the molecular signatures that could distinguish TRD and non-TRD patients.

To achieve this, we first analyzed RNAseq and smallRNAseq data from blood samples of TRD and non-TRD MDD patients. *Functional enrichment analysis* was performed to uncover biological functions, and it revealed that responders have an increased expression of genes related to the immune system, stress response and apoptotic cell death. Conversely, TRD patients manifest a transcriptomic profile enriched in functional terms associated with ribosome-related functions and translation machinery. In line with these findings, there is increasing evidence supporting the notion that impaired mRNA translation is a shared characteristic in multiple psychiatric disorders.

By deciphering the intricate molecular profiles linked to treatment response and incorporating multi-dimensional data, this research lays the groundwork for personalized interventions in the management of MDD and TRD, thereby advancing the field of psychiatric medicine.

33
GENE ANNOTATION VISUALIZATION USING THE PYRANGES LIBRARY

Ester Muñoz (1,2) , Nadezhda Makarova (1) , Max Ticó (1) , Marco Mariotti (1) (1) Department of Genetics, Microbiology and Statistics, Universitat de Barcelona, Barcelona, Spain (2) Universitat Pompeu Fabra, Barcelona, Spain

Pyranges is a Python data structure for handling genomic intervals in a fast and memory-efficient way. The module offers numerous convenient methods, including read/write, overlap, intersection, sorting, sequence retrieval, and many others. Gene annotations, mapped reads, protein domains, and results from Blast or analogous software are some examples of sequence intervals which can be represented as PyRanges objects. Every interval in a PyRanges object is defined, at minimum, by its chromosome and start and end coordinates. Optionally, the strand (+ or -) can also be present, as well as an arbitrary number of meta-data fields. Lately, we are adding customizable visualization functionalities for PyRanges objects, allowing a better understanding of the data. The visualization is based on two alternative engines, the widely used python packages Matplotlib or Pyplot, offering both the same outcome and options. Familiarization with the outcome display is simplified due to its analogy to other genome browsers such as UCSC. Altogether, this makes PyRanges a convenient and efficient tool for genomic data handling, exploration and analysis

34

PANGENOME ANALYSIS OF 288 STRAINS OF VIBRIO CHOLERAE REVEALS COMPLEX DINAMICS AND GLOBAL INTERACTION DURING THE 7TH PANDEMIC'S WAVE IN AMERICA.

Míriam Muñoz-Lapeira1 , Jaime Martínez-Urtaza2 1. IRTA-Food Safety and Functionality Program, Finca Camps i Armet, 17121 Monells (Girona) 2. Department of Genetics and Micriobology, UAB, 08193 Cerdanyola del Vallès (Barcelona) Presenter e-mail: miriam.munoz@irta.cat

The 1990s cholera outbreak in America was a rapid and devastating pandemic that brought this pathogen in the continent after nearly 100 years. It was first detected in Perú and quickly spread but its origin is still object of debate. A collection of 284 Vibrio cholerae genomes were analyzed. 120 were sequenced in "Instituto Nacional de la Salud" in Perú, 34 were retrieved from NCBI's Pathogen Detection project and 130 were published in Domman, 2017. We used an Anvi'o workflow for pangenome analysis and two PCA for visualization, one based on accessory genes and the second on core-genome SNPs, followed by a population genetics analysis. These graphic representations show two generally differentiated groups, one from Asian ancestry that gathers Mexican strains from the 2000s, and other that places Nigerian, Ugandan and Peruvian samples. However, there is a close relationship between Mexican, Brazilian, and Bolivian strains from that decade and strains from Angola, Mali, Burkina Faso or Cote d'Ivoire. In addition, Pi diversity's values were around 0, and Tajima's T around 6 - 7. These results are consistent with clonal expansion but also reiterate the close relationship between Africa's sequences, expected to be evolving independently. Until now, the established theory explains that there were two introductions, one from Africa to Perú and a second one from China to Mexico in the 2000s (Domman, 2017). This work challenges this theory and shows a more complex relationship, indicating at least another introduction from Africa to Mexico in the 90s.

35

AI AND DEMOGRAPHIC INFERENCE IN STRUCTURED POPULATIONS

Alba Nieto

Next generation sequencing data has sparkled a revolution in population genetics, enabling a fine-scale investigation of the evolutionary history of species through the use of an unprecedented amount of data. The field of conservation genetics aims to make use of genetic information to assess the risk of species extinction and develop effective conservation policies based on historical demographic data, benefiting greatly from understanding the demographic history of populations. By gaining insights into the demographic dynamics of each population, we can better identify threats and implement targeted conservation measures to protect endangered species. Yet, understanding the history of species requires a primary focus on investigating realistic demographic models.

The most widely used inference algorithms can predict the variation of coalescent rate (CR) through time in a non-parametric manner (i.e. perform without requiring any distribution modeling this variation). Methods such as PSMC, SMCpp, and Stairway Plot 2 employ various population genetics summary statistics obtained from genomic data and may perform differently depending on i) the true demography of the species under investigation and ii) the time frame considered.

If panmixia (i.e. random mating) is respected, changes in CR correspond to changes in population size over time. Nevertheless, species usually display genetic structure, most often resulting from an organization in meta-population(s). In such cases, the spatial organization of population structure and its strength deeply impact the CR. Hence, interpreting the result of an inference of the variation of CR becomes non-trivial. For instance, under a n-island model (i.e. a simple structured scenario of n populations exchanging the same amount of migrants over time), we observe an increase in the CR in recent times that should not be interpreted as a decline in effective population size. This evidently presents a threat to any endeavor aimed at identifying endangered species or applying conservation measures based on the observed genetic diversity of that species. Hence, there is a need for novel approaches to accurately interpret the coalescence rate and integrate this metric with other population genetic statistics.
In the present PhD I aim to develop new AI based tools to infer the demographic history of a species from its observed genetic variation, incorporating the inferred CR from available algorithms with statistics derived from linkage disequilibrium and detecting population substructure. As a first step, I am evaluating the accuracy and robustness of non-parametric algorithms in estimating the variation of the CR by testing them on simulated sequences under different demographic scenarios.
Next, I will investigate the impact of population structure on the variation of the coalescent rate, focusing on three demographic parameters: the population fragmentation (i.e. number of demes), the deme size, and the connectivity (i.e. migration rate) among them. We will measure the accuracy of SNIF, an available algorithm that uses the inference on the CR performed by PSMC to predict changes in connectivity. Then, I will propose a new Deep Learning algorithm as an inferential tool that integrates information from different summary statistics, such as site frequency spectrum (SFS) and the coalescent rate itself, to provide predictions on population dynamics.

36

MOLECULAR CHARACTERIZATION OF U_6 INVERSION BREAKPOINTS IN *DROSOPHILA SUBOBSCURA*

Kenia M. Delgado[1], Dorcas J. Orengo[1, 2]
[1] *Universitat Oberta de Catalunya.*
[2] *Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, i Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.*

Chromosomal inversions are the structural variation more common within species and also play an important role in speciation. They produced by the breakage of a chromosomal fragment and its reinsertion in the inverted orientation. Chromosomal inversions do not change the gen content of the chromosomes, but they block recombination between homologous chromosomes in heterokaryotypes.
The existence of polytene chromosomes in Diptera has allowed the chromosomal inversion polymorphism to be widely studied in Drosophila. The species *D. subobscura* stands out for having a very rich chromosomal inversion polymorphism in its five acrocentric chromosomes. Studies of the variation of the inversion polymorphism of this species, throughout space, time and altitude, support the idea that the chromosomal inversion polymorphism is maintained by natural selection. Furthermore, it has been observed that the variations described in the

Palearctic region (its original area of distribution) have been rapidly replicated in the colonized areas of both Americas.
It is of great interest to physically delimit the inversions and know which genes are included in the inverted regions in order to understand how natural selection acts through them. To date, some polymorphic inversions of chromosomes A, E and O of *D. subobscura* have been molecularly characterized using laborious chromosome walking and in situ hybridization techniques. Along with massive sequencing, some methods have been developed that can help detect structural variants without the need for these laborious techniques. Here we show how we obtained the breakpoints of a U chromosome inversion by mapping Illumina pair-end reads on a reference genome.

37

EVOLUTION OF NUCLEAR RECEPTOR PROTEIN SEQUENCES AND DNA BINDING FEATURES (MOTIF AND SITES) IN CRUSTACEAN GENOMES: A CASE STUDY OF THE NR2B FAMILY

Murat Tugrul[1] and Ferran Palero[2]

[1]Institute of Biology, Freie Universitat Berlin, Berlin, Germany
[2]Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia, Paterna, Spain.
Presenter email: murat.tugrul@fu-berlin.de

Nuclear receptors (NRs), a special class of transcription factors (TFs), control gene expression in response to signalling molecules and environmental stimuli, influencing physiology and fitness. NR evolution is crucial for metazoan diversification and adaptation, yet the coevolution of NR protein sequences, DNA binding preferences and sites remains poorly understood. We focus on a key family NR2B (usp/rxr) and on crustaceans, a highly diverse group of invertebrates occupying various ecological niches, with available whole genome sequences and transcriptomic data. Our findings reveal that the evolution of NR2B protein sequences is closely aligned with the phylogeny of Crustacea, in agreement with a common ancestor shared between Copepoda and Cirripedia, while Euphausiacea and Decapoda diverged later within Malacostraca. Moreover, using a machine learning approach, we predict binding motifs for various taxa, revealing mostly conserved motifs with limited evolutionary change. However, interestingly, motif similarity between species shows unpredictability, loosely following main phylogenetic clades. Furthermore, we explore the binding site evolution within crustacean genomes. We find slight motif changes trigger numerous transcription factor binding site (TFBS) turnovers, potentially rewiring gene regulatory networks and leading to distinct phenotypes. This study provides new insights into the coevolution of nuclear receptors and their DNA binding features in crustaceans, enhancing our understanding of the molecular mechanisms governing their development and adaptation to diverse ecological conditions.

38

THE HUMAN ORAL MICROBIOME IMPLICA3ONS IN ALZHEIMER'S DISEASE PROGNOSIS

Sara Peregrina

Alzheimer’s disease (AD) is the most common type of demenFa, with no cure so far, but few medicines to alleviate the symptoms, which would be more opFmal if the disease was

detected at an early stage. Chronic inflammaFon is the first sign of disease progression, influenced potenFally by microbiome changes. Previous studies have shown associaFons between periodonFFs, an infecFon of the gums due to alteraFon of the oral microbiome, and Aß-pepFde plaques in the brain, one of the causes of AD. Our 16S metabarcoding results based on V3-V4 hypervariable region showed no differences in diversity and overall composiFon comparing the different diagnosFc groups across the AD progression: Mild CogniFve Impairment (MCI), ObjecFve DemenFa (OD), Severe CogniFve Decline (SCD) and AD, but we observed more similarity within the SCD group as compared between the other diagnosis. Our differenFal analysis confirmed greater differences when comparing SCD with the others and detected more differences from SCD to AD, the last step of the disease progression, including taxa such as Prevotella nanceiensis, Prevotella den0cola and Anaeroglobus geminatus. Finally, we automated a dbBact knowledge database search to biologically interpret the results, and link them with previous knowledge. We detected almost 60% of the species listed as differenFally abundant as previously associated with the term periodonFFs, a condiFon that was extensively linked with AD disease. On the whole, these results open the door to further studies to validate the findings and assess their potenFal roles as biomarkers to be able to detect this disease before developing full-blown AD, as it would improve the prognosis and quality of life of paFents.

39

CHANGES WITH AGE IN THE CORE HUMAN GUT MICROBIOME OF INDIVIDUALS FROM A MEDITERRANEAN COHORT

Samuel Piquer-Esteban[1,2], Susana Ruiz-Ruiz[2,3], Wladimiro Diaz-Villanueva[1,3], Vicente Arnau[1,3], Andrés Moya[1,2,3]

1. Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain
2. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region(FISABIO), Valencia, Spain
3. Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública(CIBEResp), Madrid, Spain
Presenter e-mail: Samuel.Piquer@uv.es

The human gut microbiome is key to understanding the health status of its host. It plays a crucial role during an individual's lifetime in the development of the immune system, the prevention of infections, nutrient acquisition, among others. This set of microorganisms is not fixed in time and its establishment follows the process of an ecological succession, changing dramatically during the first years of life until becoming increasingly complex during adulthood to later losing diversity during elderhood. While there are studies that focus on the analysis of the microbiota with age, few evaluate the changes in prevalence and conservation of taxa throughout life. In the present work, we characterized through 16S rRNA gene sequencing the gut microbiome from healthy individuals of three well-defined age groups from a Mediterranean cohort, including Infants, Adults, and Elders (IAE cohort) that were sampled during a period of two years. To better evaluate the taxa identified at high, medium, and low prevalence, only genera detected in all, 80%, and 50% of the samples considering all time points per age group were evaluated, respectively. Our results show the presence of both a set of conserved core taxa throughout life, and others with consistent changes in prevalence with age, representing possible age-related-markers.

40

GALEON - A BIOINFORMATIC TOOL FOR THE ANALYSIS AND VISUALIZATION OF GENE CLUSTERS

Vadim A. Pisarenco (1,2) , Joel Vizueta (3) and Julio Rozas (1,2) (1) Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain. (2) Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain. (3) Villum Centre for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark.

Gene clusters, groups of genes encoding functionally-related proteins, are commonly found in eukaryotic genomes. One of the most abundant types are gene families, those encoding a set of homologous genes commonly originated through gene duplication often via unequal crossing-over. As a result, they are found arranged in tandem in the genome. Despite the increasing number of whole genome information for many species, the comprehensive evolutionary analysis of gene family members remains largely unexplored. Two primary, not exclusive, challenges hinder this exploration: the analysis of large gene family sizes, and those of very recent (young) family members. These issues stem from limitations in current assemblies (including many of those based on long reads) that can not accurately assemble extensive stretches of repetitive DNA regions (such as those of recently originated copies). Current sequencing techniques (based on long reads and chromatin contacts) offer a promising avenue for addressing these challenges. The discovery and visualization of gene clusters is usually case-specific, involving arbitrary criteria and custom bioinformatic pipelines. To overcome such limitations, we present GALEON, a user-friendly bioinformatic tool to identify, analyze and visualize physically clustered gene family genes in chromosome-level genomes. The software uses simple input file formats with gene coordinates (BED or GFF3), and protein sequence data. Specifically, the gene cluster analysis is assessed by analyzing the distribution of pairwise physical distances between the gene set's members and the average genome gene density. GALEON also allows the study (and comparison) of two gene families at once and, if the protein sequence data is provided, can also explore the relationship between physical and evolutionary distances. Overall, GALEON represents a novel tool for the study of clustered genes to gain valuable insights into the origin, evolution and function of gene families, while also providing an utility to evaluate the local quality of assembled genomic regions.

41

DISCOVERING NOVEL BACTERIAL SYMBIONTS IN NUDIBRANCHS

Dafne Porcel Sanchis1 , Samuel Piquer-Esteve1 , Vicente Arnau1 , Wladimiro DiazVillanueva1 , Maria Dzunkova1,2,3
1. Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain
2. Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA 3. Department of Energy Joint Institute, Berkeley, CA, USA

Soft-bodied marine animals, such as marine sponges and nudibranchs, use bioactive molecules to protect themselves from their predators. Their microbiomes are seen as a possible source of new bioactive compounds. Sequencing of the 16S rRNA gene amplicons represents the first insights into the presence of novel bacterial groups that might be further investigated for

their biosynthetic potential. In contrast to marine sponges, the nudibranch microbiome still contains a large portion of unknown bacteria. Characterization of the nudibranch microbiome is challenging due to under representation of its symbiotic bacteria in conventional databases. In our previous single-cell genomics study of nudibranch microbiome, we detected a new member of Candidatus Tethybacteriales (Candidatus Doriopsilibacter californiensis), an uncultured order of endosymbiotic microbes recently discovered in marine sponges which is not yet included in conventional databases. To improve classification of novel bacterial groups in nudibranch samples an enriched database is created, adding novel marine 16S rRNA bacterial sequences from Ca. Tethybacteriales order to the SBDI Sativa curated 16S GTDB database (SBDI; 2021). The first dataset employed is composed by Doriopsilla fulva samples from mantle, gills and internal organs. Results suggest that D. fulva mantle is the reservoir of Candidatus Doriopsilibacter californiensis and suggests a symbiotic relationship. The second dataset is formed by samples of mantle and mucus from 20 different nudibranch genera. The data depics that a large proportion of nudibranch microbiome is formed by uncharacterized bacterial species, and some of them dominate the bacterial composition in these samples, suggesting a functional relationship. Core analysis also suggests other characterized bacteria are present in high prevalence and abundance in samples of certain nudibranch species, e.g. Vibrio species are found in all Doris pseudoargus nudibranch mantle samples, as well as Endozoicomonas in Trapania maculata skin samples. Keywords: Nudibranch; Tethybacterales; 16S rRNA; simbiont; database.

42

GENOME STRUCTURE AND TE PREFERANTIAL INSERTION IN THE MOSS PHYSCOMITRIUM PATENS

Marc Pulido

The moss Physcomitrium patens is a plant model species used in evolutionary studies due to its phylogenetic position as a basal clade of land plants. Since becoming the first bryophyte with a sequenced genome, our understanding of plant genomes has greatly advanced. The latest P. patens genome assembly has 481 Mb, and like to other plant genomes of similar size, it is formed by approximately 60% of repetitive sequences, predominantly LTR-Retrotransposons. Interestingly, this genome displays an even distribution of gene and TE-rich regions, different from the pattern observed in flowering plants, where TE-rich regions are mostly concentrated in pericentromeric regions. Furthermore, similar genome structures have been observed for the liverwort Marchantia polymorpha or the hornwort Anthoceros agrestis, indicating that this could be a common trait for all bryophytes. These TE-rich regions are predominantly occupied by a single gypsy LTR-RT family, RLG1, which seems to target heterochromatin for integration. However, in spite of the generally even TE distribution in the genome, a single copia LTR-RT family, RLC5, clusters at a single location in each chromosome that coincides with the centromere, suggesting strong specificity of insertion. Our research seeks to understand the molecular mechanisms driving the preferential insertion of these two LTR-RT families. Through transposition experiments in the model plant Arabidopsis thaliana, we aim to mobilize these elements and analyze the distribution of new insertions to see whether they retain their capacity to target specific genomic regions. This investigation will shed light into the structural changes in plant genomes since the early divergence of land plants.

43

METAGENOMIC SOFTWARE FOR ITS TAXONOMIC ANNOTATION

Álvaro Redondo-Río Barcelona Institute for Research in Biomedicine (IRB), Carrer de Baldiri Reixac, 10, 08028 Barcelona, Spain & Barcelona Supercomputing Center (BSC-CNS), Carrer de Jordi Girona, 29, 31, 08034 Barcelona, Spain.

The ITS region is a widely used molecular marker for studying fungal diversity and taxonomy, making it a fundamental tool in shotgun mycobiome research. Accurate annotation of ITS amplicons is crucial for obtaining meaningful insights from metagenomic data. However, the relatively slower development of reference databases and the complexity of fungal taxonomy makes this annotation more challenging and less precise than that of bacterial 16S. A recent publication this year by Odom A. et al. showed that metagenomic profiling pipelines improve taxonomic classification of 16S sequences relative to amplicon-tailored tools. To test if the same applies to ITS sequences, and in the lack of any comprehensive comparison, we present a methodology for benchmarking bioinformatic approaches to ITS taxonomic annotations, enabling researchers to make informed decisions about the best annotation strategy. Our methodology comprises four key steps: 1. Data Collection: First, we gathered publicly available ITS sequences from reported mock communities. The resulting dataset encompasses a broad range of fungal taxa to represent the complexity of real-world mycobiomes. 2. Annotation Tool Selection: We then selected a set of amplicon sequence annotation tools, together with other tools made for metagenomic annotation to incorporate tools that are not tailor-made for amplicon sequences. 3. Database Selection: As the reference database is also a crucial choice for taxonomic annotation, multiple sources were used. Amplicon-specific databases were combined with generalist databases that include whole genome sequences. 4. Analysis Metrics: To evaluate the performance of each tool-database combination, we defined a set of evaluation metrics. These metrics will help quantitatively compare the performance of different annotation methods. By following this methodology, we will evaluate the informatic analysis of ITS sequence annotations. The results obtained will enable us to make informed decisions about the choice of annotation tools and strategies for their specific microbiome research projects. Furthermore, this comparative approach promotes transparency and reproducibility in the field of microbiome informatics, facilitating the advancement of our understanding of microbial communities and their roles in various ecosystems and health-related studies.

44

THE BIOGENOME PORTAL: A WEB-BASED PLATFORM FOR BIODIVERSITY GENOMICS DATA MANAGEMENT

Emilio Righi, CRG

In the field of biodiversity genomics, efficient data management and integration are critical to advancing scientific understanding and conservation efforts. Here we present a web-based application tailored to the specific needs of biodiversity genomics projects. The application integrates user-generated data with public datasets, including taxonomy and geographic co-ordinates, facilitating comprehensive research and collaboration.

In particular, this web application could play an important role within the Earth Biogenome Project network by facilitating the generation of sequencing status reports to be sent to Genomes on a Tree and by providing real-time monitoring of the INSDC submission status for target organisms, thereby simplifying project coordination.

Key features of the application include the ability for users to upload spreadsheets of sample metadata to facilitate data entry and organization. In addition, the system supports the linking of customized data to species profiles, including essential information such as photos, common names, links to relevant publications, and any other metadata relevant to the research.

In summary, this web application is an invaluable tool for biodiversity genomics projects, providing researchers with the means to efficiently manage, integrate, and scale data while contributing to the broader mission of understanding and conserving Earth's biodiversity.

An instance of this application containing all the data generated under the Catalan Initiative for the Earth Biogenome Project can be found at: https://dades.biogenoma.cat/

45

A ROBUST STATISTICAL FRAMEWORK FOR GENE-WISE SINGLE-CELL DIFFERENTIAL EXPRESSION META-ANALYSIS IN THE CONTEXT OF POPULATION-BASED SINGLE-CELL STUDIES

Aida Ripoll-Cladellas[1*], Marc Jan Bonder[2,3,4], Maria Sopena[1], single-cell eQTLGen consortium, Lude Franke[2], Monique G.P. van der Wijst[2] & Marta Melé[1]

1. Life Sciences Department, Barcelona Supercomputing Center, 08034, Barcelona, Catalonia, Spain
2. Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
3. Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
4. Genome Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Presenter e-mail: aida.ripoll@bsc.es

Single-cell RNA sequencing (scRNA-seq) has enabled deciphering the human transcriptome at an unprecedented resolution. Its popularity for studying how inter-individual variation affects expression has grown tremendously now that scale, cost, and sensitivity have significantly improved. To fully leverage these emerging population-based scRNA-seq data resources, we have founded the single-cell eQTLGen consortium (sc-eQTLGen), aimed at pinpointing the cellular contexts in which disease-causing genetic variants affect gene expression. Our consortium builds on a federated structure that 'brings the algorithm to the data', thereby overcoming the necessity to share privacy-sensitive data, while concurrently reducing the computational load needed for processing all datasets together.

Here, we expand the sc-eQTLGen meta-analysis setup to pinpoint the cellular contexts in which specific individual traits (such as age, sex, or ethnicity) or different environmental conditions (such as immune stimulation) affect gene expression. To this end, we have developed a statistical framework to conduct a cell-type-specific gene-wise single-cell differential expression meta-analysis (SiGMeta-DE). As a proof of concept, we applied this framework to three peripheral blood mononuclear cell datasets spanning 809,353 cells from 173 different donors to study how sex (discrete) and age (continuous) affect gene expression. Our approach overcomes several limitations. First, we have used MAST, a two-part Hurdle generalized linear

model with random effects for individuals to account for both zero inflation and pseudoreplication bias. Second, by using a meta-analysis approach, we avoid sharing privacy-sensitive and re-analyzing previously processed data, which can often be difficult and cumbersome.

We show that our meta-analysis methodology substantially increased the statistical power to detect differentially expressed (DE) genes. Sex-DE genes detected in the individual datasets were all located in the sex chromosomes as these have generally large effect sizes. Our meta-analysis identified additional sex-DE genes, including many with smaller effect sizes located in autosomal chromosomes. In addition, the meta-analysis considerably increased the number of age-DE genes up to hundreds. Newly identified age-DE genes overlapped previously identified age-related genes but many are novel.

Together, our meta-analysis framework overcomes the difficulties that have previously hampered studying inter-individual variation in gene expression at single-cell resolution and allows the identification of DE effects that would otherwise remain hidden. Our approach provides a solid framework to associate single-cell molecular phenotypes (such as single-cell chromatin accessibility or single-cell DNA methylation) with demographic traits or environmental conditions in future studies.

46

APPLICATION OF MACHINE LEARNING MODELS IN THE IDENTIFICATION OF PEDIATRIC LIVER CANCER BIOMARKERS THROUGH GENE EXPRESSION PROFILING

DAVID RUBIO MANGAS1, Álvaro del Río2, Laura Royo2, Montse Domingo-Sàbat2, Juan Carrillo-Reixach2, Alfonso Valencia1, Carolina Armengol2, Davide Cirillo1

1 Barcelona Supercomputing Center (BSC), Barcelona, 08034, Spain
2 Childhood Liver Oncology Group, Germans Trias i Pujol Research Institute (IGTP), Translational Program in Cancer Research (CARE), Badalona, 08916, Spain

Pediatric liver tumors, such as pediatric hepatoblastoma (HB) and pediatric hepatocellular carcinoma (HCC), are rare but are increasing in incidence. HB occurs in young children (<3 years), whereas HCC affects older ages (>8 years) and adolescence. The category of hepatocellular malignant neoplasm not otherwise specified (HEM-NOS) falls between HB and HCC in histopathologic complexity. Due to its rarity (~1 case per million children), investigation of its biology and lack of biomarkers in clinical diagnoses make it difficult to approach.

This study aims to identify biomarkers for molecular stratification in pediatric liver cancers through an extensive gene expression panel of genomic regions using the Nanostring-nCounter assay.

Exploration of the data using Kruskal-Wallis and Dunn tests led to the identification of regions with statistical significance in diagnosis. Using these regions, machine learning models were designed specifically tailored to the dataset, taking into account its limited sample size and class inequalities. These models, which encompassed methods such as multinomial logistic regression, were subjected to training and evaluation using robust approaches suitable for this context, such as the assignment of weights to classes, leave-one-out cross-validation and the use of the F1-score.

The resulting molecular stratification via Nanostring revealed promising correlations between predictor variables and categories (HB, HCC and HEM-NOS), offering a comprehensive and predictive framework. This work constitutes a significant effort to address the challenges of machine learning in rare diseases, such as data sparsity and sampling imbalances, along with difficulties in data validation.

The clinical utility of the Nanostring panel reinforces its diagnostic potential in pediatric patients with liver cancer, suggesting practical applications for a more accurate identification and management of these cases.

47

DEVELOPMENT OF AN EASY-TO-USE BIOINFORMATIC TOOL FOR ANALYZING THE NON-CODING REGION OF ANCIENT HUMAN MITOCHONDRIAL DNA USING NEXT-GENERATION SEQUENCING

Daniel R. Cuesta Aguirre, Assumpció Malgosa, Cristina Santos

The study of genomes and transcriptomes in living or dead organisms is a common practice thanks to the Next-Generation Sequencing (NGS), that allow the increasing of throughput and dept coverage by reducing costs and time in contrast to Sanger sequencing. Moreover, mitochondrial DNA (mtDNA) could increase the success of the experiment due its great number of copies. However, with some degraded or ancient samples an amplicon-based NGS techniques are required to obtain enough data to be analysed. There are pipelines that allow the study of mtDNA samples and others based on amplicons. Withal, its installation and manipulation could be difficult for non-expert users. Moreover, they are not able to deal with both ancient and amplicon-based sequencing challenges at once. Therefore, our main goal was to generate an easy-to-use bioinformatic tool to analyse the Non-Coding Region of ancient human mtDNA. We compared MarkDuplicates from Picard and dedup parameter from fastp, two tools created to remove duplicates. Moreover, we analysed different threshold of PMDtools, a tool designed to extract reads with post-mortem degradation. We also correlated the depth coverage of each amplicon with its level of damage. Being dedup a better option to work with amplicons and PMDS=1 the main threshold to differentiate between present-day and ancient samples when working with PMDtools, these two bioinformatic tools were added to a pipeline designed to obtain both haplotype and haplogroup. A negative correlation was found between number of reads and percentage of damage for amplicons. Furthermore, information about the quality and possible contamination of the sample is generated. Our pipeline is designed to help people to both analyse and understand their own samples, but some manual analyses are required to totally understand each one.

ACKNOWLEDGMENTS
This research was supported by the Ministerio de Ciencia e Innovación (Ref. PGC2018-096666-B-100), By the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya and the Fons Social Europeu. DRCA is a PhD fellow from the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) (2020 FI_B1 00641). CS and AM are members of the GREAB supported by the Generalitat de Catalunya (Ref. 2017 SGR 1630; 2021 SGR 00186).

48

GORILLA POPULATION GENOMICS USING NON-INVASIVE SAMPLES

Irune Ruiz-Gartzia1 , Harvinder Pawar1 , Esther Lizano1,4 and Tomas Marquès1,2,3,4 1. Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain 2. Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain 3. CNAG, Centre Nacional d'Anàlisis Genòmic, Barcelona, Spain 4. Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain

Human pressure on wild populations has increased over the last 200 years. As such, gorillas have been impacted, resulting in a drastic population decline and habitat fragmentation across most of its range (Humle et al., 2016; Maisels et al., 2018). The genus gorilla has two species, the eastern and the western, and each species is divided into two subspecies, three of them labelled as critically endangered and the mountain gorillas as endangered. Genomic analyses of wild endangered populations are benefitting from the advantages of non- invasive samples. This type of samples allows to cover a larger amount of the entire distribution of current wild populations. To study the genetic variability among gorilla (sub)species and populations, almost 300 faecal samples have been collected over the last few years. To enrich in endogenous DNA, we applied target capture methods to capture the chromosome 21 of gorilla faecal samples obtaining a large number of SNPs. This dataset is being analysed together with available high-quality whole genomes and new cross river genomes which have been generated from hair samples. In addition, natural history collections worldwide house millions of specimens, constituting a unique repository of biodiversity. In this study we include museum samples to increase the gorilla population distribution representation. Moreover, these specimens add a temporal dimension to the genetic studies of endangered species. Our main goals are: 1) determine the genetic diversity of each gorilla (sub)species at population level, 2) identify possible genetic structure within and among geographic range, 3) determine the degree of separation and gene-flow among (sub)species taxonomic classifications, and 4) integrate the data with the largest available data on present-day variation for this genus.

49

COMPLETING THE GLOBAL COLONIZATION HISTORY OF DROSOPHILA MELANOGASTER USING POPULATION GENOMICS FROM POOL SEQUENCING DATA

Carlos M. Tinedo (1), Carlos E. Arboleda-Bustos (2), Dorcas Orengo (1,3), Alejandro SánchezGracia (1,3) (1) Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain; (2) Grupo de Neurociencias, Instituto de Genética, Universidad Nacional de Colombia, Bogotá, Colombia; (3) Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.

Drosophila melanogaster is one of the most popular organisms in biological research, specially in the field of genetics. This insect, also known as the fruit fly, has become a cosmopolitan organism by colonizing every continent except Antarctica as an human commensal. Originally from Africa, we now find stable populations at almost every latitude or altitude on the planet, in a wide range of climatic conditions. Although the global colonization history of this species has been extensively studied in the majority of continents, the origin and genetic composition of the populations form South America is still completely unknown. Recently, we have obtained pooled samples from natural populations of this species from Colombia, Ecuador, and Brazil. In this study, we present the preliminary population genomics analysis of these

samples, which is only the first step in a much larger project that aims to answer these and other questions about the origin of these populations and the genomics basis of adaptation.

50
EVOLUTION OF EPIGENOMICS AND GENE REGULATION IN THE PRIMATE RADIATION

Maria Torralvo 1 , David Juan 1 and Tomas Marques-Bonet 1,2,3,4 1. Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain 2. Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain 3. Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain 4. Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain

Historically, one of the primary objectives of evolutionary biologists has been to understand the molecular basis underlying phenotypes. The importance of gene regulation in primate phenotypic diversity and species-specific adaptation was first discussed by King and Wilson in 1975. A hypothesis that gained credibility after the emergence of numerous pieces of evidence on the significant genetic similarity among phenotypically divergent species and on the role of epigenetic differences in primate evolution. DNA methylation is a stable epigenetic modification involved in the regulation of numerous biological processes, conditioning cell, tissue, and organism phenotypes. However, the dynamics of DNA methylation evolution are still poorly understood. In primates, comparative studies on gene regulation have been limited to a few model species, primarily because of the difficulty of obtaining new samples and ethical considerations. For the IV IBE PhD Symposium, I will briefly present the research project I am currently working on, which includes the most extensive dataset of whole genome DNA methylation patterns across the primate phylogeny to date. The results of this analysis will help to comprehend better the evolution of the genome and epigenome in primates, delivering a valuable integrative resource of great interest for the scientific community.

51
TOWARDS A GOLD-STANDARD WORKFLOW FOR THE PROFILING OF CELL-FREE RNA IN BLOOD PLASMA

Cristina Tuñí i Domínguez 1 , Lluc Cabús 1 , Phil Sanders 1 , Marc Weber 1 , Julien Lagarde 1 1. Flomics Biotech SL, Carrer Pujades 94-96, 08005, Barcelona, Spain. Presenter e-mail: cristina.tuni@flomics.com

Liquid biopsies have become increasingly important diagnostic tools for early cancer detection due to their high sensitivity and minimal invasiveness. The use of plasma cell-free RNA (cfRNA) as a biomarker for such purposes, while promising, presents several important technical challenges. These challenges include the acquisition of high-quality, diverse cfRNA molecules, mitigating genomic DNA (gDNA) contamination in sequencing libraries, and the optimisation of bioinformatic pipelines. Another challenge is that a systematic large-scale comparison of cfRNA-Seq experimental workflows is currently lacking. The existence of this comparison could set a gold-standard workflow for the profiling of cell-free RNA in blood plasma. In this study, we address these challenges by presenting the results of a comprehensive analysis, comparing quality control metrics across both in-house and publicly available cfRNA-Seq datasets. Despite a universally reported high rate of exon mapping, a proxy indicator of cfRNA quality, metrics like the percentage of mapped reads and library complexity exhibited high variability across studies. We included in-house data in the comparison to assess the improvement yielded by several generations of our own experimental protocol. The

comparative analysis of our data revealed that iterative updates to our own protocol consistently yielded improvements in data quality. Our study highlights the importance of continuous quality control and iterative protocol optimisation in cfRNA-Seq to improve the reliability of nucleic acid detection. The advancements we demonstrate in cfRNA processing and analysis not only improve the quality and fidelity of sequencing data but also provide the foundation for more accurate diagnostics methods. By establishing benchmarks for sample quality and bioinformatics analysis, we facilitate the advancement of non-invasive diagnostic methods that could improve early cancer detection and patient monitoring. This sets the stage for future studies that could validate our enhanced protocol across different populations and cancer types, ultimately leading to better patient outcomes through earlier intervention.

52

RECOVERY OF ASSEMBLED GENOMES FROM METAGENOMES (MAGS) IN EXTREME INDUSTRIAL ENVIRONMENTS

Juan Valero Tebar1 , Maria Dzunkova1 Presenter e-mail: juanvate@alumni.uv.es 1. Institute for Integrative Systems Biology, University of Valencia and Consejo Superior de Investigaciones Científicas (CSIC), 46980 Valencia, Spain

Water pollution by hexavalent chromium is an environmental concern due to its high toxicity and persistence in aquatic ecosystems, requiring innovative bioremediation solutions. In this study we sequenced metagenomes from four basins of a Tunisian tannery that accumulate wastewater at different stages of the wastewater treatment process. Assembly and binning were performed on the contigs to recover microbial genomes from metagenomic sequences (MAGs) and classify them taxonomically. The contigs were assembled using metaSPAdes with two sets of k-mers {21, 33, 55, 77} and {21, 33, 55, 77, 99, 127} to study the effect of different assembly parameters on the MAG generation. Afterwards, Bbmap was used for mapping the reads to these contigs, and MetaBAT 2 for the contigs binning. The obtained MAGs were filtered with CheckM based on their quality and contamination, and GTDB-Tk was used for taxonomic classification. In addition, a coassembly of all samples was performed, and the resulting MAGs were compared with MAGs from individual assemblies. In total, 142 different MAGs were obtained, highlighting the presence of a wide variety unidentified species, genera, families, and even phylum of bacteria and archaea. We observed differences in microbial composition in the four tannery basins. The next step is to identify metabolic pathways that could be used to develop new bioremediation technologies