# X Bioinformatics and Genomics Symposium
# (BGS-X)

Organized by:
Cavanilles Institute of Biodiversity and Evolutionary Biology (ICBiBE)
Institute for Integrative Systems Biology (I2SysBio)
Bioinformatics and Genomics Section of the SCB
Bioinformatics Barcelona – BIB

Sponsored by:

## *Charles Darwin* Hall

## Campus de Burjassot (UVEG)

Avinguda Vicent Andrés Estellés, Burjassot
December 15th - 16th, 2022

ORGANIZING COMMITTEE:

Ana Conesa (I2SysBio)
Ferran Palero (ICBiBE)

SUPPORT:

Mario Cáceres (UAB)
Roderic Guigó (CRG-UPF)
Ana Ripoll (UAB, BIB)
Mariàngels Gallego (SCB)
Maite Sánchez (SCB)
Alexandre Mestre Pérez (ICBiBE)
Sara Tortosa Martínez (ICBiBE)

https://www.iec.cat/jornades/genomica2022.asp

# FOREWORD

The ***Bioinformatics and Genomics Symposium*** is a conference held every year since 2013 with the aim of promoting the development of bioinformatics and genomics and their application. The main objective of this meeting is to promote scientific interaction between different groups, presenting the excellent research carried out in bioinformatics and genomics, as well as advances in our understanding of the genome in humans and other organisms, new technological developments, applications, and the impact of genomic data in evolutionary biology, among other areas. Originally designed to bring together world specialists in the subject, this symposium has already been successfully organized on 9 occasions, having fallen the responsibility as hosts of its 10th edition to the researchers Ferran Palero (ICBiBE, UVEG) and Ana Conesa (I2SysBio, CSIC-UVEG). This year we intend to gather in Valencia more than 150 scientists and experts from several countries, who will be able to participate in the form of oral presentations and posters, as well as social events to promote relationships and future collaborations between attendees. We hope you will have a fruitful visit to Valencia and enjoy the meeting!

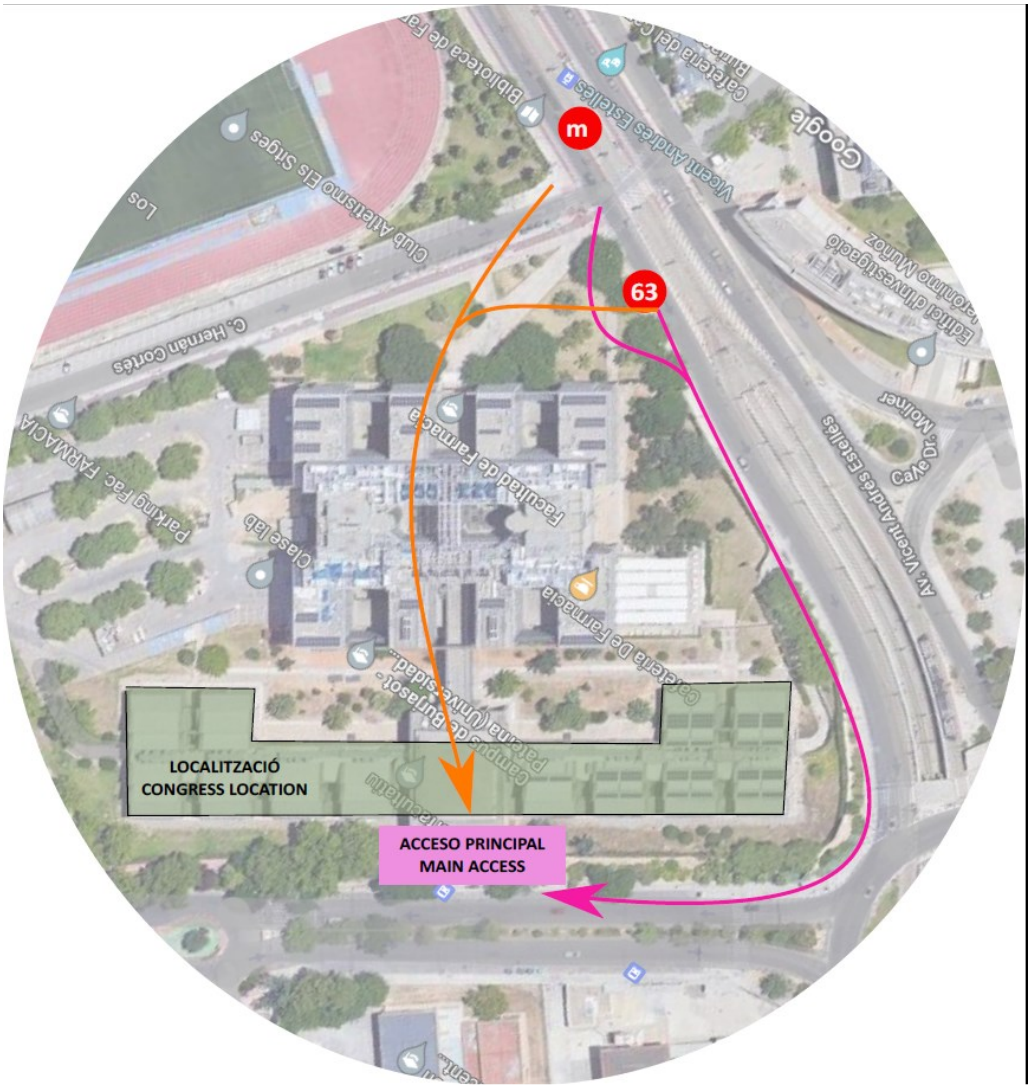
The organizing Team

# Site Plan: Everything you Need to Know

Buses should be the preferred option if you are travelling from the “Estació del Nord” train station to the Aulari Interfacultatiu (Campus de Burjassot): https://goo.gl/maps/GDcZNf71onmvmXjd9

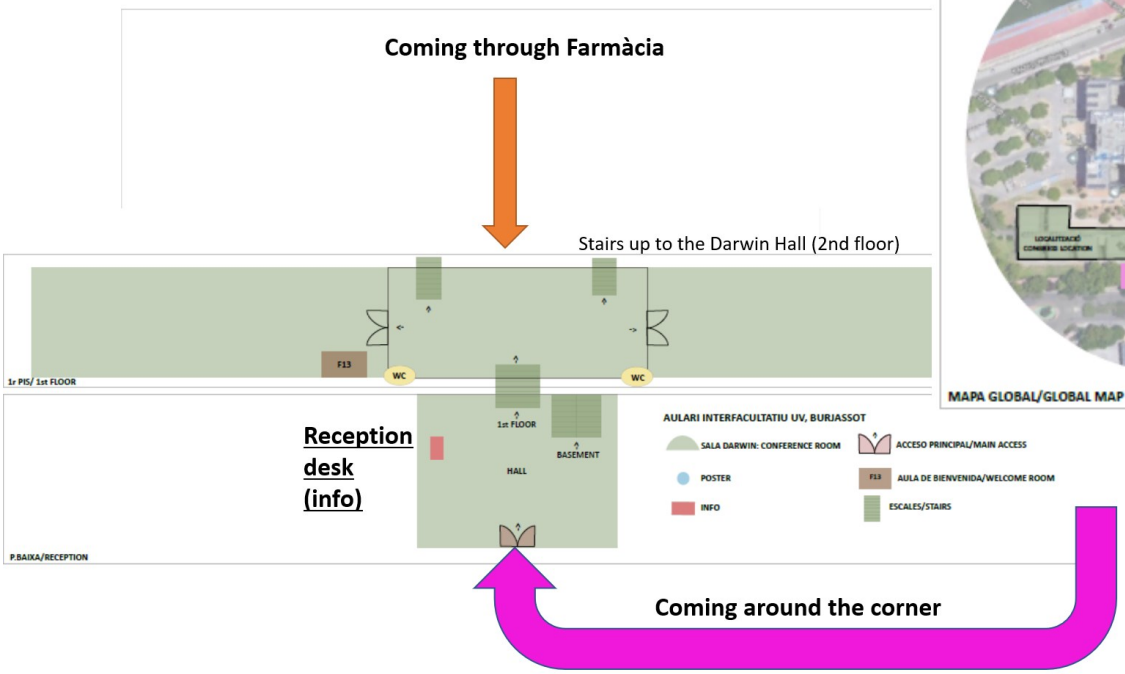
1. Bus: Line 63, Stop “Campus Burjassot” (https://www.emtvalencia.es/ciudadano/index.php?lang=en)
2. Metro/Tram: Line 4, Stop “Vicent Andrés Estellés” (https://www.metrovalencia.es/ca/index.php/)

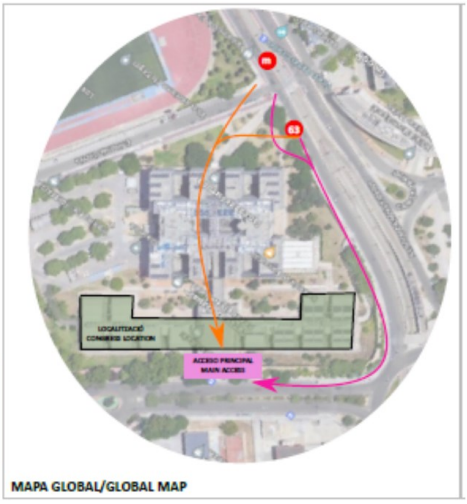
Once you get to the “Campus Burjassot” (Bus) or “Vicent Andrés Estellés” (Tram) final stops, you just need to walk through the Facultat de Farmàcia or around the corner to reach the Aulari Interfacultatiu main Access, which is where the welcome/reception desks will be:



The Aulari Interfacultatiu will host every activity during the Symposium except the Gala Dinner (which will happen right in Valencia`s city heart, at the Octubre Centre de Cultura Contemporània, Carrer de Sant Ferran, 12; https://octubre.cat/).

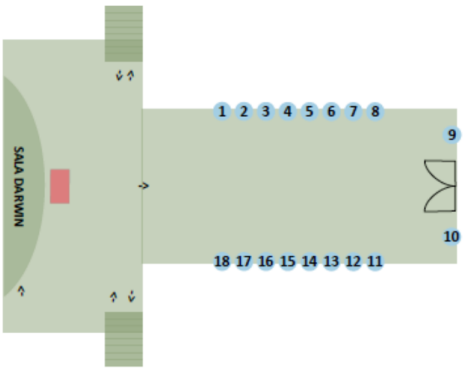
## ARRIVAL AND RECEPTION



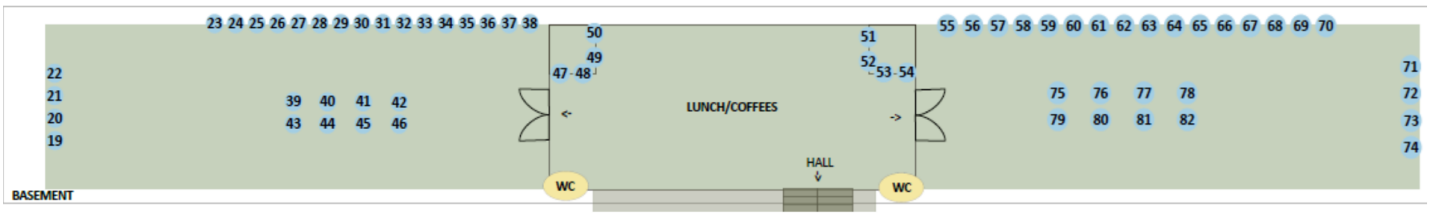




## POSTER SESSIONS AND COFFEE BREAKS

**Top (2nd floor)**




**Bottom (Basement)** => All coffee breaks and lunch breaks will happen here!




*Poster numbers are shown within blue circles.

# PROGRAM AT A GLANCE

## *December* 15th

12:00 – 13:00 Pick up documentation at the Aulari Interfacultatiu (Campus de Burjassot).

13:00 – 14:00 Welcome lunch/reception and opening of the symposium.

### SESSION I.

14:00 - 14:45 **Invited Lecture: Mark Blaxter (Sanger).** 500 completed genome sequences from the Tree of Life: New genomes yield new biology.

14:45 - 15:00 **Maria Diaz Ros (UAB).** Long-term inversion recurrence and segmental duplication conservation during mammalian evolution.

15:00 - 15:15 **Carles Galià Camps (UB).** A worldwide invasive pangenome sheds light on multiple molecular evolutionary processes.

15:15 - 15:30 **José Miguel Serradell Noguera (UPF).** North Africa demographic history through an ABC-DL approach**.**

15:30 - 16:00 Coffee Break

### SESSION II.

16:00 - 16:15 **Jorge García Calleja (UPF).** Exploring genetic adaptation in the iberian peninsula using a genome wide scan of positive selection in the GCAT data.

16:15 - 16:30 **Francesc Ganau Penella (CNAG).** Can GWAS data be used to detect adaptation among human complex phenotypes?

16:30 - 16:45 **Sergio Andreu Sánchez (Groningen U.).** Genetic, environmental and intrinsic determinants of the human antibody epitope repertoire.

16:45 - 17:00 **Winona Oliveros Diez (BSC).** Systematic characterization of regulatory variants of blood pressure genes.

17:00 - 17:15 **Lorena Alonso (BSC).** TIGER: The gene expression regulatory variation landscape of human pancreatic islets.

17:15 - 17:30 **Sponsored talk: Lluc Cabús (Floomics).** Plasma cell-free RNA monitoring for the early detection of complex diseases.

17:30 - 19:00 Cocktail and free poster viewing with authors

20:00 - 22:00 Dinner at the Octubre Centre de Cultura Contemporània (https://octubre.cat/)

# *December* 16th

## SESSION III.

9:30 - 10:15 **Invited Lecture: Julio Saez-Rodriguez (Heidelberg U.).** Knowledge-based machine learning to extract disease mechanisms from single-cell multi-omics data.

10:15 - 10:30 **Aida Ripoll Cladellas (BSC).** A fast and robust statistical framework for gene-wise single-cell differential expression meta-analysis.

10:30 - 10:45 **Paola Corbín Agustí (UV).** Genome-scale metabolic models and the discovery of new metabolic pathways.

10:45 - 11:00 **José Camacho Páez (Granada U.).** Improved statistical inference in omics analysis with Variable-selection Anova Simultaneous Component Analysis (VASCA).

11:00 - 11:30 Coffee Break

## SESSION IV.

11:30 - 11:45 **Ferriol Calvet Riera (CRG).** A universal protein-coding gene finder.

11:45 - 12:00 **Michal Zasisza Alvarez (UB).** A machine learning approach to RNA-editing.

12:00 - 12:15 **Alejandra González Sánchez (Vall Hebron).** Recombinant characterization of circulating non-polio enterovirus in Europe.

12:15 - 12:30 **Julia Hillung (I2SysBio).** Accumulation dynamics of DVGS during experimental evolution of betacoronaviruses.

12:30 - 12:45 **Miguel Álvarez-Herrera (I2SysBio).** Real-time surveillance of SARS-CoV-2: effects of variation in the spike N-terminal domain of variants of concern

12:45 - 13:00 **Sponsored talk: Evan Floden (Seqera Labs).** Building foundational open science tools for the data analysis life cycle**.**

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

## SESSION V.

14:30 - 15:15 **Invited Lecture: Yolanda Sanz (IATA).** Gut microbes - many ways of speaking to the human host

15:15 - 15:30 **Trishla Sinha (Groningen U.).** Lifelines NEXT: temporal development of the gut microbiome in relation to health and environment in 713 mother–infant pairs

15:30 - 15:45 **Olfat Khannous (BSC).** Microbiome profiling from fecal immunochemical test reveals microbial signatures with potential for colorectal cancer screening.

15:45 - 16:00 **Giuseppe D'auria (FISABIO).** The impact of third generation sequencing on the study of microbial diversity: filling the short-reads gap

16:00 - 16:30 Coffee Break

## SESSION VI.

16:30 - 16:45 **Ana Conesa (I2SysBio).** Benchmarking of Long read transcriptomics methods for transcriptome identification and quantification: the LRGASP project.

16:45 - 17:00 **Gazaldeep Kaur (CRG).** Digging into the hidden layer of the human and mouse transcriptomes

17:00 - 17:15 **Francisco J. Silva (UV).** Tissue specific expression of antimicrobial peptide genes in *Blattella Germanica*

17:15 - 17:30 Award to the best oral communication and poster and end of the symposium.

# ABSTRACTS

# ABSTRACTS

*Oral presentations*

## O-001: LONG-TERM INVERSION RECURRENCE AND SEGMENTAL DUPLICATION CONSERVATION DURING MAMMALIAN EVOLUTION

Maria Diaz-Ros[1], Mario Cáceres[1,2]

1. Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain
2. ICREA, Barcelona, Spain
Presenter e-mail: maria.diaz07@estudiant.upf.edu

Genome evolution occurs through different processes, such as small sequence changes and larger rearrangements, which are typically considered to be unique. In particular, inversions are an important type of structural variant involved in phenotypic differences, speciation and sex chromosome evolution in diverse organisms. Recently, it has been found that human inversions generated by non-allelic homologous recombination (NAHR) between inverted repeats (IRs) show high levels of recurrence, but few studies have been made in other species, partly due to a lack of available high-quality genome assemblies. To understand better how this recurrence is generated and its potential functional role, in this work we take advantage of the newest genomic data in multiple mammals to investigate in detail IR conservation and inversion recurrence over long evolutionary distances in a large set of 58 human validated inversions mediated by NAHR. By analyzing five relatively close primate species, above half of the inversions showed recurrence. When the analysis was extended to 13 species across the mammalian phylogeny with longer divergence times, we found IR conservation in 18 of the 44 analyzed inversions (41%), corresponding mostly to segmental duplications (SDs) and blocks of repetitive elements, and independent recurrent events were observed for eight of them, especially in chromosome X. Interestingly, conserved SDs tend to show considerable variation among species, but the two copies maintain high identity within each species, suggesting elevated divergence of these regions by sequence exchange and rearrangement. Our results therefore demonstrate that SD conservation and inversion recurrence are relatively frequent during mammalian evolution, opening interesting questions about the nature and functional consequences of this unusual phenomenon.

**O-002: A WORLDWIDE INVASIVE PANGENOME SHEDS LIGHT ON MULTIPLE MOLECULAR EVOLUTIONARY PROCESSES.**

Carles Galià-Camps [1,2], Tilman Schell [3], Cinta Pegueroles [1,2], Marta Pascual [1,2], Xavier Turon [4], Carola Greve [3,*], Carlos Carreras [1,2,*].

1. Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Avinguda Diagonal 643, 08028 Barcelona, Spain.
2. Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona (UB), Barcelona, Spain.
3. LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberganlage 25, 60325 Frankfurt am Main, Germany.
4. Centre d'Estudis Avançats de Blanes (CEAB, CSIC), Accés Cala St. Francesc 14, 17300 Blanes, Girona, Spain.
Presenter e-mail: cgaliacamps@gmail.com

Biological invasions are a major threat to biodiversity. Thus, monitoring genomic features of invasive species is crucial to understand their population structure and adaptation processes in order to plan future management measures. However, the genomic resources of invasive species are scarce, hampering the study of the molecular bases for their success. Here, we present the pangenome of *Styela plicata*, one of the most widespread invasive species, by using whole genome sequencing data from 24 individuals from six populations distributed worldwide. Thanks to the newly generated reference genome, we characterized polymorphic inversions in four chromosomes, accounting for ~15% of the genome size. These inversions are spread over populations, and contain shared variants which prevent the detection of population structure, which is recovered when these structural variants are removed from the analyses. Once properly assessed, we detected the main biogeographic barriers, and could characterize each region's population structure. Moreover, we recovered three major mitogenomic clades, one of them with a large insertion of 1000pb with several partial genes duplicated. Finally, we tracked down the gene functions 1) in each chromosomal inversion, 2) driving genetic differentiation between biogeographic regions, and 3) underlying mito-nuclear interactions. Our results suggest that the structural and genetic richness yielded by this species might be responsible for its success when invading new habitats. The pangenome presented here will allow a better comprehension of invasive species, and is a starting point for understanding how intraspecific genomic and structural variants are related to species population structuring and adaptation processes.

## O-003: NORTH AFRICA DEMOGRAPHIC HISTORY THROUGH AN ABC-DL APPROACH

Jose Miguel Serradell[1], Oscar Lao[1] and David Comas[1]

1. Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Departament de Medicina i Ciències de la Vida, Carrer del Doctor Aiguader 88, Barcelona 08003, Spain
Presenter e-mail: josemiguel.serradell@upf.edu

Previous studies conducted on North African human populations have identified a complex demographic scenario in the region due to the location as a crossroad between the Mediterranean Sea, Middle East and the Sahara desert. The presence of an autochthonous genetic component, plus extensive migrations and gene flow from the Middle East, Europe and Sub-Saharan Africa, have shaped the genetic composition of its people through time.
We built a demographic model based on deep learning in an Approximate Bayesian Computation (ABC-DL) framework to deduce the evolutionary history of North Africa populations, from the origin of the population to infer the effect of the admixture events from the surrounding populations, using whole genome data. The ABC framework eludes the need of estimating the likelihood of the data given the demographic model, like in traditional Bayesian methods, which could become too difficult in complex demographic scenarios like the one in North Africa. Our results support a model where North Africans present a most recent common ancestor with Middle Easterns and that increasingly complex models, with multiple demographic events, perform better than simpler ones in explaining the data.

**O-004: EXPLORING GENETIC ADAPTATION IN THE IBERIAN PENINSULA USING A GENOME WIDE SCAN OF POSITIVE SELECTION IN THE GCATDATA**

Jorge García[1], Xavier Martí-Pérez[1], Rafael de Cid[2], Francesc Calafell[1], Elena Bosch[1]

1. Institute of Evolutionary Biology (UPF-CSIC), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain
2. Genomes for Life-GCAT lab Group, Institute for Health Science Research Germans Trias i Pujol (IGTP), Badalona 08916, Spain
Presenter e-mail: jorge.garciac@upf.edu

Positive (adaptive) selection leaves particular patterns of genetic variation in our genomes. These genomic footprints can be detected at genome-wide level without requiring any prior knowledge of the selective pressures at play and independently of their corresponding genetic content. Here, we analyzed signatures of recent positive selection on the Catalan Genome (GCAT) dataset, a dataset of 800 complete genomes, sequenced at ~30X coverage from volunteers residing in Catalonia (Valls-Margarit et al. 2022), which we use as a proxy for the Iberian population. With the Singleton Density Score statistic, we detected some well-known European adaptive signals related to diet (*SLC22A5*), pigmentation (*OCA2*) and immunity (*HLA*) replicating in the Iberian Peninsula. However, when comparing the extended haplotype homozygosity of the Iberian population with that of other Europeans with the XP-EHH statistic, we also detected significant adaptive signals private of Iberians in the immunologic genes *LILRA1*, *LILRA2* and *LILRB1*. This result suggests an Iberian-specific pathogen-driven selection pressure. By using genealogy-based methods, such as Relate and CLUES, we now plan to date and trace each selective event and the specific trajectories of the subjacent adaptive variants, to investigate whether they coincide with past pandemics in the Peninsula.

Reference:

Valls-Margarit et al. (2022). GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. *Nucleic Acids Research* 50(5): 2464–2479, https://doi.org/10.1093/nar/gkac076

## O-005: CAN GWAS DATA BE USED TO DETECT ADAPTATION AMONG HUMAN COMPLEX PHENOTYPES?

Francesc Ganau[1]

1. Center for Genomic Research (CRG), Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain
Presenter e-mail: francesc.ganau@cnag.crg.eu

Statistical tests aiming to detect fingerprints left by past strong selective pressures on human genomes have been successfully used to detect several loci associated to recent human adaptation (<100,000 years). However, linking such events to particular phenotypes is more contentious. Furthermore, the power for detecting positive selection is potentially limited for complex phenotypes, as these are the product of a large number of variants with low biological relevance for the phenotype. Genome-Wide Association Studies (GWAS) provide statistically powerful phenotype-genotype associations for human complex phenotypes, and they have been suggested as an alternative for identifying candidate loci that can be used to study complex phenotype adaptation.
However, GWAS are more powerful at identifying markers at intermediate frequencies and low biological impact in the phenotype in the discovery population. A basic question is then to which extend such bias can affect the conclusions obtained from statistical tests used for positive selection.
Our study evaluates how useful are GWAS data when studying how adaptation shaped worldwide human inter-population differentiation. We used hair colour as our proof-of-concept phenotype, as 1) published evidence suggests hair colour was differentially selected in Europe in recent Prehistory and henceforth the phenotype shows large differences between human populations, 2) large-scale GWAS support it being highly polygenic, yet 3) it behaves as a quasi-Mendelian trait, with few loci counting for large proportion of the variance of the trait and a large number of loci contribute in a modest, almost biologically negligible way.
To study the genetic variation of this trait, we used the GWAS-significant loci from the study published in Nature in 2018 by Morgan et al, and retrieved the genètic variation of these loci in 26 worldwide populations available in 1000Genomes.
First, we observed that, contrary to what would be expected, statistically significant GWAS loci with large biological impact in the phenotype tends to be poorly differentiated among populations, whereas markers explaining modest proportions of the variance of the phenotype tend to strongly variate among populations. This paradox has to be interpreted in terms of the initial bias introduced by the GWAS SNP discovery. Second, if instead of using measures of genetic differentiation as proxy for positive selection for these loci we consider selective signals of iHS and Fay & Wu's H, we see that population clustering by hair-significant loci does not differ from clustering by neutral sites matched by MAF. Hence, caution is required when interpreting classical signals of positive selection using GWAS data, as even complex traits with large inter-population variance cannot produce significant signals using this approach.

## O-006: GENETIC, ENVIRONMENTAL AND INTRINSIC DETERMINANTS OF THE HUMAN ANTIBODY EPITOPE REPERTOIRE

Sergio Andreu-Sánchez[1,2]*, Arno R. Bourgonje[3]*, Thomas Vogl[4,5,6], Alexander Kurilshikov[1], Sigal Leviatan[4,5], Angel J. Ruiz Moreno[1], Shixian Hu[1,3], Trishla Sinha[1], Arnau Vich Vila[1,3], Ranko Gacesa[1,3], Elisabeth Brouwer[7], Adina Weinberger[4,5], Cisca Wijmenga[1], Gerard Dijkstra[3], Eran Segal[4,5], Rinse K. Weersma[3]*, Jingyuan Fu[1,2]*, Alexandra Zhernakova[1]*

1 Department of Genetics, 2 department of Pediatrics, 3 department of Gastroenterology and Hepatology, 7 department of of Rheumatology and Clinical Immunology, University of Groningen, University Medical Center Groningen.
4 Department of Computer Science and Applied Mathematics, 5Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel
6 Diagnostic and Research Institute of Hygiene, Microbiology and Environmental Medicine, Medical University Graz, Graz, Austria
Presenter e-mail: s.andreu.sanchez@umcg.nl

The human antibody repertoire can be perceived as a journal that records an individual's current and past immune responses to thousands of antigens that they have been exposed to. Currently, we have little knowledge about genetic and environmental determinants shaping the human antibody repertoire and its relation with disease, especially in the context of immune response to the gut microbiome. To fill this knowledge gap, we applied the state- of-the-art phage immunoprecipitation sequencing (PhIP-Seq) technology to characterize serological antibody repertoires against 344,000 peptides derived from commensal gut microbes, pathogens and dietary antigens in a broadly phenotyped, population-based cohort (LifeLines-DEEP, n=1,443) and a patient cohort with inflammatory bowel disease (IBD) (1000IBD, n=497).

We demonstrated that the antibody repertoire is individual specific, consistent over time (4 years follow-up) and is similar within individuals who are co-housing. We identified co-occurring networks of antibody-bound peptides with phylogenetically distinct origins, including those from gut microbiome, but with highly conserved motifs which might highlight a role of residing microbiota in the development of immune diseases via bacterial mimicry. Genetic analyses showed multiple human genetic variants shaping the antibody repertoire of the population, mainly at *HLA*, *IGHV* and *FUT2* regions. In addition, phenotypic factors including age, cell counts, sex, smoking behavior and allergies, were associated to 544 antibody-bound peptides. In patients with IBD, a total of 373 peptides appeared to be significantly enriched or depleted in IBD patients in comparison to population controls. These peptides mainly belonged to bacterial flagellins, were predominately enriched in patients with Crohn's disease and were able to accurately discriminate Crohn's disease patients from population controls (AUC = 0.89).

Our results indicate that human antibody epitope repertoires are shaped by host genetics, commensal microbes and environmental exposures like co-housing and smoking. We further highlight the role of previously undescribed antigens in relation to diseases, suggesting the potential value of measuring antibody-bound peptides for disease diagnostics and understanding their pathophysiological implications.

## O-007: SYSTEMATIC CHARACTERIZATION OF REGULATORY VARIANTS OF BLOOD PRESSURE GENES

Winona Oliveros[1+], Kate Delfosse[2+], Daniella F. Lato[2], Katerina Kiriakopulos[2,5], Milad Mokhtaridoost[2], Abdelrahman Said[2], Brandon J. McMurray[2], Jared W.L. Browning[2,5], Kaia Mattioli[3], Guoliang Meng[4], James Ellis[4,5], Seema Mital[2,6,7], Marta Melé[1*+], Philipp G. Maass[2,5*+]

1. Life Sciences Department, Barcelona Supercomputing Center, 08034 Barcelona, Catalonia, Spain
2. Genetics & Genome Biology Program, The Hospital for Sick Children, Toronto, Ontario, M5G 0A4, Canada
3. Division of Genetics, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA
4. Developmental and Stem Cell Biology Program, The Hospital for Sick Children, Toronto, Ontario, M5G 0A4, Canada
5. Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
6. Ted Rogers Centre for Heart Research, Toronto, Ontario, M5G 1X8, Canada
7. Department of Pediatrics, The Hospital for Sick Children, University of Toronto, Toronto, Ontario, M5G 0A4, Canada
+ authors contributed equally
* corresponding authors: Philipp.maass@sickkids.ca; marta.mele@bsc.es
Presenter e-mail: winona.oliveros@bsc.es

High blood pressure (BP) is the major risk factor for cardiovascular disease. Genome-wide association studies have identified genetic variants for BP, but functional insights into causality and related molecular mechanisms lag behind. We functionally characterize 4608 genetic variants in linkage with 135 BP loci in vascular smooth muscle cells and cardiomyocytes by massively parallel reporter assays. Some loci harbor hundreds of regulatory variants (i.e. ULK4, MAP4, CFDP1, PDE5A) suggesting that multiple variants may drive genetic association. Regulatory variants are enriched in repeats and alter cardiovascular-related transcription factor motifs. We develop an heuristic scoring system to define likely causal variants for BP genes at GWAS loci. Modeling allelic ratios with CRISPR prime editing we are able to link regulatory variants to KCNK9, SNFX2 and PCGF6 target genes that we further confirm using higher-order genome organization data. Our systems-level approach provides a catalogue of functionally relevant variants and their genomic architecture for a better understanding of blood pressure gene regulation.

**O-008: TIGER: THE GENE EXPRESSION REGULATORY VARIATION LANDSCAPE OF HUMAN PANCREATIC ISLETS**

Lorena Alonso[1,*], Anthony Piron[2,3,*], Ignasi Morán[1,*], Marta Guindo-Martínez[1], Sílvia Bonàs-Guarch[4,5], Goutham Atla[4,5], Irene Miguel-Escalada[4,5], Romina Royo[1], Montserrat Puiggròs[1], Xavier Garcia-Hurtado[4,5], Mara Suleiman[6], Lorella Marselli[6], Jonathan L.S. Esguerra[7], Jean-Valery Turatsinze[2], Jason M. Torres[8,9], Vibe Nylander[10], Ji Chen[11], Lena Eliasson[7], Matthieu Defrance[2], Ramon Amela[1], MAGIC[24], Hindrik Mulder[12], Anna L. Gloyn[9,10,13,14,15], Leif Groop[7,12,16], Piero Marchetti[6], Decio L. Eizirik[2,17], Jorge Ferrer[4,5,18], Josep M. Mercader[1,19,20,21,*], Miriam Cnop[2,22,*], and David Torrents[1,23,*]

1. Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain
2. ULB Center for Diabetes Research, Université Libre de Bruxelles, Brussels 1070, Belgium
3. Interuniversity Institute of Bioinformatics in Brussels (IB2), Brussels 1050, Belgium
4. Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain
5. Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM) Barcelona 08013, Spain
6. Department of Clinical and Experimental Medicine and AOUP Cisanello University Hospital, University of Pisa, Pisa 56126, Italy
7. Unit of Islet Cell Exocytosis, Lund University Diabetes Centre, Malmö 214 28, Sweden
8. Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK
9. Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7LF, UK
10. Oxford Centre for Diabetes, Endocrinology, and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 7LE, UK
11. Exeter Centre of Excellence for Diabetes Research (EXCEED), University of Exeter Medical School, Exeter EX4 4PY, UK
12. Unit of Molecular Metabolism, Lund University Diabetes Centre, Malmö 214 28, Sweden
13. Division of Endocrinology, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94304, USA
14. NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford OX3 7DQ, UK
15. Stanford Diabetes Research Centre, Stanford University, Stanford, CA 94305, USA
16. Finnish Institute of Molecular Medicine Finland (FIMM), Helsinki University, Helsinki 00014, Finland
17. WELBIO, Université Libre de Bruxelles, Brussels 1050, Belgium
18. Section of Epigenomics and Disease, Department of Medicine, Imperial College London, London SW7 2AZ, UK
19. Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
20. Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
21. Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
22. Division of Endocrinology, Erasmus Hospital, Université Libre de Bruxelles, Brussels 1070, Belgium

23. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain
24. Members of the MAGIC consortium
*. These authors contributed equally
Presenter e-mail: lorena.alonso@bsc.es

The broad use of Genome Wide Association Studies (GWAS) have facilitated the discovery of a large list of variants associated with complex disorders. However, the vast majority of these signals still remain without functional interpretation, thus, representing a challenge for the research community. To approach this problem, gene expression variation and regulatory regions analyses have enhanced the detection of numerous associations between variants and their change in expression, and the creation of wide catalogues of disease-related regulatory elements, such as enhancers and promoters. Remarkably, the transcriptomic and epigenetic study of disease-related tissues facilitates the understanding of the molecular mechanisms underlying GWAS disease-susceptibility loci. This is the case of pancreatic islets from which progressive failure and dysfunction is related to the development of Type 2 Diabetes (T2D). Despite its interest to gain insights of the disease, the many complexities surrounding the accessibility and analysis of pancreatic islets has limited the advance on the study of the genetic and regulatory landscape of human islets and T2D. Here, within the context of the T2DSystems, a Horizon2020 Project, we developed the Translational human Islet Genotype tissue-Expression Resource (TIGER), a large human islets regulatory expression database (http://tiger.bsc.es). This database integrates, in a unique platform, the results obtained from the performance of extensive expression, expression quantitative trait loci (eQTL), and combined allele-specific expression (cASE) in 514 human islets samples, with publicly available summary statistics results from islets analyses, including expression array, regulatory elements, and other gene, variant, and disease functional information. As a result, the TIGER platform contains information for more than 27 million variants and 59,625 genes, and facilitates the search at the level of the gene and at the level of the variant. It encloses tools for visualising, querying, and downloading human islet data enhancing the study of T2D and other islet-related diseases. Therefore, representing a formidable resource to interrogate the molecular aetiology of beta-cell failure.

## O-009: PLASMA CELL-FREE RNA MONITORING FOR THE EARLY DETECTION OF COMPLEX DISEASES

Lluc Cabus[1], Cristina Tuñi[1], Cristina Hernando-Davalillo[2], Adrián Alcalá[2], Mercè Bolasell[2], Patricia Fernández-Pareja[2], Gisela Márquez[2], Gregorio A. Nolasco[3], Joao Curado[1], Julien Lagarde[1], Jennifer Perez-Boza[1], Mercedes Serrano[3], Esther Lizano[1]

1. Flomics Biotech, PRBB Building, Doctor Aiguader 88, 08003 Barcelona (Spain)
2. Department of Genetic and Molecular Medicine and Pediatric Institute of Rare Diseases, Hospital Sant Joan de Déu, Barcelona, Spain.
3. Neuropediatric Department, Hospital Sant Joan de Déu, Barcelona, Spain. U-703 Centre for Biomedical Research on Rare Diseases (CIBER-ER), Instituto de Salud Carlos III, Barcelona, Spain.
Presenter e-mail: lluc.cabus@flomics.com

Cell-free RNA (cfRNA) is released into the circulation by active secretion, apoptosis or necrosis. In blood, cfRNA is either encapsulated inside microvesicles or forming ribonucleoprotein complexes, increasing its stability. Traditionally, cfRNA has been used as an analyte for cancer detection, but since it can derive from any tissue, its analysis can be applied to a wide spectrum of diseases. MicroRNAs (miRNAs) are especially useful, due to their high stability in plasma.
PMM2-CDG is a rare disease where patients develop a cerebellar syndrome, causing long-term disability. It is often detected during childhood and when the symptoms are detected, its progression is irreversible. Therefore, early diagnosis is crucial to improve patient prognosis.
Here we analyze the plasma circulating miRNome to generate a machine learning (ML) model able to discriminate between PMM2-CDG patients and healthy controls with an AUC higher than 0.9 in both training and validation cohorts. Using regularization methods and backward stepwise selection, we achieve a minimal signature with great diagnostic value.
This is the first study proposing the detection of plasma miRNAs as a non-invasive method for the early detection of PMM2-CDG, offering a non-invasive diagnosis for children with presymptomatic neurological diseases.

## O-010: A FAST AND ROBUST STATISTICAL FRAMEWORK FOR GENE-WISE SINGLE-CELL DIFFERENTIAL EXPRESSION META-ANALYSIS

Aida Ripoll-Cladellas[1], Monique G.P. van der Wijst[2], Marc Jan Bonder[2,3,4], single-cell eQTLGen consortium, Lude Franke[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. European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany

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

In the last decade, there has been a huge effort to build human cell atlases from different tissues using single-cell transcriptomics data. These atlases provide a deep molecular characterization of cell types but often profile many cells from one or a few donors, thus hampering the study of inter-individual variability. Recent technological advances enable single-cell transcriptome sequencing of many cells from multiple individuals offering the opportunity to investigate cell-type-specific gene expression responses to individual traits (such as sex or age). However, most current single-cell methods do not address the hierarchical structure (i.e., cells from the same individual are not statistically independent; referred to as subsamples, or pseudoreplicates) of single-cell data. Here, we propose to use two-part Hurdle generalized linear models with random effects for individuals to account for both zero inflation and within-sample correlation in single-cell RNA-seq data. Towards this end, we develop a statistical framework to conduct a gene-wise single-cell differential expression meta-analysis across several peripheral blood mononuclear cells datasets from the single-cell eQTLGen consortium. Our approach increases the statistical power to find cell-type-specific genes differentially expressed with sex and age while reducing the running times and memory consumption of single-cell mixed models. Altogether, this study addresses the pseudoreplication bias in single-cell differential expression analysis and highlights the need to “bring the algorithm to the data” in the era of upcoming consortia.

## O-011: GENOME-SCALE METABOLIC MODELS AND THE DISCOVERY OF NEW METABOLIC PATHWAYS

Paola Corbín-Agustí[1], Arián Ferrero-Fernández[1], Miguel Álvarez-Herrera[1], Juli Peretó[1,2]

1. Institute for Integrative Systems Biology I2SysBio (Universitat de València-CSIC), C. Catedrático Agustín Escardino Benlloch 9, 46980, Paterna, Spain
2. Department of Biochemistry and Molecular Biology, Universitat de València, C. Dr. Moliner 50, 46100, Burjassot, Spain
Presenter e-mail: paola.corbin@uv.es

Genomic-scale metabolic models (GEMRE) based on constraints are useful and robust computational tools for in silico systemic approaches to metabolic networks. The applications of this methodology include the prediction and phenotypic characterization of the modelled biological system and the discovery of new metabolic pathways. We have recently identified two bacteria capable of living on acetate as the sole carbon source lacking any of the previously described metabolic solutions to produce biomass from acetate: the phytopathogen *Xylella fastidiosa* and the polyextremophilic bacterium *Exiguobacterium* sp. Helios isolated from solar panels. In this work, the GEMRE of the two bacteria were reconstructed and analyzed. With this purpose, the four steps of reconstruction were followed: (i) automatic draft reconstruction, (ii) manual refinement, (iii) conversion into a computable mathematical model, and (iv) network evaluation and exploitation. The reconstruction of the GEMREs suggests two alternative and different, yet undescribed pathways for the growth on acetate. A series of experimental validations are under way in order to confirm the two hypotheses. In conclusion, we observe a widely multiplicity of pathways that make possible to live on acetate and how the models are a suitable basis for understanding fundamental metabolic features, thus increasing biological knowledge.

Acknowledgements. We are grateful to ADM-Biopolis, AEI-FEDER (grant RT RTI2018- 095584-B-C41), CSIC (PIE202020E120) and Iber-Xyfas network (CYTED 119RT0569) for financial support. PC-A is a recipient of an ACIF predoctoral contract (Generalitat Valenciana).

## O-012: IMPROVED STATISTICAL INFERENCE IN OMICS ANALYSIS WITH VARIABLE-SELECTION ANOVA SIMULTANEOUS COMPONENT ANALYSIS (VASCA)

José Camacho[1], Oliver Aleksandrei Polushkina, Carolina Gómez-Llorente[1], Caridad Díaz[1], Pedro Sánchez-Rovira[2]

1. Universidad de Granada, Granada, Spain
2. Fundación MEDINA Centro de Excelencia en Investigación de Medicamentos Innovadores en Andalucía, Granada, Spain
3. Medical Oncology Unit, University Hospital of Jaén, Jaén, Spain
Presenter e-mail: josecamacho@ugr.es

Statistical inference is ubiquitous in genomics and other -omics. The state-of-the-art is based on the procedure to control the the False Discovery Rate (FDR) by Benjamini and Hochberg [1] and the optimized version of the FDR by Storey and Tibshirani [2], often referred to as the q-value. Both methods are widely popular (> 90k and >10k references, respectively, according to Google Scholar) and are often referred to as main promoters of the widespread application of genomics in science.

The FDR and the q-value are suitable inference tools when large numbers of potential features (genes, SNPs, protein/metabolite abundances, etc.) are studied. The FDR or q- value is computed from the original p-value of a feature considering the number or distribution of the other p-values. Given that a p-value is intrinsically univariate, recomputation according to the FDR or the q-value remains univariate, and therefore it does not take into account the possibly complex multivariate structure underlying the data.

A recent multivariate alternative to the FDR was proposed in [3]. The Variable-selection ANOVA Simultaneous Component Analysis (VASCA) method is an extension of ASCA that applies variable-selection using a step-up procedure inspired in [1]. In this presentation, we introduce the VASCA methodology and show that it provides higher statistical power than alternative methods in several (multi)-omics studies.

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300.
[2] Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16), 9440-9445.
[3] Camacho, J., et al. (2022). Variable-selection ANOVA Simultaneous Component Analysis (VASCA). Submitted to Bioinformatics
[4] Smilde, A. K. et al. (2005). ANOVA-Simultaneous Component Analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics, 21(13), 3043–3048.

## O-013: A UNIVERSAL PROTEIN-CODING GENE FINDER

Ferriol Calvet[1,2], Jaume Reig[1], Emilio Righi[1], Francisco Cámara[1], Roderic Guigó[1,2]

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

The Earth Biogenome Project will sequence the genome of 1.8M eukaryotes. Identifying the genes in these genomes is essential to uncovering the species' biology and increasing our knowledge of life on Earth. To capture the full transcriptional complexity, gene annotation methods incorporate deep RNA sequencing into complex pipelines. These require substantial resources, available only at a few sites, and are only partially successful, as illustrated by the fact that the human gene set has not yet been finalised. Moreover, functional inference on the biology of genomes can only be made from coding genes. Given the strong imprint that coding regions leave in genome sequences, *ab initio* methods can produce decent predictions of coding genes without the need for transcriptome data. We have shown that when considering all splice isoforms of a protein-coding gene, there is only marginal gain in the functional assignment over considering just one. Considering these results, we used the program Geneid in conjunction with a non-redundant set of proteins, and Diamond for protein-DNA comparisons, to efficiently predict the dominant isoform of protein-coding genes. We show that our method, GeneidX, is comparable to top *ab initio* programs, but orders of magnitude faster in vertebrates' and arthropods' genomes annotated in Ensembl. Our implementation automatically estimates parameters required by Geneid making GeneidX a training-free program. The minimal computational requirements and its portability and scalability make GeneidX carbon footprint-aware and address the equity concerns in genome diversity projects, contributing to empowering local communities to perform genome analysis. Also, we are running our pipeline to generate the genome annotations for all the species under the EBP umbrella and we are distributing them publicly through a data portal.

## O-014: A MACHINE LEARNING APPROACH TO RNA-EDITING

Michał Zawisza Alvarez[1], Ferran Reverter Comes[1], Esteban Vegas Lozano[1], Jordi Garcia Fernández[1], Carlos Herrera Úbeda[1]

1. Departament de Genètica, Microbiologia i Estadística, Facultat de biología, Universitat de Barcelona
Presenter e-mail: m.zawisza@ub.edu

RNA-editing is a post-transcriptional modification of the RNA where particular nucleotides undergo a chemical change, thus changing the original sequence. Together with alternative splicing, RNA-editing is one of the mechanisms that allows for a flexible genome with a single gene being able to produce different mature transcripts. While there are other types of editing across Eukaryotes, our focus is on the adenosine to inosine editing mediated by the ADAR protein family, that is the most prominent form of editing in metazoans. Synthesis-based RNA sequencing techniques cannot directly detect inosines, appearing as guanosines in the sequencing output. A standard way of detecting editing involves comparing the reads of DNA and RNA sequencing from the same individual organism in order to confidently classify the A-to-G changes detected as polymorphisms or A-to-I editing. This approach relies on the availability of RNA and DNA reads from the same individual, which may not be readily available and may require specific sequencing process. Thus, an alternative approach for genome-wide detection of RNA-editing, requiring only the RNA sequences, could be useful for certain cases. We will present several results of a novel machine-learning approach for A-to-I RNA-editing detection, based on the RNA secondary structure, as well as a comparison between using neural networks and using a random forest algorithm for this approach. We think this provides an alternative method for A-to-I RNA-editing that also reveals clues on how ADAR selects which adenosines will be edited.

## O-015: RECOMBINANT CHARACTERIZATION OF CIRCULATING NON-POLIO ENTEROVIRUS IN EUROPE

Alejandra González-Sánchez[1,2], Cristina Andrés[1,2], Maria Piñana[1,2], Narcís Saubí[1,2], Manuel García-Ulloa[3], Andrés Santos[3], Jaime Martínez-Urtaza[3], Andrés Antón[1,2]

1 Respiratory Viruses Unit, Virology Section, Microbiology Department, Vall d'Hebron Hospital Universitari, Vall d'Hebron Institut de Recerca (VHIR), Vall d'Hebron Barcelona Hospital Campus, Passeig Vall d'Hebron 119-129, 08035 Barcelona, Spain
2 Centro de Investigación Biomédica en red de Enfermedades Infecciosas CIBERINFEC, Instituto Carlos III, Madrid, Spain.
3 Department of Genetics and Microbiology, Universitat Autònoma de Barcelona, Barcelona, Spain.
Presenter e-mail: alejandra.gonzalez@vallhebron.cat

Enteroviruses are non-enveloped viruses with a positive single-stranded RNA genome, which belong to the *Picornaviridae* family. Those infecting humans are classified in four species, enterovirus A, B, C and D. Among the main forces shaping the evolution of enteroviruses, recombination and point mutations can be observed at high rates. Recombination is not only important for relevant antigenic changes and hence, its association with epidemic waves, but also with worsening the clinical outcomes. Nevertheless, little is known about the underlying recombination mechanism. In this study, we analysed 737 whole genome sequences (351 EV-A, 161 EV-B, 24 non-polio EV-C, 201 EV-D) sampled between 1966 and 2022 from 25 European countries, of which 63 were sequenced and assembled in our hospital, with the aim of detecting and characterizing recombination events. Our bioinformatic analysis identified 3 recombination hotspots, one at the beginning of the 5' UTR region and two flanking the capside proteins, and 1 recombination cold spot spanning almost the entire region of the capside proteins in circulating enteroviruses across Europe. These findings could help elucidate recombination dynamics, deepening the genetic mechanism of evolution of these viruses associated with a high morbidity and mortality. The location of hotspots may suggest capsid proteins exchange between enterovirus types, allowing binding to a wider range of tissue receptors, in addition to changes in antigenic features, and thus, leading to more severe clinical outcomes.

## O-016: ACCUMULATION DYNAMICS OF DVGS DURING EXPERIMENTAL EVOLUTION OF BETACORONAVIRUSES

Julia Hillung[1], María J. Olmo-Uceda[1], Santiago F. Elena[1,2]

1. Institute for Integrative Systems Biology (I2SYSBIO), CSIC-UV, C/Catedrático Agustín Escardino Benlloch 9, 46980 Paterna (Valencia)
2. Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, United States of America
Presenter e-mail: julia.hillung@csic.es

Replication of RNA genomes by virus-encoded RNA dependent RNA Polymerases (RdRP) often results in hypermutated or in truncated versions of genome (DVGs). Coinfection with a helper virus that provides missing protein(s) in trans, may result in amplification and transmission of DVGs. DVGs depend on helper virus for propagation, but in coinfection they can interfere in the replication of infectious virus, cause faster immune responses and also promote viral persistence or/y evolution. However, dynamics of accumulation of DVGs during viral infection remains elusive.

To address this question, we perform an experimental evolution of betacoronaviruses Human Corona Virus OC43 (HCoV-OC43) and Murine Hepatitis Virus (MHV) in cell culture and at high and low multiplicity of infection (MOI) conditions. Subsequently we performed high throughput sequencing of equidistant time series of resulting viral populations. We reconstructed DVGs from obtained short reads and their accumulation among passages were analyzed.

We found that DVGs evolved to more net diversity and more abundance. Most abundant DVG types were deletions and insertions. Especially at high MOI some deletions had only transitory existence, while other won at abundance. Genomes of HCoV-OC43 carrying deletions had bimodal size distribution at low MOI, with short and long genomes being common and almost no changes in their size distribution during the viral evolution. In samples at high MOI, the size distribution was multimodal shifted over time to longer DVGs.

In MHV genomes with deletions the size distribution shifted from bimodal distribution to unimodal distribution, being large genomes most common as well as at high as at low MOI. Hot-spot regions for deletions evolved at high MOI were found and they stretched mainly over regions of structural and accessory proteins.

In conclusion, we showed the pervasive formation of DVGs during betacoronavirus evolution, with the dynamics being MOI- and virus-dependent.

## O-017: REAL-TIME SURVEILLANCE OF SARS-COV-2: EFFECTS OF VARIATION IN THE SPIKE N-TERMINAL DOMAIN OF VARIANTS OF CONCERN

Miguel Álvarez-Herrera[1], Paula Ruiz-Rodríguez[1], Beatriz Navarro-Domínguez[1], Joao Zulaica[1], Daniel del Hoyo[4], María Alma Bracho[1,2], Álvaro Chiner-Oms[3], Ron Geller[1], Carlos Óscar S. Sorzano[4], Fernando González-Candelas[1,2], Iñaki Comas[2,3], Mireia Coscollá[1]

1. Institute for Integrative Systems Biology, FISABIO joint unit (I2SysBio-CSIC-UV, FISABIO)
2. CIBER in Epidemiology and Public Health (CIBERESP)
3. Institute of Biomedicine of Valencia (IBV-CSIC)
4. National Centre for Biotechnology (CNB-CSIC)
Presenter e-mail: m.alvarez.herrera@csic.es

Genomic surveillance has become a powerful asset to monitor SARS-CoV-2 evolution in real-time and detect the emergence and spread of lineages of concern. In 2021 alone, sudden global takeover by novel variants of concern (VOC) happened three times, starting with the Alpha variant, then Delta, and lastly Omicron. Interestingly, these three variants belong to different clades that diverged early in the pandemic. In fact, Alpha and Delta only share one substitution on the S gene, which gathers the most diversity. Hence, the need to understand the generation of variability to tackle the current state and future progression of the pandemic. Here we show the co-occurrence of mutational markers located at the N-terminal end of the S gene from the B.1.1.7 lineage (Alpha VOC) the B.1.617.2 lineage (Delta VOC) and the B.1 lineage (Omicron VOC). We have defined a series of haplotypes containing marker mixtures, identified haplotype-matching sample on a worldwide dataset, and assessed the abundance of each genotype, their spatiotemporal distribution, and their phylogenies in their local geographic context. To address the impact of each combination of markers on their fitness, we have also estimated the minimum number of independent emergences and their implication in transmission events. The performance of certain haplotypes does not fit neutrality when considering their relative abundance, the expected emergence rate, and genetic distance. Additionally, we have assessed differences in infectivity, antibody escape, cell-cell fusion, thermal stability and *in silico* molecular dynamics attributable to the remodeling of the N-terminal domain of the Omicron VOC. Strikingly, some combinations of mutations have never been sequenced. However, the most abundant ones have emerged multiple times, and been implicated in transmission events, with one of them clearly outperforming the rest. The lack of some of the marker mixtures point to strong epistatic interactions between said markers. Our framework is ready to be implemented for any other context, genomic region, or variants of interest.

## O-018: LIFELINES NEXT: TEMPORAL DEVELOPMENT OF THE GUT MICROBIOME IN RELATION TO HEALTH AND ENVIRONMENT IN 713 MOTHER–INFANT PAIRS

Trishla Sinha[1], S. Brushett[1,2], A. Fernandez-Pato[1], S. Garmaeva[1], M. F. Brandao Gois[1], J. Spreckels[1], S. Andreu-Sánchez[1,3], A. Vich Vila[1,4], R. Gacesa[1,4], A. Kurilshikov[1], S. Jankipersadsing[1], M. Kruk[1], C. Bang[5], A. Franke[5], M. Kroon[4], M. Swertz[1], J. Dekens[1,6], J. Sikkema[6], F. Kuipers[2], M. Reijneveld[4], J. Prins[7], S. Gordijn[7], S. Scherjon[7], M. Yassour[8,9], C. Wijmenga[1], J. Fu[1,2], A. Zhernakova[1]

1. Departments of Genetics, 2. Health Sciences, 3. Pediatrics, 4. Gastroenterology and Hepatology, University Medical Center Groningen, Groningen, Netherlands; 5. Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany; 6. Center of Development and Innovation, 7. Departments of Obstetrics and Gynecology, University Medical Center Groningen, Groningen, Netherlands; 8. Faculty of Medicine and 9. The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel.
Presenter e-mail: t.sinha@rug.nl

The establishment of the gut microbiome in early life plays a crucial role in health. Despite its importance, current knowledge on the environmental factors modulating strain-level colonization and their dynamics over time is limited. To bridge this gap, we collected 4,369 longitudinal fecal samples (starting from 12 weeks of pregnancy up to 1 year after birth) from 713 mother–infant pairs in the Lifelines NEXT cohort, a prospective birth cohort from the Northern Netherlands. Shotgun metagenomic sequencing was used to characterize the gut microbiome at strain-level. From all participants, we acquired >1000 meta-data records, including birth factors, maternal and infant diet, pregnancy complications, medications, allergies, diseases, and environmental exposures. Generalized additive models were used to relate gut microbial features with the available phenotypes. In the first 3 months of life, the overall gut microbiome composition was strongly shaped by birth factors such as mode of delivery, duration of active pushing during birth and number of previous pregnancies. Within-individual alpha diversity increased significantly during the first year of life, was largely shaped by infant feeding mode and medications such as remedies to reduce gastrointestinal gas formation and was associated with stool frequency (FDR<0.05). After introduction of solid food, alpha diversity was mainly influenced by infant diet (FDR<0.05). Co-assembly of longitudinal samples from the same families showed mother to infant transmission of bacterial strains belonging to *Bacteroides fragilis*, *B. uniformis*, *B. dorei*, *Bifidobacterium adolescentis*, *Parabacteroides distasonis* and *Parabacteroides merdae* species throughout the first year of life. In contrast, strain-sharing was observed less frequently for *Bifidobacterium breve* and *Klebsiella variicola*, suggesting different seeding sources for these strains. In vaginally born infants, the dominant strains of *Escherichia coli*, *Bifidobacterium longum* and *B. dorei* were significantly closer to their mother's strains when compared to infants born via cesarian section indicating a disturbed gut-to-gut strain-transmission in infants born by cesarean section. In conclusion, our high-resolution longitudinal characterization of the maternal and infant gut microbiome, coupled with detailed phenotypic data, enabled us to identify novel patterns of microbial transmission, early-life microbiome assembly and the environmental and health factors that shape human gut microbiome development. This provides a better understanding of how the gut

microbiome is established and developed and will pave the way to discovering methods to modify the gut microbiome, such as maternal and infant lifestyle and dietary interventions.

**O-019: MICROBIOME PROFILING FROM FECAL IMMUNOCHEMICAL TEST REVEALS MICROBIAL SIGNATURES WITH POTENTIAL FOR COLORECTAL CANCER SCREENING**

Olfat Khannous-Lleiffe[1,2], Jesse R. Willis[1,2], Ester Saus[1,2], CRIPREV Consortium, Victor Moreno[3], Sergi Castellví-Bel[4,] Toni Gabaldón[1,2,5,6]

1. Barcelona Supercomputing Center (BSC-CNS). Carrer de Jordi Girona, 29, 31, 08034 Barcelona.
2. Institute for Research in Biomedicine (IRB). Carrer de Baldiri Reixac, 10, 08028 Barcelona.
3. Catalan Institute of Oncology, Bellvitge Biomedical Research Institute (IDIBELL), Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP) and University of Barcelona, 08908 Barcelona, Spain.
4. Gastroenterology Department, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Hospital Clínic, University of Barcelona, Barcelona, Spain.
5. Institució Catalana de Recerca i Estudis Avançats (ICREA). Pg. Lluís Companys 23, 08010 Barcelona, Spain.
6. Centro Investigación Biomédica En Red de Enfermedades Infecciosas (CIBERINFEC), Barcelona, Spain.
Presenter e-mail: olfat.khannous@bsc.es

Colorectal cancer (CRC) is the third most common cancer and the second leading cause of cancer deaths worldwide. Early diagnosis of CRC, which saves lives and enables better outcomes, is generally implemented through a two-step population screening approach based on the use of Fecal Immunochemical Test (FIT) followed by colonoscopy if the test is positive. However, the FIT step has a high false positive rate, and there is a need for new predictive biomarkers to better prioritize cases for colonoscopy. Here we used 16S rRNA metabarcoding from FIT positive samples to uncover microbial taxa, taxon co-occurrence and metabolic features significantly associated with different colonoscopy outcomes, underscoring a predictive potential and revealing changes along the path from healthy tissue to carcinoma. Finally, we used machine learning to develop a two-phase classifier which reduces the current false positive rate while maximizing the inclusion of CRC and clinically relevant samples.

**O-020: THE IMPACT OF THIRD GENERATION SEQUENCING ON THE STUDY OF MICROBIAL DIVERSITY: FILLING THE SHORT-READS GAP**

Giuseppe D'Auria[1,3], Elena Buetas[2], Marta Jordán López[4], Andrés López Roldán[4], Griselda De Marco-Romero[1], Llucia Martínez-Priego[1], Alex Mira[2,3] and Miguel Carda-Dieguez[2].

1. Servicio de Secuenciación y Bioinformática, Fundación Para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO-Salud Pública), 46022 València, Spain.
2. Genomics and Health Department, FISABIO Institute, Valencia, Spain.
3. CIBER in Epidemiology and Public Health (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain.
4. University of Valencia, Valencia, Spain
Presenter e-mail: dauria_giu@gva.es

Amplicon sequencing is used to obtain the taxonomic annotation of the bacteria present in a given sample, and the count of the times they are found. From the 1990s until the advent of 2nd generation methods, the subject of sequencing was the whole 16S rDNA gene (~1500bp) enclosed within the historic primers 27F-1492R (*E. coli* coordinates). The process passed through the DNA extraction, gene amplification, the cloning of the amplicons in commercial plasmid, the transformation, clones selection and sequencing of each insert using Sanger method (1st generation). During the first decade of 2000, the appearance of 2nd generation sequencers allowed to skip the cloning step, sequencing directly the amplicons and producing hundreds of thousands of sequences in a single experiment. These approaches were limited in amplicon size reducing the taxonomic "signal" to less than 500 bp (1/3 of the gene sequenced using Sanger sequencing). As a result, we have been populating databases with partial sequences for more than 10 years.
Here we present a comparison of oral and gut communities using short-reads from Illumina and full-length 16S from PacBio sequencing methods. DNAs from saliva, subgingival (dental) plaque and feces of 9 volunteers were extracted. In parallel, regions V3-V4 and V1-V9 were amplified and sequenced by Illumina Miseq and PacBio Sequel II sequencers, respectively. Reads were annotated against SILVA.v138 database using DADA2 and QIIME2 pipelines. PacBio sequencing showed, a higher number of reads assigned to the genus (99.41% vs 93.58% for traditional Illumina V3-V4 sequencing) and to the species level (75.92% vs 55.44%). Regarding the composition, more than 85% of abundant genera (those at >0.1% abundance) were shared between platforms in all types of samples. However, relevant genera such as *Fusobacterium* (6.13% vs 10.14%) were significantly reduced in saliva (adjusted p-value<0.05) using PacBio technology whereas *Streptococcus* (19.85% vs 14.03%) was increased. Our results suggest a high similarity of the overall microbial composition between platforms at the genus level but with some inconsistencies that will need to be evaluated with mock communities. We conclude that, since the advent of third generation sequencers, we have re-gained the possibility to sequence the full length of the 16S rDNA gene, coming back to the first generation standard but with the high-throughput typical of parallel sequencing methods.

**O-021: BENCHMARKING OF LONG READ TRANSCRIPTOMICS METHODS FOR TRANSCRIPTOME IDENTIFICATION AND QUANTIFICATION: THE LRGASP PROJECT.**

Ana Conesa, on behalf of the LRGASP consortium*
Institute for Integrative Systems Biology, Spanish National Research Council, Paterna, Spain.
*LRGASP Consortium: Francisco J. Pardo-Palacios, Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E. Loveland, Matthew S. Adams, Gabriela Balderrama-Gutierrez, Amit K. Behera, Maite De María, Jose M. Gonzalez, Toby Hunt, Julien Lagarde, Haoran Li, Cindy E. Liang, AndreyD. Prjibelski, Leon Sheynkman, David Moraga-Amador, If Barnes, AndrewBerry, Muhammed Hasan Çelik, Natàlia Garcia-Reyero, Stefan Goetz, Liudmyla Kondratova, Jorge Martinez-Tomas, Carlos Menor, Jonathan M. Mudge, Alejandro Paniagua, Marie-Marthe Suner, Hazuki Takahashi, Alison D. Tang, Ingrid Ashley Youngworth, Piero Carninci, Nancy Denslow, Roderic Guigó, Margaret E.Hunter, Hagen U. Tilgner Barbara J. Wold, Christopher Vollmers, Adam Frankish, Kin Fai Au, Gloria M. Sheynkman, Ana Conesa, Ali Mortazavi, Angela N. Brooks.

Long reads, single molecule sequencing methods featured by Pacific Biosystems and Oxford Nanopore, have created new opportunities for the analysis of the transcriptomes as they enable the detection of fulllength transcripts and the identification of alternative isoforms. Novel analysis algorithms have been developed to analyze these data. Using these resources, studies in different species showed that the repertoire of alternative transcripts that are expressed in any tissue is larger than previously anticipated. However, long reads methods have also their limitations. Library preparations do not always render fulllength molecules and the single molecule technologies have higher error rates than Illumina. As these approaches are increasingly being used to define transcriptome catalogues, there is an urgent benchmarking necessity to establish the accuracy of both the technologies and bioinformatics tools for transcriptome detection and quantification.
The LRGASP project is an international initiative for such effort. We generated long reads data in three species, on different biological samples -including spike-ins and simulated data- and using four different library preparation methods and subjected them to sequencing by Pacbio and Nanopore. Data was made public to the bioinformatics community for prediction of transcript models with and without the utilization of a reference annotation, and for quantification. A total of 12 bioinformatics labs submitted nearly 200 transcriptome predictions for evaluation. Quality of the transcriptome identity was evaluated using SQANTI framework and performance metrics were obtained on spike-ins, simulated data, and a set of 50 GENCODE-manually curated loci. A selection of novel transcript predictions was experimentally validated.
Our results indicate that a large diversity in the predicted transcriptome exists when comparing methodologies, both experimental methods and analysis algorithms. Specifically, the detection of novel transcripts seems to be particularly challenging. Moreover, our data reveals that the library preparation protocol, not only the sequencing technology is critical for data quality, being read length and quality, rather than quantity, most important. Our results also reveal that long reads-based quantification is possible when a sufficient number of reads is available, but that accurate identification of the transcriptome without the use of a reference annotation is difficult. Finally, our results confirm the complexity of the expressed transcriptome and suggest that novel strategies should be envisioned to describe the dynamics of gene expression. The

LRGASP is the largest effort to date for the benchmarking of long reads sequencing methods for transcriptome analysis.

## O-022: DIGGING INTO THE HIDDEN LAYER OF THE HUMAN AND MOUSE TRANSCRIPTOMES

Gazaldeep Kaur*[1], Silvia Carbonell-Sala*[1], Carme Arnan[1], Emilio Palumbo[1], Barbara Uszczynska-Ratajczak[2], Julien Lagarde[3], Rugile Dragunaite[4], Roderic Guigó[1]

1. Computational Biology of RNA Processing, Bioinformatics and Genomics programme, CRG-Centre de Regulació Genòmica, Dr. Aiguader, 88, 08003, Barcelona, Catalonia, Spain
2. Computational Biology of Noncoding RNA, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland
3. Flomics, Dr. Aiguader, 88, 08003, Barcelona, Catalonia, Spain
4. Laboratory of Molecular Neurooncology, Neuroscience Institute, Medical Academy, Lithuanian University of Health Sciences, Kaunas, Lithuania

* *Both authors contributed equally*

Presenter e-mail: gazaldeep.kaur@crg.es

A comprehensive transcriptome-scale view is essential for understanding the intricacies of biological functioning at the genomic level. GENCODE, originating from the pilot phase of ENCODE, is an international consortium with an effort towards generating high-confidence annotations for human and mouse genomes. A major effort is towards the improvement and expansion of all gene feature annotations with a special attention to protein-coding, non-coding loci and pseudogenes in these genomes. In this direction, GENCODE develops experimental approaches to target transcripts from genomic regions that have not been well studied by standard transcriptome approaches, while producing high-quality full-length transcript models that require minimal manual editing. Herein, we describe our improved Capture Long Sequencing (CLS) workflow with a motive to increase the representation of these specific transcriptional elements in the sequenced genome to aid their detection. The accurate annotation of these rare elements is fundamental for the downstream studies to correctly assign the functions helping to elucidate the regulatory mechanisms these elements are involved in. In this effort, we designed the broadest catalog that includes poorly annotated but crucial regulatory elements predicted to transcribe for lncrna, putative secondary elements, enhancers and small RNA precursors, giving the most accurate view of the long noncoding transcriptome to date. The improved captrap-CLS methodology, that specifically enriches 5'-capped, full-length rnas, was used to capture and enrich full-length transcripts from this catalog of elements across various important tissues and development stages in human and mouse. We use lyric, an in-house pipeline to further aid the detection of high confidence transcripts as it incorporates experimental information for identification of full-length transcripts. The results indicate a good enrichment of the reads mapping to the target catalog elements after capture, confirming the efficiency of the captrap-CLS methodology. Downstream analysis reveals detection of a set of transcripts that are only detected post-capture, further confirming the efficiency of this method in identification of transcripts that have gone undetected so far. In addition, we discover completely novel full-length transcript models in the largest ever Effort to access the largely unexplored fraction of the human and mice transcriptomes.

## O-023: TISSUE SPECIFIC EXPRESSION OF ANTIMICROBIAL PEPTIDE GENES IN *BLATTELLA GERMANICA*

Francisco J. Silva[1,2], Leo Zuber[1], Rebeca Domínguez-Santos[1,2], Amparo Latorre[1,2], Carlos García-Ferris[1,2,3]

1 Institute for Integrative Systems Biology (I2SysBio), University of Valencia and Spanish Research Council, 46980 Paterna (Valencia), Spain
2. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region, 46020 Valencia, Spain
3. Department of Biochemistry and Molecular Biology, University of Valencia, 46100 Burjassot (Valencia), Spain
Presenter e-mail: francisco.silva@uv.es

The genome of the cockroach *Blattella germanica* harbors a large number of canonical antimicrobial peptide (AMP) genes (39 in total) distributed in five families: attacins, defensins, drosomycins, termicins and blattellicins. The latter is a new protein family evolved in the *Blattella* lineage. The expression of AMP genes is required to control pathogenic infections but also to maintain the healthy microbiota or to potentially control the obligate endosymbiont *Blattabacterium cuenoti* located in the fat body. Here, by analyzing the expression of these AMP genes in six adult tissues (salivary glands, foregut, midgut, hindgut, Malpighian tubules and fat body) and the hemolymph, which contains circulating hemocytes, we have revealed that two samples (salivary glands and hemolymph) display a total AMP gene expression two or three orders of magnitude higher than those observed in the other analyzed tissues, being midgut and hindgut the tissues with the lower expressions. In salivary glands, the high expression would help to control pathogens ingested with the food and would avoid their transit to the midgut or hindgut where they could compete with the commensal/beneficial microbiota. On the other hand, the expression in hemolymph would serve to distribute AMPs to the different organs of the insect. The reason why midgut and hindgut are the tissues with the lower AMP gene expressions could be associated with avoiding damage to the microbiota. The analysis of specific AMP gene expression revealed that some genes were specifically highly expressed in a single sample, while a few display relevant expressions in the seven samples. It was not possible to identify any specific AMP gene whose expression could be related with the control of the endosymbiont in bacteriocytes (a type of fat body cells). Moreover, an additional differential expression analysis of adult fat bodies between control and quasi-aposymbiotic individuals (treated with rifampicin to reduce the endosymbiont load) was performed and analyzed in order to detect noncanonical AMP genes and to identify the most relevant genes involved in the host-endosymbiont relationship.

## P-001: GENOME-WIDE FUNCTIONAL CHARACTERISATION OF PROTEIN ISOFORMS

Savvas Kourtis[1], Sara Sdelci[1], Georg Kustatcher[2]

1. Center for Genomic Research (CRG), Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain
2. University of Edinburgh, King's Buildings, EH9 3DL, Edinburgh United Kingdom
Presenter e-mail: savvas.kourtis@crg.eu

Although virtually all genes produce multiple protein isoforms (due to alternative splicing, alternative translation start sites, post-translation cleavage etc.), a proteome-wide functional characterisation of these isoforms is missing. Protein isoforms that only differ at their n-terminal sequences, can hugely vary in their biological function, such as the two protein products of Deoxyuridine 5'-triphosphate nucleotidohydrolase (DUT), one of which localises in the nucleus, where the other localises in the mitochondria. By applying our novel statistical model, we can de-convolute peptides shared by the two protein isoforms, and thus distinguish and characterise them. Using protein co-regulation we can infer the function of each protein isoform, proteome-wide. Currently, there is ambiguity in the field of subcellular localisation of proteins, as multiple groups demonstrate that approximately 50% of the proteins exist in multiple compartments. However, the protein sequence coverage of these studies is not deep enough to capture differences in isoforms. Using our method, we can show that while a gene can make proteins that are in multiple compartments, in fact, each subcellular compartment only contains one isoform, thus changing the dogma for protein subcellular localisation.

## P-002: CHANGING TERPENOID BIOSYNTHESIS IN RICE THROUGH SYNTHETIC BIOLOGY

Orio Basallo[1], Lucia Perez[1], Abel Lucido[1], Albert Sorribas[1], Alberto Marin-Saguino[1], Ester Vilaprinyo[1], Paul Christou, Teresa Capell[1], Rui Alves[1]

1. Universitat de Lleida
Presenter e-mail: rui.alves@udl.cat

Squalene and many high valued chemicals in the pharmaceutical, biotechnological, cosmetic, and biomedical industries belong to the terpenoid family. Biosynthesis of these chemicals relies on polymerization of IPP and/or DMAPP monomers. The monomers are synthesized by plants using two alternative pathways: a cytosolic MVA pathway and a plastidic MEP pathway.

Plants need IPP/DMAPP to produce many developmental hormones, and this creates a flux bottleneck that makes it unviable to redirect IPP/DMAPP towards production of non-cognate plant chemicals. Developing plants for use as a platform to produce high value terpenoids is an important biotechnological goal that requires increasing their IPP/DMAPP production flux. Rice is a potential platform for achieving this.

With this in mind, we created mutant rice lines that implement additional alternative biosynthetic pathways for IPP/DMAPP production through synthetic biology. These lines express three different versions of an additional MVA pathway in the plastid, in addition to the normal endogenous pathways. We collected data for changes inmacroscopic and molecular phenotypes, gene expression, isoprenoids and hormone abundance in those lines.

Integrating the molecular and macroscopic data to understand how rice accommodates the new pathway is complex. To achieve that integration and generate understanding about the effects of engineering the exogenous pathways in the mutant rice lines, we develop and analyze data-centric, line specific, multilevel mathematical models. These models connect the effects of variations in hormones and gene expression to changes in macroscopic plant phenotype and metabolite concentrations within the MVA and MEP pathways of WT and mutant rice lines.

Our models allow us to understand how an exogenous IPP/DMAPP biosynthesis pathway affects the flux of biosynthesis of terpenoid precursors. They also allow us to quantify the effect of hormonal regulation on the alternative IPP/DMAPP biosynthetic pathways, enabling the prediction of macroscopic plant characteristics from molecular data. In addition, the line-specific models are a tool that can prioritize the best reengineering strategy of the exogenous pathway in order to achieve specific metabolic goals.

## P-003: RTAPAS: AN R PACKAGE TO TRACK CO-EVOLUTION OF HOSTS AND PARASITES

Mar Llaberia-Robledillo[1], José Ignasi Lucas-Lledó[1], Boris R. Krasnov[2], Oscar Pérez-Escobar[3], Juan Antonio Balbuena[1]

1. Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia, PO Box 22085, ES-46071 Valencia, Spain
2. Mitrani Department of Desert Ecology, Swiss Institute of Dryland Environmental and Energy Research, Jacob Blaustein Institute of Desert Research, Ben-Gurion University of Negev, Sede Boqer Campus, IL-84990 Midreshet Ben-Gurion, Israel
3. Royal Botanic Gardens, Kew, Richmond TW9 3AB, UK.
Presenter e-mail: mar.llaberia@uv.es

Host-parasite (HP) interactions involve the association of two organisms over a long evolutionary time. Thus, the diversification of host and parasites is rarely independent. Phylogenetic congruence quantifies the extent to which each node and branch-length in a host phylogenetic tree maps on a corresponding position in the parasite phylogenetic tree. We present an R package, Rtapas, that provides a new framework to map phylogenetic congruence (or incongruence) on a tanglegram, thereby providing insight into cophylogenetic patterns across the HP evolutionary history. To test the effectiveness of the package, we applied the analysis to a community of 130 species of small mammals and 202 species of flea parasites.

A tanglegram reflecting the congruence of the interactions between mammals and fleas its produced. The degree of congruence in hosts is concentrated in clades, whereas in fleas it is distributed across the entire tree. This suggest that within a flea family some clades exhibit some phylogenetic conservatism in relation to their hosts, whereas other clades do not, probably due to host switching.

Rtapas provides an efficient tool to gauge large and complex HP evolutionary associations. This R package facilitates and speeds up cophylogenetic analysis between two evolutionary histories, as it can handle large phylogenies reducing computational time.

## P-004: GENERATING DE NOVO GENOME ASSEMBLIES OF NON-MODEL VERTEBRATE SPECIES

Marc Palmada-Flores[1], Adrián Talavera[1], Noemi Alonso[2], Oscar Díez[2], Aryn Wilder[3], Olivier Fedrigo[4], Erich Jarvis[4,5,6], Salvador Carranza[1], Tomàs Marquès-Bonet[1,7,8,9]

1. Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain
2. Fundación para la Conservación del Quebrantahuesos, Zaragoza, Spain
3. San Diego Zoo Institute for Conservation Research, San Diego Zoo Global, Escondido, USA
4. The Vertebrate Genomes Lab, The Rockefeller University, New York, USA
5. Howard Hughes Medical Institute, Chevy Chase, USA
6. Laboratory of Neurogenetics of Language, The Rockefeller University, New York, USA
7. Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain
8. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
9. Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain

Presenter e-mail: marc.palmada@upf.edu

Multiple collaborative genomic projects are arising with the aim to sequence a wide variety of species at high quality in order to learn more about biodiversity, their biology and ultimately to help with their conservation.

The current state of sequencing technologies of long-read, 3D and genomic maps allow for the generation of high quality genomic data for non-model vertebrates which, together with improvements in the software, finally makes possible the generation of de novo high-quality genome assemblies.

Within the framework of different genomic initiatives we have sequenced and assembled at chromosome-level different vertebrate species including a mammal (red-ruffed lemur (*Varecia rubra*)), a bird (bearded vulture (*Gypaetus barbatus*)) and a reptile (snub-nosed viper (*Vipera latastei*)).

We are going to present the different approaches and workflows used to obtain the different genome assemblies and we are going to describe the resulting genome assemblies.

**P-005: THE LANDSCAPE OF EXPRESSION AND ALTERNATIVE SPLICING VARIATION ACROSS HUMAN TRAITS**

Raquel Garcia-Perez[1], Jose Miguel Ramirez[1], Aida Ripoll-Cladellas[1], Ruben Chazarra-Gil[1], Winona Oliveros[1], Oleksandra Soldatkina[1], Mattia Bosio[1], Paul Joris Rognon[2,3], Salvador Capella[1], Miquel Calvo[4], Ferran Reverter[4], Roderic Guigó[5], François Aguet[6,10], Pedro G. Ferreira[7,8,9], Kristin G. Ardlie[6], Marta Melé[1,6]

1. Department of Life Sciences, Barcelona Supercomputing Center (BCN-CNS), Barcelona, Catalonia, 08034, Spain
2. Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Catalonia, 08005, Spain
3. Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Catalonia, 08034, Spain
4. Statistics Section, Faculty of Biology, Universitat de Barcelona (UB), Barcelona, Catalonia, 08028, Spain
5. Bioinformatics and Genomics, Center for Genomic Regulation, Barcelona, Catalonia, 08003, Spain
6. Broad Institute of Harvard and MIT, Boston, MA, USA
7. Department of Computer Science, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007, Porto, Portugal
8. Laboratory of Artificial Intelligence and Decision Support, INESC TEC, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal
9. Institute of Molecular Pathology and Immunology of the University of Porto, Institute for Research and Innovation in Health (i3s), R. Alfredo Allen 208, 4200-135, Porto, Portugal
10. Illumina Artificial Intelligence Laboratory, Illumina, Inc., San Diego, CA 92121, USA
Presenter e-mail: jose.ramirez1@bsc.es

Understanding the consequences of individual transcriptome variation is fundamental to deciphering human biology and disease. Demographic traits such as ancestry, sex, age and BMI, and clinical traits such as diabetes, simultaneously affect gene expression and alternative splicing variation. However, how these variables mechanistically interplay to ultimately define an individual's phenotype is not well understood. Here, we implement a statistical framework to quantify the joint contribution of these four demographic and 17 clinical traits as drivers of gene expression and alternative splicing variation across 46 human tissues and 781 individuals from the Genotype-Tissue Expression project. We demonstrate that demographic traits have additive and tissue-specific contributions to expression variability, but trait interactions are rare. Variation in splicing is dominated by ancestry and is under genetic control in most tissues, with ribosomal proteins showing many tissue-shared splicing events, raising the possibility that specific aspects of the ribosome machinery might differ between human populations across most tissues. Our analyses reveal important contributions of clinical traits to tissue transcriptome variation. Type 1 and 2 diabetes affect multiple tissues, particularly the tibial nerve, where we further show that the nerve damage induced by diabetes shares similarities with that of biological ageing. Furthermore, we use machine learning techniques trained on histological images to classify diabetic tibial nerve samples, and, importantly, identify new genes related to diabetic neuropathy. Overall, our multi-tissue and multi-trait approach provides an extensive characterization of the main drivers of human transcriptome variation in health and disease.

**P-006: IN-SILICO ANALYSIS OF PH-DEPENDENT LIQUID-LIQUID PHASE SEPARATION IN INTRINSICALLY DISORDERED PROTEINS**

Carlos Pintado-Grima[1], Oriol Bárcenas[1], Salvador Ventura[1]

1. Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Barcelona, Spain
Presenter e-mail: Carlos.Pintado@uab.cat

Intrinsically disordered proteins (IDPs) are essential players in the assembly of biomolecular condensates during liquid–liquid phase separation (LLPS). Disordered regions (IDRs) are significantly exposed to the solvent and, therefore, highly influenced by fluctuations in the microenvironment. Extrinsic factors, such as pH, modify the solubility and disorder state of IDPs, which in turn may impact the formation of liquid condensates. However, little attention has been paid to how the solution pH influences LLPS, despite knowing that this process is context-dependent. In this work, we have conducted a large-scale in-silico analysis of pH-dependent solubility and disorder in IDRs known to be involved in LLPS (LLPS-DRs). We found that LLPS-DRs present maximum solubility around physiological pH, where LLPS often occurs, and identified significant differences in solubility and disorder between proteins that can phase-separate by themselves or those that require a partner. We also analyzed the effect of mutations in the resulting solubility profiles of LLPS-DRs and discussed how, as a general trend, LLPS-DRs display physicochemical properties that permit their LLPS at physiologically relevant pHs.

## P-007: BIOINFORMATIC ANALYSIS THROUGH QIIME2 FOR THE DETECTION OF POTENTIALLY PATHOGENIC BACTERIAL SPECIES IN ORGANIC VEGETABLES

Lara Soler[1], Yolanda Moreno[1], Laura Moreno-Mesonero[1,2], Inmaculada Amorós[1], José Luís Alonso[1], María Antonia Ferrús[2]

1. Research Institute of Water and Environmental Engineering (IIAMA), Universitat Politècnica de València. Camino de Vera s/n, 46022, Valencia, Spain.
2. Biotechnology Department, Universitat Politècnica de València, 46022, Valencia, Spain.
Presenter e-mail: ymoren@upv.es

Vegetables that are consumed raw are potential vehicles for the transmission of any type of microorganisms capable of causing human disease. Free-living amoebae (FLA) are ubiquitous protozoa found in many ecosystems and can serve as hosts to pathogenic bacteria. So far, data of the FLA microbiome in vegetables remain scarce. Thus, in this work a preliminary characterization of the microbiome of FLA isolated from organic vegetables (lettuce, spinach and cabbage) and strawberries were carried out to know their possible implications for consumers. The bacterial microbiome of FLA was obtained after amplicon sequencing using Illumina MiSeq platform and pair-end protocol. Raw data was analyzed using QIIME2. Sequences were first imported. Then, forward and reverse sequences were merged, quality and chimera-checked, and clustered into Amplicon Sequence Variants (ASVs) using the DADA2 algorithm. ASVs underwent a taxonomic assignment using "classify-consensus-blast" plugin and SILVA SSU v138.1 database. Alpha-diversity metrics, beta diversity metrics unweighted UniFrac and Principal Coordinate Analysis (PCoA) were estimated using q2-diversity after samples were rarefied. Beta diversity analysis was computed to establish the differences of microbial communities between the different types of organic fresh produce. Finally, the core microbiome was obtained to explore the bacterial taxa shared by all the samples. Moreover, the identification of the protozoa *Acanthamoeba* spp. and *Vermamoeba vermiformis* was carried out by qPCR. Both *Acanthamoeba spp*. and *V. vermiformis* were identified in 65% and 25% of the samples, respectively. The most abundant bacteria of the FLA microbiome were *Pseudomonas*, *Flavobacterium* and *Prosthecobacter* in all samples. Bacteria not previously related to FLA such as *Bacillus* or *Brevundimonas* have been detected in this work. Moreover, some identified genera as part of the FLA microbiome are of public health interest because they contain species that are potentially pathogenic to humans such as pseudomonas or flavobacterium. Although these are found in low abundance values.

**P-008: A NEW PIPELINE FOR DETECTING TRANSPOSON DOMESTICATION EVENTS IN TAXONOMICALLY RESTRICTED GENES**

Carlos Herrera-Ubeda[1,2], Jordi Garcia-Fernandez[1], Peter W H Holland[2]

1. Department of Genetics, Microbiology & Statistics, and Institute of Biomedicine (IBUB), University of Barcelona, 08028, Barcelona, Spain.
2. Department of Biology, University of Oxford, Oxford, OX1 3SZ, UK.
Presenter e-mail: carlos.herrera@ub.edu

Newly evolved genes in a given lineage showing no homologs in other taxa are known as “orphan” or “taxonomically restricted” genes. One of the most striking sources for the birth of lineage-restricted genes is the molecular domestication of transposable element (TE) proteins into novel coding genes, which are sometimes involved in the appearance of clade-specific traits and even true evolutionary novelties. Thanks to the increasingly growing genomic data publicly available from projects like the Darwin tree of life, we developed a pipeline capable of identifying putative TE domestication events within taxonomically restricted genes across hundreds of species. Here, we focused on the available genomes of the Lepidoptera order to better understand the mechanics behind the appearance of taxonomically restricted genes by TE domestication.

## P-009: CHARACTERIZATION OF *ARABIDOPSIS* PEPTIDOME AND ITS ROLE IN FLOWER DEVELOPMENT

Raquel Álvarez-Urdiola[1], José Tomás Matus[1,2], José Luis Riechmann[1,3]

1. Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Edifici CRAG, Campus UAB, 08193 Cerdanyola del Vallès, Barcelona, Spain.
2. Institute for Integrative Systems Biology (I2SysBio), Universitat de València-CSIC, Paterna, 46908, Valencia, Spain.
3. Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.
Presenter e-mail: raquel.alvarez@cragenomica.es

A substantial but uncharted fraction of plant proteomes is composed of small proteins (peptidome), of roles and functions -for the most part- yet to be discovered. Several putative peptides encoded in short Open Reading Frames (sORFs) have been found in intergenic regions and transcripts previously identified as non-coding by integrative analysis using omics-based approaches. In order to characterize the *Arabidopsis* peptidome and, in particular, its possible involvement in flower development, we performed tandem mass spectrometry of inflorescences of the floral homeotic mutants *apetala1*, *apetala2*, *apetala3*, *pistillata*, and *agamous*, and wild-type plants of *Arabidopsis thaliana*. For peptide identification we created an extensive database that includes putative sORF-encoded peptides (SEPs) from intergenic regions, UTRs, 'non-coding' RNAs and other transcripts. By combining the datasets from individual / wild type comparisons we were able to identify 132 novel peptide candidates expressed in flowers, from which 60 peptides are predicted to be specifically expressed, or at least enriched, in one type of floral organ. Around 25% of the novel SEPs belong to putative gene families in *A. thaliana* and almost 80% have putative homologs in other plant species such as *A. lyrata*, *Brassica oleracea* and *Camelina sativa*.

## P-010: NONUNIQUE UPGMA CLUSTERINGS OF MICROSATELLITE MARKERS

Natàlia Segura-Alabart[1], Francesc Serratosa[1], Sergio Gómez[1], Alberto Fernández[2]

1. Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Spain
2. Departament d'Enginyeria Química, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Spain.
Presenter e-mail: natalia.segura@gmail.com

Agglomerative hierarchical clustering has become a common tool for the analysis and visualization of data, thus being present in a large amount of scientific research and predating all areas of bioinformatics and computational biology. In this work, we focus on a critical problem, the nonuniqueness of the clustering when there are tied distances, for which several solutions exist but are not implemented in most hierarchical clustering packages. We analyze the magnitude of this problem in one particular setting: the clustering of microsatellite markers using the Unweighted Pair-Group Method with Arithmetic Mean. To do so, we have calculated the fraction of publications at the Scopus database in which more than one hierarchical clustering is possible, showing that about 46% of the articles are affected. Additionally, to show the problem from a practical point of view, we selected two opposite examples of articles that have multiple solutions: one with two possible dendrograms, and the other with more than 2.5 million different possible hierarchical clusterings.

## P-011: FixNCut: DIGESTING FIXED TISSUES FOR SINGLE-CELL GENOMICS

Laura Jiménez-Gracia[1,*], Domenica Marchese[1,*], Juan C. Nieto[1], Ginevra Caratú[1], Xavier Hernando-Momblona[2], Elisa Melón-Ardanaz[3], Victoria Gudiño[3], Joana P. Bernades[4], Florian Tran[4,5], Laura Katharina Sievers[4,5], Stefan Schreiber[4,5], Maarten van den Berge[6,7], Tessa Kole[6,7], Petra L. van der Velde[7,8], Martijn C. Nawijn[7,8], Philip Rosenstiel[4], Eduard Batlle[2,9,10], Azucena Salas[3], Holger Heyn[1,11,12], Luciano G. Martelotto[12,13,14,†]

1. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.
2. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri i Reixac 10, 08028 Barcelona, Spain.
3. Department of Gastroenterology, IDIBAPS, Hospital Clínic, Barcelona, Spain.
4. Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany.
5. University Hospital Schleswig-Holstein (UKSH), Campus Kiel, Kiel, Germany.
6. Department of Pulmonary Diseases, University of Groningen, University Medical Center Groningen, Groningen, Netherlands.
7. Groningen Research Institute for Asthma and COPD, University of Groningen, Netherlands.
8. Department of Pathology & Medical Biology, University of Groningen, Netherlands
9. Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Barcelona, Spain.
10. ICREA, Barcelona, Spain.
11. Universitat Pompeu Fabra (UPF), Barcelona, Spain.
12. Omniscope, Parc Científic de Barcelona, Edifici Torres R+D+I, C/ Baldiri Reixac 4-8, 08028 Barcelona.
13. Adelaide Centre for Epigenetics (ACE), University of Adelaide, South Australia, Australia.
14. South Australian immunoGENomics Cancer Institute (SAiGENCI),
University of Adelaide, South Australia, Australia.
* These authors contributed equally to this work.
† Joint senior authors.
Presenter e-mail: laura.jimenez@cnag.crg.eu

The expansion of single-cell applications in the era of personalized medicine and the increasing complexity and decentralization of current studies require disconnecting the time and site of sampling from downstream processing steps. In addition, cellular transcriptomic profiles are dynamic and can change under external stressors, altering the natural state of the cells. In light of this, early sample preservation improves robustness and reproducibility by avoiding artifacts introduced during sample handling. Here, we present FixNCut, an approach for reversibly fixing tissue followed by dissociation, that overcomes limitations in the generation of single-cell data. As a proof of concept, we have showcased our sample processing approach prior to single-cell RNA sequencing for mouse lung and colon tissue, sample types with inherent sampling challenges. Further, we processed human lung and colon from different pathological contexts to provide further evidence of the clinical value of this methodology. We showed that this reversible fixation followed by dissociation strategy enables the removal of time and location constraints while preserving the RNA integrity, library complexity and cellular composition. Moreover, fixed cells can be stained with antibodies, as proved by cytofluorimetric analysis, thus allowing FACS sorting or hashing prior to single-cell analysis. In principle, FixNCut is

potentially compatible with multiple standard single-cell technologies, making it a versatile protocol to enable robust and flexible study designs.

## P-012: PREDICTION OF BACTERIAL INTERACTOMES BASED ON GENOME-WIDE COEVOLUTIONARY NETWORKS: AN UPDATED IMPLEMENTATION OF THE CONTEXTMIRROR APPROACH

Miguel Fernández[1], Camila Pontes[1], Victoria Ruiz-Serra[1], Alfonso Valencia[1]

1. Barcelona Supercomputing Center (BSC), Life Science, 08034, Barcelona, Spain.
Presenter e-mail: miguel.fernandez@bsc.es

Interacting proteins tend to experience reciprocal evolutionary change to retain biological functions. This coevolution, quantified as the degree of similarity between the phylogenetic trees of protein families, has been demonstrated to be an indicator of protein-protein interaction. The observed degree of resemblance between phylogenetic trees is the result of the concerted action of many factors, not necessarily related with the interaction, such as the effect of background evolution on protein families because of underlying speciation events.

Previously proposed methods treat every protein pair as an independent interaction partner; however, proteins interact with many others and its adaptation is influenced by them. The ContextMirror approach uses the coevolutionary network of an organism (network of similarities between all pairs of proteins in a genome) to evaluate coadaptation by integrating the influence of every interactor on a given protein pair. This has been proven to be a significant enhancement in the accuracy of protein-protein interaction inference since its publication in 2008. In this review we present an updated version of the ContextMirror approach, better fitted to the vast amount of available data and compliant with the latest findings on the field. The testing of this implementation with a reduced set of proteins yielded a 90.4% precision and 38.7% recall when compared against experimentally validated interactions.

## P-013: IMPLEMENTATION OF CELLULAR TRANSPORT MECHANISMS WITHIN A MULTISCALE SIMULATION FRAMEWORK

Othmane Hayoun-Mya[1], Arnau Montagud[1], Miguel Ponce-de-León[1], Alfonso Valencia[1,2]

1. Barcelona Supercomputing Center (BSC), Plaça d'Eusebi Güell, 1-3, 08034 Barcelona, Spain
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain
Presenter e-mail: ohayoun@bsc.es

Agent-based models (ABMs) constitute an advantageous approach for multicellular systems modeling. Nonetheless, cell and tissue ABMs usually do not include detailed transport mechanisms which interface between the agent or cell and its chemical microenvironment. Given that these mechanisms are a key biological process, we set to develop and implement a compendium of general transport mechanisms into a cellular ABM-based multiscale software: PhysiCell. We consider six different mechanisms which encompass both passive (simple and facilitated diffusion through carriers and channels) and active transport (primary and secondary). Transport mechanisms were described as systems of Ordinary Differential Equations (ODEs), conceptually grounded on Fick's Diffusion Laws and Michaelis-Menten kinetics. We assessed density equilibrium dynamics in each mechanism through simple experiments, confirming saturable, nonlinear dynamics. We also performed unit testing on simple models to check for unexpected dynamics, and a PhysiCell-based template of drug resistance acquisition was then built by combination of three models. We also present ongoing work that combines molecular transport and boolean modeling within PhysiCell in order to recapitulate different experimental drug synergies in a gastric cancer cell line. We argue that our models improve PhysiCell base code, which accounts only for linear transport dynamics. Furthermore, we made an effort to assemble modular models with interoperable units, which will facilitate connection with specific agent rules and will make them more user-friendly by simplifying model fitting with experimental data.

**P-014: EpiGe-App: A WEB SERVER FOR THE GENERATION OF AUTOMATED MEDULLOBLASTOMA SUBGROUP PREDICTIONS DIRECTLY FROM QUANTITATIVE PCR EXPERIMENTS**

Joshua Llano-Viles[1,2,3,4], Soledad Gómez-González[3,4], Cinzia Lavarino[3,4], Alexandre Perera-Lluna[1,2,3]

1. B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain
2. Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain
3. Institut de Recerca Pediàtrica Hospital Sant Joan de Dèu, Esplugues de Llobregat, Barcelona, Spain
4. Pediatric Cancer Center Barcelona (PCCB), Hospital Sant Joan de Déu, Barcelona, Spain.
Presenter e-mail: joshua.llano@upc.edu

Classification of medulloblastoma (MB), the most common malignant pediatric brain tumor, into clinically relevant molecular subgroups (WNT, SHH and non-WNT/non-SHH) is crucial to personalize tumor treatment and optimize patient care. Genome-wide DNA methylation profiling enables precise classification of this tumor. However, the implementation of this classification approach in daily clinical practice is challenging for many centers worldwide. Here, we present EpiGe-App (https://www.epige.irsjd.org), a web server dedicated to classify MB tumors using a six-cytosine methylation signature from methyl-genotyping PCR data. EpiGe-App contains an integrated computational pipeline which performs a pre-processing analysis and applies machine learning to classify a sample into a MB subgroup. At the beginning of the pipeline, quality control and cytosine amplification data are extracted from the text file obtained by qPCR analysis. Next, logistic regression is applied on the amplification data to predict the methylation status and a Hamming distance-based score is used to assign the molecular subgroup. Finally, EpiGe-App generates a downloadable PDF file with high-resolution figures that summarize the results of the analysis. Overall, EpiGe-App is a unique, simple, accurate and cost-effective solution for the classification of MB tumors into clinically relevant subgroups using qPCR technology, which is a widespread molecular technique.

## P-015: CHARACTERIZATION OF SEX-SPECIFIC DIFFERENCES AND LIPIDOME VARIANCE IN DIABETES MELLITUS THROUGH MULTIVARIATE ANALYSIS

Maria Barranco-Altirriba[1,2,3,4], Esmeralda Castelblanco[5], Paola Quifer[1,6], Oscar Yanes[6,7], Joana Rossell[1,6], Minerva Granado-Casas[1,6], Ralf Weber[8], Emilio Ortega[9,10], Núria Alonso[6,11], Alexandre Perera-Lluna[2,3,4], Didac Mauricio[1,6,12]

[1] Sant Pau Biomedical Research Institute (IIB Sant Pau), Hospital de la Santa Creu i Sant Pau, Endocrinology and Nutrition, Barcelona, Spain
[2] B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain
[3] Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain
[4] Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain
[5] Department of Internal Medicine, Endocrinology, Metabolism and Lipid Research Division, Washington University School of Medicine, St Louis, MO, USA·
[6] Center for Biomedical Research on Diabetes and Associated Metabolic Diseases, CIBERDEM, Barcelona, Spain
[7] Institute of Health Research Pere Virgili (IISPV) & University Pere i Virgili, Metabolomics Platform, Reus, Spain
[8] Phenome Centre Birmingham and School of Biosciences, University of Birmingham, Bioinformatics, Birmingham, United Kingdom
[9] Institut d'Investigacions Biomèdiques August Pi i Suñer, Hospital Clinic, Endocrinology and Nutrition, Barcelona, Spain
[10] Center for Biomedical Research on Physiopathology of Obesity and Nutrition, CIBEROBN, Madrid, Spain
[11] Institute for Health Science Research Germans Trias i Pujol (IGTP), University Hospital and Health Research Institute Germans Trias i Pujol, Endocrinology and Nutrition, Badalona, Spain
[12] Faculty of Medicine, University of Vic (UVIC/UCC), 08500 Vic, Spain.
Presenter e-mail: maria.barranco@upc.edu

Lipid metabolism disruption in diabetes has been widely recognized, but the lipidome changes between Type 1 diabetes (T1D) and Type 2 diabetes (T2D) have not been fully characterized, nor have been studied the sex-specific differences in the lipidome due to diabetes.

Here, we applied multivariate linear regression analysis to untargeted lipidomics data of 360 subjects (91 T1D, 91 T2D and 178 controls) without cardiovascular and/or chronic kidney disease. The data were acquired using ultra-high performance liquid chromatography-electrospray ionization tandem mass spectrometry (UHPLC-ESI-MS/MS) and LipidSearch was used to identify MS/MS spectra. Multiple linear regression models were used to assess the association between each lipidic feature and each condition. In order to determine sex-specific differences related to diabetes, the interaction between diabetes and sex was considered. The models were adjusted by 15 different variables and diabetes duration was also included in the comparison between T1D and T2D.

A total of 1411 and 831 unique significant features were found in all 9 analyses in positive and negative ionization modes, respectively. The significant features that had been previously

identified using LipidSearch represented 70 unique lipid subspecies from 15 unique lipidic classes. Lysophosphatidylcholines (LPC) and Ceramides (Cer) showed opposite effects in T1D and T2D and several ether glycerophosphoplipids were significantly reduced in T2D. Regarding sex-specific differences, men presented a higher metabolism alteration in T1D than women. Otherwise, women showed larger disruption in lipid metabolism in T2D than men and a greater difference between T1D and T2D metabolism than their male analogues. Specifically, Ceramides and phosphatidylcholines (PC) showed more important diabetes-associated differences due to sex.

Our results show an extensive characterization of lipid metabolism in subjects with T1D and T2D, as well as sex-specific lipidome changes associated to diabetes.

## P-016: STUDYING TYPE 2 DIABETES MELLITUS SUB-TYPES FROM LONGITUDINAL HEALTH RECORDS USING UNSUPERVISED DEEP LEARNING TECHNIQUES

Enrico Manzini[1,2,3,a], Bogdan Vlacho[4,a], Josep Franch-Nadal[4,5,6], Joan Escudero[7], Ana Génova[7], Elisenda Reixach[8], Erik Andrés[8], Israel Pizarro[9], José-Luis Portero[9], Dídac Mauricio[4,5,10,b], Alexandre Perera-Lluna[1,2,3,b]

1. B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain
2. Networking Biomedical Research Centre in the subject area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain
3. Institut de Recerca Sant Joan de Deu, Esplugues de Llobregat, Barcelona, Spain
4. DAP-Cat Group, Unitat de Suport a la Recerca Barcelona, Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain
5. CIBER of Diabetes and Associated Metabolic Diseases (CIBERDEM), Institutode Salud Carlos III (ISCIII), Spain
6. Primary Health Care Center Raval Sud, Gardenia d'Atenció Primària, Institut Català de la Salut, Barcelona, Spain
7. Grupo Pulso, Spain
8. Fundació TIC Salut Social, Departament de Salut, Generalitat de Catalunya, Barcelona, Spain
9. Novo Nordisk, Spain
10. Department of Medicine, University of Vic - Central University of Catalonia, Vic, Spain
a. These authors have contributed equally to this work and share the first authorship
b. These authors have contributed equally to this work and share senior authorship
Presenter e-mail: enrico.manzini@upc.edu

Type2 Diabetes Mellitus (T2DM) is a chronic disease that shows highly heterogeneous progressions between patients, with different associated complications and response to therapies. Since the recent publication of Ahlqvist et al. (Ahlqvist al., 2018) of a first attempt to clusterize diabetic subjects using six clinical parameters, several efforts have been made to investigate the heterogeneity of individuals with T2DM.

Using a Kernelized-Autoencoder, we mapped 5 years of data in the format of Electronic Health Records (EHRs) of 11.028 patients in a latent space. The model was trained assuring that the vectors in the latent space do embed similarities and differences in the input data. Once we obtained the vectors from the latent space, we used classical algorithms (k-mean, CLARA and hierarchical clustering) to identify different T2DM sub-types.

This unsupervised deep learning clustering algorithm suggested 7 different clusters of T2DM subjects, representing different evolutions of the disease.

We observed different mean ages among the clusters (ranging from 65.3±11.6 to 72.8±9.4) and different diabetes duration, with subjects in clusters B (Hypercholesteraemic) and E (Hypertensive) having shorter diabetes duration (9.2±3.9 and 9.5±3.9 years respectively). Differences were also observed in terms of glycaemic control, ranging from a mean glycated hemoglobin of 7.99±1.42% in Cluster G (Metabolic) to 7.04±1.11% in cluster E (Hypertensive)

The deep learning algorithm used in this study confirms the heterogeneity of T2DM and offers the possibility of investigating different trajectories of T2DM disease from routinely collected EHRs.

**P-017: INFERRING EVOLUTIONARY ADAPTATION ACROSS THE HUMAN DEVELOPMENTAL IMMUNE SYSTEM.**

Irepan Salvador-Martínez[1,2], Jesus Morga-Moreno[3], Juan Nieto[1], Aina Vaquer[2], Marta Coronado-Zamora[2,4], Antonio Barbadilla[2], David Enard[3], Holger Heyn[1]

1. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain.
2. Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Barcelona, Spain
3. University of Arizona Department of Ecology and Evolutionary Biology, Tucson, United States.
4. Institute of Evolutionary Bioloy, CSIC, Universitat Pompeu Fabra. Barcelona, Spain
Presenter e-mail: irepan.salvador@cnag.crg.eu

The human immune system (HIS) is continuously evolving due to natural selection (NS). In a population, NS increases the frequency of genetic variants that improve survival, leaving adaptive signatures in its genome.

The recent publication of the developmental HIS atlas provided the most comprehensive transcriptional characterisation of immune cell types in different organs and ontogenic stages.

Using this cell atlas and state-of-the-art population genomic methods, we inferred adaptation in the HIS cell types. Our results show differential adaptation patterns in human immune cell types at an unprecedented resolution, identiying cell types in both myeloid and lymphoid compartments with high rates of adaptive substitution.

## P-018: SINGLE-CELL RNA AND T CELL RECEPTOR SEQUENCING ANALYSIS IN PLEURAL EFFUSION TO MONITOR THE EVOLUTION OF THE IMMUNE SYSTEM DURING IMMUNOTHERAPY IN LUNG CANCER

Paula Nieto[1], Juan C. Nieto[1], M. Alba Sorolla[2], Anabel Sorolla[2], Mara Mulet[3], J. Manuel Porcel[2,3], Silvia Vidal[4], Holger Heyn[1]

1. Centro Nacional de Análisis Genómico CNAG-CRG, Single-Cell Genomics, Barcelona, Spain.
2. Research Group of Cancer Biomarkers, Institut de Recerca Biomèdica de Lleida, Lleida, Spain.
3. Pleural Medicine Unit, Internal Medicine Service, Hospital Universitario Arnau de Vilanova de Lleida, Lleida, Spain.
4. Institut de Recerca de l'Hospital de la Santa Creu i Sant Pau, Grupo de Investigación en Enfermedades Inflamatorias, Barcelona, Spain.
Presenter e-mail: paula.nieto@cnag.crg.eu

Patients with advanced lung cancer stages can suffer malignant pleural effusions (PE). This pleural effusion is used as a diagnostic tool and can be considered malignant PE when a cytology or pleural biopsy culture shows malignant tumoral cells. In these patients, immune checkpoint inhibitors have become the main therapeutic strategy. In this study, we aim to analyze the TCR repertoires from paired PE, malignant pleural biopsies, and peripheral blood samples in a longitudinal study of probands following immune checkpoint therapy with anti-PD-L1 (ongoing). We have analyzed a pilot cohort of four malignant PE patients (induced by lung cancer) using single-cell RNA sequencing techniques as well as single-cell TCR sequencing (over 18000 total cells) and we characterized the populations in their TIMEs (both immune and non-immune compartments). Furthermore, we applied computational methods for the joint analysis of TCR sequences and gene expression, which make possible to understand not only clonal expansion, but also how the expansion profiles differ between patients and how they might relate to prognosis. We have identified seven major populations present, in different proportions, in all patients: monocytes, macrophages, dendritic cells, T cells, B cells, fibroblasts and tumor cells; as well as 46 total subpopulations, not all of them shared between all patients; and all of this annotated by their top marker genes. Furthermore, we identified five clusters of patient-specific expanded TCRs, which greatly differ in their phenotypes: exhausted, effector, cytotoxic, naive and intermediate phenotypes, and could clearly be linked to treatment response and prognosis. With this pilot cohort, and thanks to single-cell RNA sequencing techniques, we showed that we could use pleural effusions to recapitulate the composition of the TIME. We can identify both differences and commonalities across patients. Furthermore, we were able to find striking differences in T cell expanding populations, which can be a promising target and predictor of treatment outcome and patient response. The goal is to apply this knowledge, together with the results of the ongoing clinical study, to train a predictive model of tumor-specific TCR sequences that we can use for precision oncology.

## P-019: DYNORPHINS AS CELL PENETRATING PEPTIDES

Èric Catalina-Hernández[1], Mario Lopez-Martin[1], Alex Peralvarez-Marin[1]

1. Universitat Autònoma de Barcelona, Medicine Faculty, Biophysics Unit, Building M, Av. Can Domènech, 08193 Bellaterra, Barcelona, Catalonia, Spain
Presenter e-mail: eric.catalina@autonoma.cat

Membrane proteins are crucial in connecting the cell with its environment. Most current drug therapeutic approaches are directed to membrane protein targets, mainly because of the involvement of these proteins in signalling cascade events, as well as in the transport between the exterior and the interior of the cell and among cellular compartments. Understanding how membrane proteins regulate all these processes will lead a to better knowledge on how to tackle diseases at the molecular level. Dynorphins are endogenous peptides and the canonical substrate for kappa-opioid receptor (KOR). Dynorphins are prohormone derived from prodynorphin (PDYN), which is cleaved into big dynorphin (BigDyn, 32 residues), and further processed into dynorphin A (DynA, 17 residues) and B (DynB, 13 residues). Pathophysiological implications for dynorphins have been described owing to their cell penetrating peptide (CPP)-like behaviour, as molecules capable of inducing membrane translocation via membrane destabilization and/or pore formation. Besides, they are related to Alzheimer's and Parkinson's diseases (AD, PD). Hence, we have studied DynA CPP activity. Thus, we have simulated DynA in 3 types of membrane (DPPC, DPPC:DOPC:CHOL, and DPPC:DOPC:DPPS:DOPS:CHOL) and have induced membrane translocation through Steered Molecular Dynamics (SMD) simulations. We observed a proteolipidic pore after the simulations, since dynorphins, as positively charged peptides, interact with the negatively charged polar head groups of phospholipids. We have computed the potential of mean force (PMF) as a way of comparing the CPP activity to peptides with similar characteristics. We are currently analysing and undergoing more simulations to extract concluding information.

## P-020: EXPLORING THE DEMOGRAPHIC HISTORY OF AMAZIGH GROUPS FROM ALGERIA

Laura Vilà-Valls[1], Amine Abdeli[2], Marcel Lucas-Sánchez[1], Asmahan Bekada[3], Traki Benhassine[2], David Comas[1]

1. Institute of Evolutionary Biology (UPF-CSIC), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain
2. Laboratoire de Biologie Cellulaire et Moléculaire, Faculté des Sciences Biologiques, Université des Sciences et de la Technologie Houari Boumediene, Alger, Algeria
3. Département de Biotechnologie, Faculté des Sciences de la Nature et de la Vie, Université Oran
1 (Ahmad Ben Bella), 1524 El M'Naouer, Oran 31000, Algeria
Presenter e-mail: laura.vila@upf.edu

Present-day North Africa is complex and rich in diversity of languages, religions and cultures, since it has been geographically strategic for many demographic movements. Traditionally, inhabitants of North Africa have been divided in two main groups, Arabs and Imazighen (sg. Amazigh). The presence of North Africa in genetic studies is scant compared to the rest of the African continent, which has received much more attention as the geographic origin of the human species. In the present project we analyzed new genome-wide array data from selected North African populations, in order to infer their population history and genetic diversity as a consequence of their demographic history. The samples under study belong to different groups of Imazighen from Algeria, including a group of Mozabites, and three Chaoui groups from different towns in the Aurès region: Batna, Khenchela and Oum El Bouaghi. Our analyses seem to indicate heterogeneity among the Amazigh populations, although little differences are found between North African groups, whether Arab or Amazigh. Mozabite and Chaoui from Khenchela show a recent admixture with a sub-Saharan source population, although this admixture event is not detected in non-Amazigh North African groups and in the groups of Batna and Oum El Bouaghi, suggesting a different demographic history between the Amazigh populations of Algeria. Moreover, the population of Khenchela exhibits on average higher levels of inbreeding, suggesting an isolation that could be the result of a bottleneck or a founder effect.

**P-021: BACTERIA FROM SALINE ENVIRONMENTS: UNDERSTANDING THEIR LIFESTYLE AND REWIRING THEIR METABOLISM**

Karina Hobmeier[1], Alberto Marin-Sanguino[2]

1. Technical University of Munich
2. Universidad de Lleida
Presenter e-mail: alberto.marin@udl.cat

Halophilic bacteria are promising candidates to expand the ever growing set of biotechnological workhorses due to their ability to grow in non-edible waters and produce compatible solutes. These are low molecular weight substances able to accumulate in high concentrations in the cytoplasm without compromising any physiological processes. Their protectant nature makes them very valuable to the pharmaceutical industry.

In this group we will summarize our experience combining mathematical modeling, bioinformatics and molecular genetics to improve the production of the compatible solute ectoine by the halophilic bacterium *Halomonas elongata*. We will emphasize how the role of mathematical modeling goes beyond the algorithmic formulation of recipes for genetic manipulation and how the dialogue between modeling and experiments can lead to a deeper understanding of metabolic organization.

**P-022: ModCRE: A STRUCTURE HOMOLOGY-MODELING APPROACH TO PREDICT TF BINDING IN CIS-REGULATORY ELEMENTS**

Fornes O[1], Meseguer A[2], Aguirre-Plans J[2], Patrick Gohl[2], Mireia-Bota, P[2], MolinaFernández R[2], Bonet J[2,3], Chinchilla, A[4], Pegenaute, F[4], Gallego, O[4], FernandezFuentes, N[5], Oliva B[2*]

1. Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, BC V5Z 4H4, Canada
2. Structural Bioinformatics Lab (GRIB-IMIM), Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona 08005, Catalonia, Spain
3. Laboratory of Protein Design & Immunoengineering, School of Engineering, Ecole Polytechnique Federale de Lausanne, Lausanne 1015, Vaud, Switzerland
4. Live-Cell Structural Biology. Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona 08005, Catalonia, Spain
5. Institute of Biological, Environmental and Rural Science, Aberystwyth University, SY23 3DA Aberystwyth, UK.
Presenter email: patrick.gohl@upf.edu

Transcription factor (TF) binding is a key component of genomic regulation. There are numerous high-throughput experimental methods to characterize TF-DNA binding specificities. Their application, however, is both laborious and expensive, which makes profiling all TFs challenging. For instance, the binding preferences of ~25% human TFs remain unknown; they neither have been determined experimentally nor inferred computationally. Here, we introduce ModCRE, a web server implementing a structure homology-modelling approach to predict TF motifs and automatically model higher-order TF regulatory complexes. Starting from a TF sequence or structure, ModCRE predicts a set of motifs for that TF. The predicted motifs are then used to scan the DNA for occurrences of each of them, and the best matches are either profiled with a binding score or collected for their subsequent modeling into a higher-order regulatory complex with DNA, as well as other TFs and co-factors. Moreover, we demonstrate that incorporating high-throughput TF binding data, such as from protein binding microarrays, addresses the protein-DNA structure scarcity problem for deriving statistical potentials. In turn, these statistical potentials are proven to be capable predictors of TF motifs. We also show the conditional advantage of using ModCRE over a nearest-neighbor approach for predicting TF binding sites as well as an improvement in prediction accuracy when using a rank-enrichment selection system. Finally, as case examples, we apply ModCRE to model the interferon beta enhanceosome and the complex of SOX2 and 11 with a nucleosome.

## P-023: MONITORING SARS-COV-2 PREVALENCE AND VARIANTS IN WASTEWATER IN VALENCIA: A WASTEWATER-BASED EPIDEMIOLOGY APPROACH

Rosa Bermejo Ramirez[1], Rubén Cañas Cañas[1], Ester Méndez Belinchón[1]

1. General de Análisis, Materiales y Servicios, SL (GAMASER – Global Omnium), Carrer del Corretger 51, 46988, Paterna, Valencia, Spain.
Presenters e-mail: rubencac@globalomnium.com, roberrami@globalomnium.com

COVID-19 pandemic has dealt an immense health and socioeconomic impact worldwide. SARS-CoV-2 surveillance programmes have allowed following the evolution of an emergent virus for the first time in history. Currently, prevalence estimation is conducted via individual diagnostic tests, which are limited and lead to other problems such as not covering the whole population and delays in the clinical information reports. Wastewater-based epidemiology (WBE) can overcome this limitations, not only by monitoring the virus circulation nearly in real-time and eliminating the sampling bias but also as a powerful epidemiologic prediction tool for the evolution of the virus and its clinical consequences. From GAMASER (Global Omnium) we have been conducting this practice, which has enabled the creation of an early warning system and the prediction of epidemiologic trends. Furthermore, we could estimate the proportion of the different SARS-CoV-2 variants using high-throughput sequencing and deconvolution algorithms based on general additive models. WBE has helped providing support to health authorities in crucial decision making moments, and the future perspectives are to monitor a great number of markers and environmental factors, such as virus and antibiotic resistance pathogens, chronic diseases biomarkers or metabolites from pharmaceutical and illicit drugs, which could contribute to evaluate the population habits and health situation from a broader point of view.

## P-024: A METAGENOMICS PIPELINE TO LINK *STREPTOCOCCUS* GENUS WITH PHAGES IN THE HUMAN MICROBIOME

Elena Cabello Yeves[1], Raúl Cabrera Rubio[2], María Carmen Collado Amores[2], Iñaki Comas[3], Alberto Marina[1]

1. Institute of Biomedicine of Valencia CSIC, Genomics and Proteomics, Valencia, Spain.
2. Institute of Agrochemistry and Food Technology, Food Biotechnology, Valencia, Spain.
3. Institute of Biomedicine of Valencia CSIC, Genomics of Tuberculosis, Valencia, Spain.
Presenter e-mail: ecabello@ibv.csic.es

The microbiota in neonatal development is a complex bacterial ecosystem influenced by various factors such as the maternal microbiota, delivery, weight, antibiotics use or diet. Because of the high probability of infection in the hospitalized neonates, antibiotics are administered for 72 hours per protocol. The efficiency of such administration is in question because of the multiresistant profile of many of these bacteria, and their unknown effects on the baby's microbiota. We usually associate infections with vertical maternal-neonatal transmission during pregnancy and childbirth, frequently caused by Streptococcus type B among others. One of the ways in which bacteria acquire these resistances to antibiotics is through horizontal gene transfer (HGT) usually from bacteriophages (phages) to bacteria. Phages encode and express auxiliary metabolic genes (AMGs) which are bacterial host genes that can impact host metabolic processes and being transferred from host-to-host through horizontal gene. Metagenomics allows the identification of AMGs from phages from a specific environment and the reconstruction of metagenome assembled genomes (MAGs) from microbes of each sample. Metagenomic data from a prospective mother-infant birth cohort in the Spanish-Mediterranean area have been used for this project to specifically look for correlations between phages and Streptococcus in the human microbiome. Here, a metagenomics pipeline has been developed to extract MAGs from Streptococcus genus and link them with phages by looking for AMGs in both bacteria and virus. To develop our understanding, results will be discussed to underpin future studies and suggest some solutions to existing challenges.

## P-025: ANALYSIS OF CANCER DRIVERS IN NON-CODING REGULATORY REGIONS

Miguel Rodriguez-Galindo[1], Donate Weghorn[1,2]

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

Tumour progression is dominated by two evolutionary forces: first mutagenesis, which provides the heritable variability where, secondly, natural selection acts. The main challenge of cancer genomics is to identify the somatic mutations that drive the tumorigenesis (drivers) from the vast majority of neutral variation (passengers).

A decade of careful surveying of tumour DNA has revealed a multitude of protein-coding drivers, several of which have been used as therapeutic targets. However, many tumors do not exhibit any of these known exonic driver events, leaving a gap in our knowledge. Recently, large efforts have been made querying the remaining non-coding part of the genome. Intriguingly, the role of non-coding somatic mutations still remains largely less well understood than its protein-coding counterpart. These regions are specially challenging: poorly annotated, dominated by abnormal mutational processes that act as confounders and with a broader mutational target that coding regions.

To address these specific challenges we developed an approach that identifies selection in regulatory regions. Regulatory units are tested as a whole to increase the statistical power and the impact of mutations is evaluated through biophysical models for transcription factor binding sites. Finally, a mutational model accounts for mutation rate variability at several scales. With this approach we find known and new putative cancer drivers on an harmonized cohort of 7498 somatic whole cancer genomes.

## P-026: IsoAnnot: A NEW PIPELINE FOR FUNCTIONAL ANNOTATION AT ISOFORM RESOLUTION

Alessandra Martínez Martín[1], Francisco J. Pardo-Palacios[1], Pedro Salguero[2], Ángeles Arzalluz-Luque[2], Lorena de la Fuente[3], Ana Conesa[1]

1 Genomics of Gene Expression Lab, Instituto de Biología Integrativa de Sistemas (CSIC-UV),Valencia, Spain
2 Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Spain
3 PerkinElmer Informatics, R&D, Tres Cantos, Spain
Presenter e-mail: alessandra.martinez.m@gmail.com

Post-transcriptional regulation (PTR) mechanisms are essential for creating transcriptome and proteome complexity in eukaryotic cells. Previous studies suggest that alternative exons modulate properties such as subcellular localization or mRNA stability. However, how these mechanisms imprint distinct functional characteristics on the resulting set of isoforms to define the observed phenotype remains poorly understood. Functional profiling is the most broadly adopted genome-wide approach for characterization of the functional relevance of gene expression regulation. However, the gene-centric nature of functional annotation resources such as Gene Ontology prevents the study of functional consequences of differential splicing in specific contexts. The emergence of third-generation sequencing technologies has allowed the characterization of full-length isoform sequences and motivates the development of methods for isoform-centered functional analysis. Here we document the development of IsoAnnot, a new pipeline for the functional characterization of isoforms. This tool considers an extensive variety of functional properties, both at RNA and protein level. Importantly, most of the functional labels are defined by protein/RNA coordinates which enables the direct mapping of splicing events to functional elements. The pipeline uses transcript sequences to construct an isoform-resolved database of functional annotations by integrating information stored in public databases such as Uniprot and PhosphositePlus and tools based on sequence-prediction such as UTRscan and RepeatMasker. The main advantage of IsoAnnot is its ability to annotate known and novel isoforms obtained from long-read sequencing. IsoAnnot have been developed.

## P-027: EVIDENCE-DRIVEN ANNOTATION OF THE *TRICHECHUS MANATUS LATIROSTRIS* GENOME USING LONG-READS

Alejandro Paniagua[1], Francisco J. Pardo-Palacios[1], Ana Conesa[1]

1. Institute for Integrative Systems Biology (CSIC-UV), Spanish National Research Council, Paterna, Valencia, Spain.
Presenter e-mail: apaniaguadepedro@gmail.com

While the production of a draft genome has become more accessible due to third generations sequencing (TGS), the structural annotation of these new genomes has not been developed at the same pace. TGS of mRNA transcripts could play a key role in the improvement of structural annotation. Previous studies have not assessed the effect of different sequencing technologies or long-reads processing pipelines in structural annotation. In this study, we evaluated the differences in the accuracy of gene prediction based on sequencing technologies, pipelines applied to the sequencing data and the approach used to incorporate this data. Nanopore and PacBio sequencing data from WTC11 human cell line were processed independently and in combination using IsoSeq3 or FLAIR pipelines. The generated transcripts were filtered to produce different sets of reliable non-redundant genes. We tested how the AUGUSTUS ab initio prediction performed on the human genome when trained with these gene sets or the BUSCO genes identified in the human genome, giving the latest approach the best results. We incorporated the raw reads, transcripts or genes generated by the different pipelines and technologies into the gene prediction process using the model trained with the BUSCO genes. The PacBio transcripts in combination with this model showed the best results. The accuracy of the predictions improved from an F-score at gene level of 0.16 in ab initio mode to an F-score of 0.40 when the PacBio data was incorporated. We applied this strategy to the annotation of the *Trichechus manatus latirostris* genome.

**P-028: SENSBIO: A BIOSENSOR DESIGN TOOLBOX FOR METABOLIC PATHWAY SCREENING AND DYNAMIC REGULATION**

Jonathan Tellechea-Luzardo[1], Pablo Carbonell[1,2]

1. Institute of Industrial Control Systems and Computing (ai2), Universitat Politècnica de València, 46022 Valencia, Spain
2. Institute for Integrative Systems Biology I2SysBio, Universitat de València-CSIC, Catedrático Agustín Escardino Benlloch 9, 46980, Paterna, Spain
Presenter e-mail: pablo.carbonell@upv.es

Allosteric transcription factors (TFs) can be used as a genetic biosensor circuit that is induced by the presence of some small chemical, allowing the repression or activation of gene expression. Several hundreds of such biosensors exist in nature, while additional synthetic ones have been also built. There is an increasing interest in using those biosensors in large-scale applications such as bioproduction for automated biofoundries, as they can provide an effective way for performing high-throughput screening and, eventually, introduce synthetic dynamic regulation in the production routes through biomolecular feedback and feedforward systems. However, the current design space for biosensors, which is focused on some specific classes of compounds and cellular processes, needs to be enlarged to ensure that biosensors can be applied within a wider range of chemical targets. Here, we introduce a computational design toolbox for biosensors based on multiple modeling strategies. The toolbox provides a two-way query system where the user can either search for the best TF candidate for a given molecule or, conversely, the most promising inducer molecule for a given TF sequence. The output is based on sequence and chemical homology, which is combined with a score from the TF-chemical latent space of a trained deep-learning model, as well as a structural docking workflow analysis. The extended design space of predicted biosensors is shown to have a larger coverage of the bioproduction space of chemicals that are of interest in the context of the bioeconomy.

## P-029: POST-TRANSCRIPTIONAL CONTROL OF A STEMNESS SIGNATURE BY RNA-BINDING PROTEIN MEX3A REGULATES ADULT NEUROGENESIS

Ana Domingo-Muelas[1-3], Pere Duart-Abadia[1-3], José Manuel Morante-Redolat[1-3], Antonio Jordán-Pla[2], Germán Belenguer[1-3], Jaime Fabra-Beser[1-3], Lucía Paniagua-Herranz[7-9], Ana Pérez-Villalba[1-3], Adrián Álvarez-Varela[4,5], Francisco M. Barriga[5], Cristina Gil-Sanz[1-3], Felipe Ortega[7-9], Eduard Batlle[4-6] and Isabel Fariñas[1-3]

1. Departamento de Biología Celular, Biología Funcional y Antropología Física, Universitat de València, Spain
2. Instituto de Biotecnología y Biomedicina (BIOTECMED), Universitat de València, Spain
3. Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED), Instituto de Salud Carlos III, Madrid, Spain
4. Institute for Research in Biomedicine (IRB), Barcelona, Spain
5. Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Instituto de Salud Carlos III, Madrid, Spain
6. Institución Catalana de Investigación y Estudios Avanzados (ICREA), Barcelona, Spain
7. Departamento de Bioquímica y Biología Molecular, Universidad Complutense de Madrid (UCM), Spain
8. Instituto Universitario de Investigación en Neuroquímica (IUIN), Universidad Complutense de Madrid (UCM), Spain
9. Instituto de Investigación Sanitaria San Carlos (IdISSC), Madrid, Spain
Presenter e-mail: antonio.jordan@uv.es

Neural stem cells (NSCs) in the adult mammalian subependymal zone maintain a glial identity and the developmental potential to generate neurons during the entire lifespan of the animal. Production of neurons from these NSCs follows an orderly pattern of cell-state progressions which allows the gradual increase in the expression of pro-neural factors along the lineage needed to trigger neuronal fates. To molecularly balance stemness and differentiation, specific molecular programs must be expressed to instruct NSCs to either maintain stemness or to progress along the neurogenic lineage. Here, we identify RNA-binding protein MEX3A as a post-transcriptional regulator acting on a set of stemness-associated transcripts at critical transitions in the subependymal neurogenic lineage. MEX3A regulates a quiescence-related RNA signature in activated NSCs that is needed for their return to quiescence, playing a role in the long-term maintenance of the NSC pool. Furthermore, it is required for the repression of the same program at the onset of neuronal differentiation. Our data indicate that MEX3A is a pivotal regulator of adult mammalian neurogenesis modulating gene expression programs at the post-transcriptional level.

## P-030: CHARACTERIZATION OF HUMAN INVERSIONS WITH OXFORD NANOPORE LONG READS

Ricardo Moreira[1], Illya Yakymenko[1], Konstantinos Karakostis[1], Andrés Santos[1,2], Miquel Àngel Senar[3], Jaime Martinez-Urtaza[1,2], Marta Puig[1,2], Mario Cáceres[1,4]

1. Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
2. Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
3. Departament d'Arquitectura d'Ordinadors i Sistemes Operatius, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
4. ICREA, Barcelona, Spain.
Presenter e-mail: Ricardo.moreira@uab.cat

Inversions are a common class of structural variants (SVs) that change the orientation of one segment of the genome, usually without a gain or loss of DNA. The study of inversions has lagged behind due to the technical difficulties in determining the orientation of the inverted segments, which are often flanked by inverted repeats (IRs). The use of new techniques able to cross these large breakpoints is contributing to increase the catalogue of human inversions. However, in most cases just a reduced number of individuals has been studied, which precludes the analysis of the effects of the detected variants. Here, we tested a novel pipeline to genotype inversions by the detection of Oxford Nanopore long reads spanning inversion breakpoints using probe sequences located both outside and inside inverted regions. This methodology was applied to interrogate 132 inversion predictions ranging from 931 bp to 2.2 Mb flanked by (IRs) up to 190 kb in a set of 18 individuals with available long read data. We detected the two orientations in 91 inversions, validating 43 new inversions previously detected with different genome-wide techniques. Moreover, 18 additional SVs were identified during the genotyping process of these regions. By comparison of the inversion genotypes with previous data of the same samples, we find that the methodology is highly accurate. We also observed a correlation of the number of genotypes obtained in a sample with the total number of reads and read length. In fact, as we expected, read length is the main limiting factor for genotyping of inversions with large complex repeats at the breakpoints. The developed methodology therefore shows great promise for the characterization of previously missed inversions in multiple individuals, contributing to a more complete picture of human genetic variation.

## P-031: TOWARDS A COMPLETE CHARACTERIZATION OF HUMAN POLYMORPHIC INVERSIONS AND THEIR FUNCTIONAL EFFECTS

Elena Campoy[1]*, Jon Lerga-Jaso[1]*, Marta Puig[1,2]*, Ruth Gómez Graciani[1], Illya Yakymenko[1], Teresa Soos[1], Alba Vilella-Figuerola[1], Ricardo Moreira[1], Alejandra Delprat[1], Marina Laplana[1], Mario Cáceres[1,3]

1. Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
2. Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
3. ICREA, Barcelona, Spain.
* Equally contributing authors.
Presenter e-mail: elena.campoy@uab.cat

Structural variation (SV) contributes substantially to genetic and phenotypic diversity, but the characterization of these variants is far from complete. Inversions are particularly interesting because of their effects on recombination. However, they are often missed due to their balanced nature, the repetitive sequences at their breakpoints and the fact that many are recurrent. The latest genomic techniques are finally allowing us to obtain a full catalogue of human inversions, although the number of inversions studied in detail is still limited. Here, by careful analysis of >350 predictions from different studies we have generated a reference benchmark of sequence-resolved and manually-annotated polymorphic human inversions (134) and inverted duplications (61). Moreover, each variant has been accurately genotyped in a large number of individuals from diverse populations, representing the most complete resource of this type of SVs to date. Among other things, our unique data set makes finally possible to analyse in depth the potential functional effects of inversions at multiple levels (including gene expression data from different tissues and cell lines and epigenetic signals, such as chromatin accessibility, histone modification and DNA methylation levels) and their association with disease or other phenotypic traits. As an example, a similar analysis of the well-known 8p23.1 and 17q21.31 inversions has found that the two inversions are associated to diverse phenotypes involving brain-related traits, red and white blood cells, lung function, anthropometric measures, male and female characteristics and disease risk. Also, the 17q21.31 inversion acts as lead eQTL of 29 genes located within or close to it. These findings highlight the important role that inversions can play in the human genome and that further investigation of their functional impact is needed.

## P-032: THE ROLE OF RECOMBINATION IN THE ORIGIN AND EVOLUTION OF HUMAN INVERSIONS

Ruth Gómez-Graciani[1,2], Antonio Barbadilla[1,2], Mario Cáceres[1,3]

1. Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
2. Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain.
3. ICREA, Barcelona, Spain.
Presenter e-mail: ruth.gomez.graciani@uab.cat

Chromosomal inversions are structural variants that inhibit recombination in heterozygosis and are involved in multiple evolutionary processes. Historically, *Drosophila* has been the paradigm of the study of inversions unique evolutionary effects, although some key characteristics of this model cannot be translated to other organisms. Here, we investigate the role of recombination in the origin and maintenance of human polymorphic inversions by combining different types of recombination maps and a highly accurate data set of 134 inversions polymorphic in multiple populations, ranging from 0.1 kb to 5.7 Mb. Using pedigree and likelihood-based population recombination maps, we found that inversion location patterns are related to the generation of inversions by non-allelic homologous recombination (NAHR) or non-homologous (NH) mechanisms: NAHR-mediated inversions are more frequent in repetitive regions, while NH inversions generate randomly in the genome. Also, NAHR inversions at frequencies above 0.2 overlap less than expected with the highest 10% and 1% recombination rates in the genome, while NH inversions overlap more than expected with the lowest 10% recombination rates. Using a published set of 813,122 crossovers and 787 aneuploid chromosomes from 20 sperm donors, we observed a decrease in crossover rates within the inverted region in heterozygotes compared to homozygotes, which is especially noticeable in inversions >25 kb. Moreover, we found positive correlations between the number of chromosome aneuploidies and the number of heterozygous inversions in the chromosome, and between the genetic length affected by inversions in heterozygosis and the number of arm losses detected on each chromosome. These results indicate that there is purifying selection against gametes carrying aberrant chromosomes resulting from a crossover between opposite orientations and inversions are more likely to survive in low-recombining contexts. Therefore, new inversions conferring positive evolutionary effects may overcome the potential fitness loss caused by their effect on fertility.

## P-033: MOLECULAR MECHANISMS UNDERLYING NEURON EVOLUTION IN CAENORHABDITIS SPECIES

Adrián Tarazona Sánchez[1], Antonio Jordan[1], Nuria Flames Bonilla[1]

1. Developmental Neurobiology Unit. Instituto de Biomedicina de Valencia (IBV-CSIC), Carrer Jaume Roig 11, 46010 Valencia, Valencia, Spain
Presenter e-mail: atsanchez@ibv.csic.es

Neurons are very specialized cell types able to perform diverse functions. How gene regulatory networs change along evolution to generate complex nervous systems with thousands of neuron types is not well understood. Here we study different *Caenorhabditis* species to provide insights to this question. The structure and development of *C. elegans* nervous system has been very well characterized. An adult hermaphrodite contains 302 neurons that are classified into 118 neuronal types. Moreover, *C. elegans* invariable developmental cellular lineage is conserved within the *Caenorhabditis* genus, facilitating the identification of homologous neurons between different species. To date, the degree of gene expression similarity or divergence among homologous neuron types has not been characterized. The present work presents an evolutionary approach that involves the use of two different species (*C. elegans* and *C. briggsae*) to find the evolutive principles and the changes in the gene regulatory programs by which neurons acquire new functions and evolve into new neuronal types. For this we propose the study of gene expression at single cell resolution using two different levels of analysis. Second larval stage animals are dissociated to obtain single cell suspensions from which transcriptome and open chromatin profiles will be obtained using scRNA-seq and scATAC-seq from 10X genomics technologies. We have already obtained the transcriptome for approximately 25.000 and 35.000 cells of *C. elegans* and *C. briggsae* ¸ respectively. We hope our analysis will allow us to better understand general principles underlying the evolution of nervous systems.

## P-034: RNAget: AN API TO SECURELY RETRIEVE RNA QUANTIFICATIONS

Sean Upchurch[1], Emilio Palumbo[2], Jeremy Adams[3], David Bujold[4], Guillaume Bourque[4], Jared Nedzel[5], Keenan Graham[6], Meenakshi S. Kagda[6], Pedro Assis[6], Benjamin Hitz[6], Emilio Righi[2], Roderic Guigó[2,7], Barbara J. Wold[1] and GA4GH RNA-Seq Task Team[8]

1. Biology and Biomedical Engineering, California Institute of Technology, Pasadena, CA 91125, USA
2. Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Catalonia, Spain
3. Ontario Institute for Cancer Research, Toronto, ON, M5G 0A3, Canada
4. Dept. of Human Genetics, McGill University, Montreal, QC, H3A 0G4, Canada,
5. Broad Institute, Cambridge, MA 02142, USA
6. Stanford University, Stanford, CA 94305, USA
7. Universitat Pompeu Fabra (UPF), Barcelona 08002, Catalonia, Spain
8. Global Alliance for Genomics and Health, Toronto, ON, M5G 0A3, Canada

Presenter e-mail: emilio.palumbo@crg.eu

Availability and implementation: https://github.com/ga4gh-rnaseq/schema

Gene expression quantification by sequencing (e.g., RNA-seq) or by hybridization (e.g., microarrays) are major contemporary research tools for phenotyping human cells and tissues. Rapidly increasing RNA data published worldwide presents compelling opportunities for large-scale data mining from multiple sources. For instance, the European Genome-phenome Archive (EGA), the database of Genotypes and Phenotypes (dbGaP), the Encyclopedia of DNA Elements (ENCODE), the Genotype-Tissue Expression project (GTEx), the National Institute of Health Genomic Data Commons (NIH-GDC), the International Human Epigenome Consortium (IHEC) among others have established large repositories intended for sharing. Yet there are unmet challenges for handling huge numbers of files from diverse sources, coupled with limitations arising from jurisdictional and consent restrictions on data access. To try to overcome these limitations, the Global Alliance for Genomics and Health (GA4GH) international consortium develops and maintains a suite of interoperable standards (Birney et al., 2017). In this context, following the GA4GH vision, a standardized API for the delivery of quantification data from all experimental types and data formats is needed for interoperability and data sharing. To this aim, here we introduce RNAget, an open standard for secure retrieval of expression quantifications drawn from multiple individual samples that is applicable to microarray data and RNAseq from bulk, pseudo-bulk single-cell, or single-cell data. Further, it generalizes to quantification matrices of other sequence-based genomics data such as ATAC-seq and ChIP-seq. This protocol allows a client to retrieve matrices containing data from multiple samples, uses existing community data formats that readily manage dense or sparse data, and provides an option for matrix slicing. RNAget is a part of a family of compatible GA4GH protocols designed to enable efficient and secure discovery and exchange of many types of primary and derived genomic data (Rehm et al., 2021).

## P-035: TOWARDS A GENERIC DATABASE FOR MOLECULAR DYNAMICS

Adam Hospital[1], Daniel Beltrán[1], Aurélien Luciani[1], Genís Bayarri[1], Josep Lluís Gelpí[2,3], Modesto Orozco[1,3]

1. Molecular Modeling and Bioinformatics Group, IRB Barcelona - Institute for Research in Biomedicine, Carrer de Baldiri Reixac 10, 08028 Barcelona, Spain.
2. Barcelona Supercomputing Center (BSC), Plaça d'Eusebi Güell 1-3, 08034 Barcelona, Spain.
3. Departament Biochemistry and Molecular Biomedicine, Univeristy of Barcelona, Barcelona, Spain.
Presenter e-mail: daniel.beltran@irbbarcelona.org

Molecular dynamics (MDs) are crucial to understand how proteins and nucleic acids, among others, behave at the atomic level. However, these models are expensive to calculate and thus it is crucial to share them along the scientific community. Our goal is to store MDs generated by any research group in a reference database and provide open web access to these results. Simulations and their metadata are processed in order to standardize and curate data. Then several analyses are performed and loaded to the database as well. The web client allows to visualize both the stored MDs and their analysis results online.

**P-036: ROSETTA STONE OF BACTERIAL COMMUNICATION. GENETIC AND STRUCTURAL CHARACTERIZATION OF THE RRNPPA FAMILY OF QUORUM SENSING.**

Alonso Felipe Ruiz[1], Eduardo P.C Rocha[2], Alberto Marina[1]

1. Instituto de Biomedicina de Valencia (IBV-CSIC), Jaume Roig Street 11, 46010, Valencia, Spain.
2. Institut Pasteur, Université de Paris Cité, CNRS UMR3525, Microbial Evolutionary Genomics, Paris, France.
Presenter e-mail: aferuiz@ibv.csic.es

Quorum sensing (QS) is a mechanism of bacterial communication, coordinating important decisions depending on bacterial population. QS regulates important processes not only in bacterial behavior but also in genetic mobile elements and host-guest interactions. In Firmicutes, cytoplasmic quorum sensing (QS) receptors of the RRNPPA family regulate important processes such as competence development, sporulation, virulence, and biofilm formation, using peptide-based communication. Despite the importance of such systems, their origin and diversification are poorly understood. In this work, we integrate structural, genomic, and phylogenetic evidence to shed light on RRNPPA protein origin and diversification.

With the combined use of hmm profiles and a large dataset of complete genomes, we analyze the taxonomical distribution of the seven members of the RRNPPA family (Rgg, ComR, PrgX, AimR, NprR, PlcR, and Rap) and inferred their phylogeny. We used an approach to combine structural and sequence data for better phylogenies. The phylogeny of these proteins show that subfamilies diversified a long time ago, resulting in key structural and functional differences. The concordance between the distribution of subfamilies and the bacterial phylogeny was somewhat unexpected, since many of the subfamilies are very abundant in mobile genetic elements, such as phages, plasmids, and phage-plasmids. We also analyzed the different architectures of the peptide pheromone, raising intriguing questions about their export and maturation. The analysis of the peptide architecture also suggests the existence of diverse roles for the RRNPPA systems. Some systems encode multiple pheromones on the same propeptide or multiple similar propeptides, suggesting that they act as “chatterers”. Many others lack pheromone genes and may be “eavesdroppers”.

**P-037: LATE-ONSET DIETARY RESTRICTON LARGELY REVERSES THE ADVERSE AGE-ASSOCIATED CHANGES IN THE FAECAL MICROBIOME AND METABOLOME: A LONGITUDINAL STUDY**

Carolina Monzó[1,2], Lisonia Gkioni[1], Jens Seidel[1], Jenny Fröhlich[1], Aleksandra Placzek[1], Ramona Hoppe[1], Andreas Beyer[2], Dario Riccardo Valenzano[1], Sebastian Grönke[1], Linda Partridge[1,3]

1. Max Planck Institute for Biology of Ageing, Cologne, Germany
2. CECAD Research Center, Cologne, Germany
3. Institute of Healthy Ageing, University College of London, United Kingdom

Presenter e-mail: carolina.monzo@csic.es

The human body is densely populated by microbes. During ageing, the composition of the microbiome changes and recent research suggests that the intestinal microbiome might play a causal role in the determination of organismal lifespan. Dietary restriction (DR) and reduced activity of the insulin/IGF1 (IIS) and mTOR networks increase lifespan and improve health in mice. However, whether changes in the microbiome induced by these longevity interventions contribute to the positive effects on survival is currently unknown. To evaluate how the faecal microbiome and metabolome change in response to longevity interventions in mice, we performed a systematic longitudinal analysis of age-related changes in the faecal microbiome and metabolome of mice under DR. In addition, to identify changes in the microbiome that are common between longevity interventions, we analysed the microbiome of mice with reduced IIS and mTOR signalling. We show that DR mitigates age-related changes in microbiome community structure, including the decline in alpha diversity, increase in beta diversity, and loss of equilibrium between *Firmicutes* and *Bacteroidetes* phyla. Furthermore, by studying the faecal microbiome and metabolome of late-life DR switches that either increase or not increase lifespan, we identified an age-dependent memory of AL feeding in the microbiome associated with lifespan. Finally, we identified bacteria that were shared between DR, reduced mTOR and IIS signalling, suggesting that these bacteria may contribute to the health benefits observed in these long-lived mouse models.

**P-038: DNAffinity: A MACHINE-LEARNING APPROACH TO PREDICT DNA BINDING AFFINITIES OF TRANSCRIPTION FACTORS**

Sandro Barissi[1,†], Alba Sala[1,†], Milosz Wieczor[1,2], Federica Battistini[1,3,*] and Modesto Orozco[1,3,*]

1. Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology. Baldiri Reixac 10–12, 08028 Barcelona, Spain
2. Department of Physical Chemistry. Gdansk University of Technology, 80-233 Gdansk, Poland
3. Department of Biochemistry and Molecular Biology. University of Barcelona, 08028 Barcelona, Spain
Presenter e-mail: alba.sala@irbbarcelona.org

Proteins are the main regulators of gene expression as they can directly or indirectly inactivate, activate, or enhance the transcription of DNA. Central elements in this regulatory system are transcription factors (TFs): modular proteins that recognize sequences of DNA (typically 6-20 base pair long), helping to recruit RNA polymerases that trigger the subsequent transcription of a nearby gene. The binding of TFs during normal cell life is difficult to predict as it is modulated by a myriad of effects, such as the presence of nucleosomes (which in general hinders TF binding), or the formation of clusters that foster cooperativity, general chromatin compaction, or even phase separation. However, a key requirement for in vivo binding is a good binding to the targeted naked DNA.

Here we present a physics-based machine learning approach to predict in vitro transcription factor binding affinities from structural and mechanical DNA properties directly derived from atomistic molecular dynamics simulations. The method is able to predict affinities obtained with techniques as different as uPBM, gcPBM and HT-SELEX with an excellent performance, outperforming existing state-of-the-art algorithms. Due to its nature, the method can be extended to epigenetic variants, mismatches, mutations, or any non-coding nucleobases. When complemented with chromatin structure information, our in vitro trained method provides also good estimates of in vivo binding sites in yeast.

**P-039: BIOINFORMATIC ANALYSES OF OLIVE POLLEN TUBE AND SEED TRANSCRIPTOMES AND THEIR INTEGRATION IN AN INTERACTIVE OLIVE EXPRESSION ATLAS**

Amanda Bullones[1,2], Noe Fernandez-Pozo[1], Elena Lima-Cabello[3], Antonio Jesús Castro[3], Juan de Dios Alché[3], M. Gonzalo Claros[1,2,4,5]

1. Institute for Mediterranean and Subtropical Horticulture "La Mayora" (IHSM-CSIC-UMA), Málaga, Spain
2. Department of Biochemistry and Molecular Biology, Universidad de Málaga (UMA), Málaga, Spain
3. Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas (CSIC), Granada, Spain
4. CIBER de Enfermedades Raras (CIBERER), Málaga, Spain,
5. Institute of Biomedical Research in Malaga (IBIMA), IBIMA-RARE, Málaga, Spain
Presenter e-mail: amandabullones@uma.es

The olive tree (Olea europaea L.) is of great socioeconomic importance due to the production of oil and table olives, and the seed is also an emergent source of nutraceuticals (Jiménez-Ruiz et al., 2020) and feed supplements (Maestri et al., 2019). Spain is the world's leading producer and, of the 272 varieties present in our country, Picual is the most cultivated and the main source of olive oil. Despite its relevance, there are many unknown aspects in its biology. Here we study the development of the Picual pollen tube to discover the biological processes involved and its potential regulation. Highly relevant transcripts like those of two RING proteins and other two Zinc-finger containing proteins seem to indicate that they play a main role in the process of mitosis II, and its cross-talk with environmental and developmental clues. In addition, Picual orthologues of the 14 olive allergens known to date were determined and their expression in the pollen tube monitored. Allergens are consistently expressed during pollen tube development, most of them at a high level. Despite Ole e 1 being the first olive pollen allergen to be identified (Villalba et al., 1993), it is the second most abundant transcript, behind Ole e 14 which was only recently described (Oeo-Santos et al., 2018), with Ole e 15 being the last allergen identified (San Segundo-Acosta et al., 2019) and the one less expressed. We are characterizing the transcriptome of Picual seeds at different developmental stages, where the biological processes of the endosperm that nourish the embryo, and the establishment of embryonic polarity are particularly overrepresented. Finally, an interactive gene expression atlas for olive tree (OliveAtlas) where expression data were mapped to the Picual reference genome and its gene model annotation was created. This tool tries to ease the lack of bioinformatics and genomics resources and assists olive research and breeding. This research was founded by grants PID2020-113324GB-I00 (AEI), P18-RT-1577 (JA, excellent research), and UMA20-FEDERJA-029 (JA), all of them co-founded by ERDF/EU.

## P-040: EXPLORING SYMBIOTIC CROSS-TALK IN *BLATTELLA GERMANICA* USING RIFAMPICIN AND METAGENOMICS

Monica Cazzaniga[1], Rebeca Domínguez-Santos[1,2], Jesús Marín-Miret[1,2], Amparo Latorre[1,2], Rosario Gil[1,2], Carlos García-Ferris[1,2,3]

1. Institute for Integrative Systems Biology (I2SysBio), University of Valencia and Spanish Research Council, 46980 Paterna (Valencia), Spain
2. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region, 46020 Valencia, Spain
3. Department of Biochemistry and Molecular Biology, University of Valencia, 46100 Burjassot (Valencia), Spain
Presenter e-mail: carlos.garcia.ferris@uv.es

The German cockroach Blattella germanica harbours two cohabiting symbiotic systems: an obligate endosymbiont, Blattabacterium, located inside bacteriocytes in the fat body, which is key in essential amino acid production and nitrogen recycling of urate deposits, and an abundant and complex gut microbiota that must play an important role (as yet unknown) in host physiology. While Blattabacterium is transmitted vertically from mother to oocytes, the gut microbiota is acquired by horizontal transmission (mainly coprophagy). In this work, we used the antibiotic rifampicin to deepen knowledge on the relationship between the host and the two symbiotic systems. Rifampicin treatment of adults in G1 does not affects the endosymbiont in the fat body but reduces notably the oocytes infection, when Blattabacterium is extracellularly located in the ovaries, blocking the vertical transmission to the next generation. Consequently, adults in G2 showed a reduction of 5-6 orders of magnitude in the Blattabacterium load (quasi-aposymbiont cockroaches), together with a development delay, and were unable to give rise to the G3. On the other hand, we performed a blind experiment where quasi-aposymbiotic nymphs (obtained as G2 after rifampicin treatment of adults in G1) were mixed with control ones, sharing environment and microbiota input (faeces for coprophagy). The mixed population was raised until the adult stage, where the two symbiotic systems were analyzed in each individual: Blattabacterium density by RT-PCR, and the hingut microbiota composition and structure by metagenomics. Our results revealed that the reduction in the endosymbiont's load –even though it is associated with a delay in the development time– does not affect the hindgut microbiota's composition and structure. We conclude that there is no cross-talk between the two symbiotic systems in B. germanica, and that the microbiota cannot compensate the essential functions performed by Blattabacterium.

## P-041: TRANSCRIPTOME INNOVATIONS IN PRIMATES REVEALED BY SINGLE-MOLECULE LONG-READ SEQUENCING

Luis Ferrández-Peral[1], Xiaoyu Zhan[2], Marina Álvarez-Estapé[1], Cristina Chiva[3,4], Paula Esteller-Cucala[1], Raquel García-Pérez[1], Eva Julià[3], Esther Lizano[1,8], Òscar Fornas[3,4], Eduard Sabidó[3,4], Qiye Li[2,5], Tomàs Marquès-Bonet[1,4,6,7,8], David Juan[1], Guojie Zhang[2,9,10,11]

1. Institute of Evolutionary Biology (UPF-CSIC), PRBB, 08003 Barcelona, Spain
2. BGI-Shenzhen, Shenzhen 518083, China
3. Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08003 Barcelona, Spain
4. Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
5. College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
6. Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain
7. CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain
8. Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, 08193 Barcelona, Spain
9. State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China
10. Section for Ecology and Evolution, Department of Biology, University of Copenhagen, DK-DK-2100 Copenhagen 2200, Denmark
11. Evolutionary and Organismal Biology Research Center, School of Medicine, Zhejiang University, Hangzhou 310058, China

Presenter e-mail: lj.ferrandez@gmail.com

Transcriptomic diversity greatly contributes to the fundamentals of disease, lineage-specific biology, and environmental adaptation. However, much of the actual isoform repertoire contributing to shaping primate evolution remains unknown. Here, we combined deep long- and short-read sequencing complemented with mass spectrometry proteomics in a panel of lymphoblastoid cell lines (LCLs) from human, three other great apes, and rhesus macaque, producing the largest full-length isoform catalog in primates to date. Around half of the captured isoforms are not annotated in their reference genomes, significantly expanding the gene models in primates. Furthermore, our comparative analyses unveil hundreds of transcriptomic innovations and isoform usage changes related to immune function and immunological disorders. The confluence of these evolutionary innovations with signals of positive selection and their limited impact in the proteome points to changes in alternative splicing in genes involved in immune response as an important target of recent regulatory divergence in primates.

**P-042: SEQUENCE AND STRUCTURE PREDICTION OF *GRAPHOLITA MOLESTA* BUSK (LEPIDOPTERA; TORTRICIDAE) CADHERIN: A CANDIDATE RECEPTOR FOR *BACILLUS THURINGIENSIS* BERLINER, CRY1A PROTEIN**

Dafne Toledo[1], M. Cristina Crava[1], Yolanda Bel[1], Baltasar Escriche[1]

1. Laboratory of Biotechnological Control of Pests, BioTecMed Institut, Genetics Department, University of Valencia, Burjassot, Valencia, Spain.
Presenter e-mail: dafne.toledo@uv.es

Cadherins are a big family of transmembrane proteins involved in crucial physiological processes such as cell structural integrity, migration, and signaling. The structure of these proteins consists of 3 domains: extracellular, transmembrane, and intracellular. A variable number of the called “cadherin repeats” (CRs) are located at the extracellular domain.
In Lepidoptera, midgut cadherins have been extensively studied for their role as receptors of the insecticidal Cry proteins, which are synthesized by the entomopathogenic bacterium *Bacillus thuringiensis* (Bt). It has been described that the interaction between Cry toxins and cadherin-like receptors promotes the oligomerization of the Cry proteins and the insertion of the oligomers in the epithelial cell membrane, leading to the disruption of gut integrity. This interaction happens between the exposed loops of Domain II of the Cry protein and the cadherin toxin binding regions, which are located in the extracellular CRs. The resistance of several lepidopteran populations to the Cry toxins has been associated with the presence of mutations in the cadherin-like genes, highlighting the importance of these receptors in the mode of action of Cry proteins, specifically in the steps prior to their insertion into the insect midgut cells.
The main objective of the present study was to identify the putative cadherin-like receptor of Cry proteins in stone-fruit crops pest *Grapholita molesta* Busk (Lepidoptera; Tortricidae). We used the cadherin of *Ostrinia nubilalis* Hübner, (Lepidoptera; Crambidae, Acc. number: AAY44392.1) as a query to search for similar sequences in the *G. molesta* genome. An mRNA of 4839 base pairs on the positive strand was obtained, formed by 29 exons coding for a protein of 1612 amino acids. This predicted protein had high homology with the cadherin 99-C from *Leguminivora glycinivorella* Matsumara (Lepidoptera; Tortricidae). From the predicted structure, the *G. molesta* protein comprises 11 CRs, a transmembrane region, and a small intracellular part. The structures of the *O. nubilalis* and *G. molesta* cadherins modeled *in silico* showed a similar three-dimensional conformation despite their low amino acid identity (21%).
This study describes for the first time a Cry1A cadherin-like candidate receptor in *G. molesta*. This work will help in the study of the mode of action of Bt insecticidal proteins in this lepidopteran pest and would provide tools for the early detection of resistance outbreaks.

**P-043: FIRST STEM FORM ENDOSYMBIONT TO CHASSIS IN SYNTHETIC BIOLOGY: HELPING *BARTONELLA QUINTANA* TO GROW MORE AND BETTER**

Emilio Garrote-Sánchez[1,2], Christian Seeger[3], Andrés Moya[1,2], Rosario Gil[1,2]

1. Evolutionary Systems Biology of Symbionts, Institute for Integrative Systems Biology (I2SysBio), University of Valencia and Spanish Research Council, 46980 Paterna (València), Spain.
2. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region (FISABIO), 46020 València, Spain.
3. Department of Cell and Molecular Biology, Uppsala University, 752 36 Uppsala, Sweden.
Presenter e-mail: emilio.garrote@uv.es

Obligatory endosymbiotic organisms, whether parasitic or mutualistic, tend to have reduced genomes compared to their free-living relatives, as a result of the evolutionary process called 'genomic reduction syndrome'. Yet, these streamlined genomes maintain those genes involved in essential functions, getting close to the definition of a minimal genome. Their characterization, as well as the possibility of optimizing them, by eliminating superfluous genes or by adding genes to complete impaired metabolic pathways, is highly relevant in the field of synthetic biology. However, most endosymbionts cannot be cultured in the laboratory, making it difficult to manipulate them.
*Bartonella quintana* Toulouse, our model endosymbiont, can be grown in culture and has the capacity to infect mammalian cells, making this bacterium a good model to design a chassis for potential biomedical applications. However, it has a very slow growth rate due to its complex nutritional requirements. In our research group, we intend to define the ideal medium composition that improves its growth efficiency, which will also have an impact on the ease of performing genomic manipulation experiments for a better characterization of the model prior of its use as an endosymbiont chassis. First, we generated a metabolic model of *B. quintana* from genomic data and, through flux balance analysis (FBA), we determined which compounds are limiting factors for its growth and are deficient in commercial media. Next, we established a protocol for culturing this bacterium using media supplemented with these compounds in different concentrations and measured its growth impact. Finally, we have performed proteomics studies to determine which changes could be linked to the results observed.

## P-044: ANALYSIS OF DIFFERENTIALLY METHYLATED REGIONS IN THE EXOME OF PATIENTS WITH TYPE II DIABETES

Celeste Moya-Valera[1], Francisco Lara-Hernández[1], Elena Quiroz[1], Veronica Lendinez[1], Rebeca Melero-Valverde[1], Wladimiro Díaz[2], Vicente Arnau[2], Lisardo Fernández[3], Juan M. Orduña[3], Sergio Valdés[4,5], Ana Lago-Sampedro[4,5], Gemma Rojo[4,5], F. Javier Chaves[1,4], Ana Bárbara García-García[1,4]

1. Unidad de Genómica y Diabetes. INCLIVA, Valencia, Spain
2. Institute of Integrative Systems Biology (I2SysBio), University of València and Consejo Superior de Investigaciones Científicas (CSIC), València, Spain, 2 Instituto de Biotecnología y Biomedicina (Biotecmed), Universitat de València, Burjassot, Spain
3. Departamento de Informática, Universidad de Valencia, Avda. Universidad, s/n, Burjassot (Valencia), Spain
4. CIBERDEM, ISCIII, Madrid, Spain
5. Biomedical Research Institute of Málaga (IBIMA), Endocrinology and Nutrition Department. Regional University Hospital of Malaga. Málaga.
Presenter e-mail: ceromova@alumni.uv.es

Type 2 diabetes (T2D) is a complex disease that accounts for approximately 90% of all Diabetes mellitus cases. In general, it is mediated by obesity-associated insulin resistance (IR) leading to pancreatic ß-cell failure. DNA methylation is one of the mechanisms involved in its development through epigenetic changes that may come from fetal and neonatal development and changes that occur throughout the lifetime of each patient. Methylation of the human exome appears to play an important role in the regulation of gene expression. Differentially methylated regions (DMRs) can have a relevant role in genetic regulation and in metabolic changes involved in T2D development. The aim of this study is to identify DMRs present in the human exome and related to T2D development in the Spanish population that may be present in the patient's DNA years before the establishment of the disease; also studying these alterations in relation to IR as an associated factor. We have analyzed the methylation in the exome of 24 patients who have developed T2D, mediated by IR or not. Bicycle and HPG-Msuite pipelines were used to detect DMRs, and the results were intersected to obtain reliable and robust results. The detected genes were selected based on statistical parameters. The genes are IRS2, ADARB2, KLK7, OAZ1, FEM1A, UTP11, DNASE2 and the pseudogene HSPD1P4. Several of these genes may have a relevant role in the development of T2D.

Funding:

Research grants PMP21/00076, PI17/00544 and PI14/00874 from the “Instituto de Salud Carlos III” (Ministerio de Ciencia e Innovación, Spanish Government), and CIBER- Consorcio Centro de Investigación Biomédica en Red-(CB07/08/0018), Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación, Spanish Government.

## P-045: BIOGENOME DATA PORTAL

Emilio Righi[1], Emilio Palumbo[1], Ferriol Calvet[1], Roderic Guigó[1]

1.Centre for Genomic Regulation (CRG), Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain
Presenter e-mail: erighi@crg.es

Since the Earth BioGenome Project (EBP) started, many genome assemblies have been generated and these will increase exponentially in the next few years. The generated sequences and their metadata are deposited in the International Nucleotide Sequence Database Collaboration (INSDC) which is a collaboration between the major biological databases: DDBJ, NCBI and ENA. However, querying these databases can often be difficult, in particular for non-technical users. We have implemented a user-friendly web application which aims to help users and researchers to manage INSDC and local data in one place: The BioGenome Data Portal.
The BioGenome Data Portal (https://github.com/guigolab/biogenome-portal ) can retrieve any existing record from the INSDC database (as assemblies, reads and biosample) via a user-friendly form, and it contains a cutting-edge Genome Browser (JBrowse2 https://jbrowse.org/jb2/ ) to help visualize sequences and annotations (in FASTA and GFF format) related to that assembly.
The BioGenome Data Portal can also be used to disseminate information on endemic species, offering the possibility of adding vernacular names of the species related to the language of that region, photos of the collected specimens and links to relevant scientific publications.
The BioGenome Data Portal can also import all the data linked to a specific INSDC BioProject accession, as the EBP or its children, and start tracking its progresses: when a new assembly is added, the portal will automatically import its metadata as well as its related reads and biosamples.
An additional functionality of the BioGenome Data Portal is to import sample's metadata from a spreadsheet file, thus helping researchers manage spreadsheet metadata in a user-friendly way.
Finally, the BioGenome Data Portal is totally configurable and portable, making it easy to run locally or to deploy it in a server machine.
A BioGenome Data Portal instance tracking the data generated under the EBP umbrella can be found at the following link: https://ebp.biogenoma.cat.

## P-046: STUDYING RELATIVE RNA LOCALIZATION - FROM NUCLEUS TO THE CYTOPLASM

Vasilis F. Ntasis[1], Roderic Guigó Serra[1,2]

1. Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Catalonia, Spain
2. Department of Experimental and Health Sciences (DCEXS), Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
Presenter e-mail: vasilis.ntasis@crg.eu

The precise coordination of important biological processes, such as differentiation and development, is highly dependent on the regulation of expression of the genetic information. The flow of the genetic information is tightly regulated on multiple levels. Among them, RNA export to cytoplasm is an essential step for the production of proteins in eukaryotic cells. Hence, estimating transcript relative localization, that is the proportion of total RNA molecules in eukaryotic cells present in the cytosol or in the nucleus, is of major significance. However, most studies with a focus on transcriptome analysis ignore subcellular RNA localization. Those with an effort to take that into account, utilize RNA sequencing (RNA-seq) in combination with cellular fractionation. Nevertheless, transcript quantification estimates obtained independently from nuclear and cytosolic RNA cannot be compared, as the total amount of RNA in each of these cellular compartments is usually unknown. Here we show that if, in addition to nuclear and cytosolic RNA-seq, whole cell RNA-seq is also performed, then accurate estimations of the relative localization of transcripts can be obtained. We first establish the theoretical basis that supports this by formalizing mathematically the relationship between the different RNA abundances. Based on this, we designed a method that estimates for every transcript its relative localization. Fianally, we evaluate our methodology first on simulated, and then on real bulk RNA-seq data from the ENCODE project.

## P-047: THE IMPORTANCE OF ROTATING SCHEDULES ON WORKERS' TRANSCRIPTOME - THE HORMONIT STUDY

Tamara Perteghella[1], Valentin Wucher[1], Roderic Guigó Serra[1].

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

Presenter e-mail: tamara.perteghella@crg.es

The impact of mistimed schedules on worker's health recently gained increased attention, being associated to higher incidence of several health impairments. Nonetheless, when it comes to the molecular mechanisms relating shift work to its pathological outcomes, there is still much to unveil. The present study investigates the impact of rotating working schedules on the transcriptome of 47 male workers from a car factory in Spain. The greatest changes in expression profiles have been reported during the night shift and impact mostly processes related to the immune system. Several genes implicated in cell chemotaxis, cytokines production and signalling are indeed down-regulated, in agreement with a decrease in the level of these analytes in the blood. Consistently, when testing for impairments in established blood transcription modules different types of cell signalling, general immune activation and inflammatory response are reported significantly down-regulated after night shift. Conversely, the modules up-regulated mainly involve the activation and differentiation of lymphocytes. While these findings are partially supported by the results of cellular deconvolution, the cell type proportions obtained align to some extent with the internal circadian phase of the subjects, being in agreement also with the adjustment of acrophase reported in several hormones (e.g., Cortisol). Despite this suggests the possibility of a partial adaptation to the mistimed working conditions, overall, many signalling and immune related processes are altered in the population, an effect that seems exacerbated when working at night.

## P-048: AURORA KINASE B INHIBITORS: A SEARCH FOR NEW COMBINATORIAL STRATEGIES TO TARGET EWING SARCOMA

María Sánchez-Jiménez[1,2], Enrique Blanco[3], Pablo Táboas[1,2], Carla Rios-Astorch[1,2], Elisabet Figuerola-Bou[1,2], Jaume Mora[1,2] and Sara Sánchez-Molina[1,2]

1. Developmental Tumor Biology Laboratory, Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat 08950 Barcelona, Spain
2. Pediatric Cancer Center Barcelona (PCCB), Hospital Sant Joan de Déu, Esplugues de Llobregat, 08950 Barcelona, Spain
3. Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, 08003 Barcelona, Spain
Presenter e-mail: maria.sanchezj@sjd.es

Ewing Sarcoma (EwS) is the second most common malignant neoplasm appearing in the bone and soft tissues that predominantly affects adolescents and young adults. EwS is an exclusive human disease cytogenetically characterized by a chromosomal translocation most commonly affecting chromosomes 11 and 22, t(11;22)(q24;q12). The resulting fusion protein, EWSR1-FLI1, has the ability to act as an aberrant transcription factor leading both to gene activation and repression of a large set of genes involved in the tumorigenic phenotype of EwS. Our group has recently demonstrated that the Polycomb subunit with E3 ubiquitin ligase activity, RING1B, co-localizes genome–wide with EWSR1-FLI1 at active enhancers having a key role in EwS tumorigenesis *in vivo*, mainly by regulating enhancers and recruiting the fusion oncogene to key targets such as *NKX2-2*, *SOX-2* and *IGF-1*.

Aurora Kinases are conserved serine/threonine kinases with high expression in many different tumors that play multiple roles in cell division and maintenance of genomic stability. We have demonstrated that EwS cell lines are specifically and highly vulnerable to inhibition by Aurora Kinase B (AURKB) inhibitors. In the present study, using ChIP-seq technology, we characterize AURKB genomic distribution and analyse its colocalization with EWSR1-FLI1 oncogene. Moreover, we described the contribution of AURKB in the activation of RING1B-EWSR1-FLI1 enhancers necessary for EwS tumorigenesis.

## P-049: METHYLCLOCK: A BIOCONDUCTOR PACKAGE TO ESTIMATE DNA METHYLATION AGE

Dolors Pelegrí-Sisó[1], Paula de Prado[1], Justiina Ronkainen[2], Mariona Bustamante[1,3], Juan R González[1,3]

1. Barcelona Research Institute for Global Health (ISGlobal), Doctor Aiguader 88, 08003 Barcelona, Spain.
2. Center for Life Course Health Research, University of Oulu, Pentti Kaiteran katu 1, Linnanmaa, Oulu, Finland.
3. CIBER in Epidemiology (CIBERESP), Av. Monforte de Lemos, 3-5. Pabellón 11. Planta 0 28029 Madrid, Spain
Presenter e-mail: dolors.pelegri@isglobal.org

Ageing is a biological and psychosocial process related to diseases and mortality. It correlates with changes in DNA methylation (DNAm) in all human tissues. Therefore, epigenetic markers can be used to estimate biological age using DNAm profiling across tissues. Alterations of the correlation between biological and chronological age (i.e. age acceleration or deceleration) have been related to several age-related diseases, pathological stages and can occur before clinical manifestation of diseases become overt. Therefore, epigenetic clocks are being considered as a tool in prevention, diagnostic and even in forensic applications. In 2021, we developed methylclock, a Bioconductor package to compute epigenetic age. Since that moment, the package is being routinely updated with new epigenetic clocks. So far, it includes 15 DNAm adult/childhood and gestational clocks. They range from the classic Hovarth's clock to more recent ones like EPIC predictor of gestational age. Methylclock has been used, among many others projects, to investigate the association between the early life exposome and epigenetic age acceleration in children using the Pediatric-Buccal- Epigenetics' clock. It has also been used to characterize the ethnic differences in DNA methylation between UK-residents, South Asians and Europeans. Another interesting study using methyclock showed that age acceleration is associated with alcohol use disorder, providing the first evidence for a recovery of this effect upon abstinence from alcohol. methylclock package is in continuous development and new clocks are added, some of them requested by researchers. Interestingly, there are some companies such as TruDiagnostic® who implemented our tool in their pipelines. In summary, methylclock enables quick and user-friendly computation of a large set of epigenetic clocks that allows the investigation of this biomarkers in different research areas aiming at improving personalized medicine and public health. In this talk, we will present how to use methyclock with several real cases.

**P-050: A PANCESTROME AND A NEW COMPLETE INFERENCE FOR AN ANCESTRAL GENOMIC SEQUENCE OF *MYCOBACTERIUM TUBERCULOSIS* COMPLEX**

Paula Ruiz-Rodriguez[1], Mireia Coscolla[1]

1. Institute for Integrative Systems Biology (I$^2$SysBio), University of Valencia-CSIC, FISABIO Joint Unit, 46980, Valencia (Spain)
Presenter e-mail: paula.ruiz-rodriguez@uv.es

*Mycobacterium tuberculosis* is currently one of the deadliest pathogens worldwide, being the second deadliest only overcome by SARS-CoV-2. The study of its diversity could help revealing keys to its virulence. *M. tuberculosis* was classified as almost a clone, but we can observe lineages that affect humans (L1-9), and lineages that mainly affect other animal hosts (A1-4). Most genetic analyses of MTBC are based on Illumina mapping, using the gold standard of a reference based on the ancestor of *M. tuberculosis*. This reference has the genomic length of the lineage 4 strain H37Rv and lacks genes that are not present in this strain. We analyzed 67 closed genomes available from public databases and 10 newly sequenced complete genomes to infer a new *M. tuberculosis* ancestor. A total of 77 closed genomes and 26,810 short-read sequences were considered to perform a Bayesian inference of the ancestral sequence harboring all MTBC diversity. Also, we generated a new annotation for this new ancestor of MTBC resulted from a pancestrome for all lineages of MTBC, which was made with multiple annotators. Finally, this new reference and pancestrome were used to infer genomic events during the diversification of host-specific *M. tuberculosis* lineages. Host-specific genomic markers have the potential to identify host-specific virulence markers.

This work is supported by the Spanish Ministry of Science, Innovation and Universities (MICIN) and co-funded by the FEDER Funds (PID2021-123443OB-I00), Ramón y Cajal fellowship (RYC-2015-18213 MICIN), and the European Comission—NextGenerationEU (Regulation EU 2020/2094), through CSIC's Global Health Program (PTI+ Salud Global).

**P-051: A WOLF IN SPIDERS' CLOTHING: GENOMIC-WIDE SCREENING UNCOVERS ADMIXTURE BETWEEN MADEIRA WOLF SPIDERS**

Alba Enguídanos García[1,2], Vanina Tonzo[3], Luis C. Crespo[1,4], Miquel A. Arnedo[1,2]

1.Departament Biologia Evolutiva, Ecologia i Ciències Ambientals, Universitat de Barcelona (UB), Avinguda Diagonal 643, 08028 Barcelona, Spain
2. Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona (UB), Barcelona, Spain
3. Centre d'Ecologie Fonctionnelle et Evolutive (CEFE-CNRS). 1919, route de Mende, 34293 Montpellier 5, France
4. Laboratory for Integrative Biodiversity Research (LIBRe), Finnish Museum of Natural History (LUOMUS), University of Helsinki, P.O. Box 17, 00014 Helsinki, Finland
Presenter e-mail: albaengarcia@gmail.com

Seven endemic species of the wolf spider genus *Hogna* Simon, 1885 have been documented in the volcanic archipelago of Madeira. Despite their colourful bodies, large size and widespread distribution across the islands, they have received little attention. A recent study, which integrated molecular markers and morphological information, supported the current delimitation of most nominal species. However, the species pair *Hogna insularum* (Kulczyński, 1899), present in all islands, and the large *H. maderiana* (Walckenaer, 1837), circumscribed to Porto Santo, constituted an exception. Although they are morphologically diagnosable and differ in their preferred habitats, mitochondrial data failed to distinguish them. Of the two main lineages recovered, one was exclusively formed by *H. insularum* haplotypes from Madeira Island, while the second one was a mixture between the remaining *H. insularum* haplotypes and those of *H. maderiana*. To test alternative hypotheses that may explain the observed inconsistencies, we generated double-digest restriction site-associated DNA sequencing (ddRADseq) markers from specimens sampled across most known populations of both species. Our genome-wide analysis revealed the existence of two different groups, but unlike the mitochondrial data, one consisted of *H. insularum* specimens from Madeira and Desertas Islands, while the second one was exclusively of *H. maderiana* specimens from Porto Santo. However, several *H. insularum* specimens from Porto Santo showed similar levels of introgression when compared to the two groups, suggesting hybridization events between the two species on this island. Similar patterns of introgression were also reported in wolf spiders from the Galapagos.

## P-052: SHOTGUN METAGENOMIC PIPELINES COMPARISON FOR ANALYSING GUT MICROBIOME IN ATOPIC DERMATITIS PEDIATRIC PATIENTS

Andrea Rodenes[1], Araceli Lamelas[1], Vanessa Illescas [1], Cristina Moya [1], Mª Carmen Solaz [1], Juan F Martínez-Blanch[1], Empar Chenoll[1]

1. Archer Daniels Midland Co-Biopolis, 46980 Valencia, Spain
Presenter e-mail: andrea.rodenes@adm.com

Human gut microbial community studies have increased faster over the last decade along with the continuous improvements in high throughput sequencing technology and shotgun metagenome bioinformatic pipelines for microbiota analysis. The choice of the analysis method will be critical in the final results since they present differences in terms of pipeline structure and databases. However, there is limited literature available comparing the features, advantages and disadvantages of each pipeline, making the pipeline choice to use often unclear. In this study, we applied Metaphlan3, Metaphlan 4, Humann3 and internal pipeline, to analyse taxonomically and functionality the faecal microbiota associated with Atopic Dermatitis pediatric patients belonging to a published clinical trial (Clinicaltrials.gov identifier: NCT02585986). The principal parameters which differed between the pipelines were the number of reads used in the estimation of microbial composition, Shannon index, Simpson index and Richness, and number of genes predicted. Our results highlight the differences in taxonomic and functional compositions of samples obtained from the different pipelines, and their impact on downstream analyses.

## P-053: IN HOUSE TOOLS FOR PUBLIC META-OMICS METADATA CURATION

Samuel Piquer-Esteban[1,2], Wladimiro Diaz-Villanueva[1], Vicente Arnau[1], 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

Despite advances in data and metadata sharing in science, interacting and working with different public datasets remains challenging for meta-omics. While it is true that there are some initiatives designed for data reuse, most are focused on human data. To the best of our knowledge, currently, there are no tools to facilitate the in-house curation process of this type of data. Here, we present a series of scripts designed to obtain and curate metadata and FASTQ files associated with research projects hosted in the European Nucleotide Archive (ENA).The main functions include metadata downloading, checking, filtering, FASTQ files downloading, reviewing, and treatment (concatenation and renaming). Furthermore, we have reviewed more than 200 ENA projects, and applied that acquired knowledge to specially designed checking scripts to accompany and guide the researcher during the curation process. The objective is, therefore, to provide a uniform and reproducible workflow to simplify the curation of public sequencing data.

## P-054: ANALYSING THE EFFECT OF AN ANTIBIOTIC TREATMENT IN THE GUT MICROBIOME OF *BLATTELLA GERMANICA*

Irene Creus-Martí [1,2], Jesús Marín-Miret[1], Andrés Moya [1,3,4,*], Francisco J. Santonja[2,*]

1. Institute for Integrative Systems Biology (I2Sysbio), Universitat de València and CSIC, València, Spain
2. Department of Statistics and Operation Research, Universitat de València, Valencia, Spain
3. Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), Valencia, Spain
4. CIBER in Epidemiology and Public Health (CIBERESP), Madrid, Spain
Presenter e-mail: icreus@alumni.uv.es

The gut microbiome is related to the health status of the host. It is known nowadays that antibiotic intake alters microbiota composition. However, there is not much knowledge about the dynamics and interactions between microbial taxa. In this work, we analyze longitudinal data of the gut microbiome in *B. germanica* treated with kanamycin. In three periods, the antibiotic was supplied, giving the microbiota time to recover from one antibiotic intake to the other. To carry out the analysis, we use two new statistical models that are not focused on pair-wise interactions and, as a result, allow extracting information about the interactions between groups of bacterial taxa and how some groups affect single taxa. These models will enable us to identify how the antibiotic affects each bacteria: some are community-dependent, others are adapted, and other works together to overcome the antibiotic. These models allow recalling that the community works together in antibiotic conditions.

## P-055: PREDICTION OF PRIMER DIMER FORMATION AND OFF-TARGET AMPLIFICATION APPLIED TO TARGETED SEQUENCING DATA

Rusu EC[1], Cedeño JM[1], Olivares MD[1], Arnau V[2], Diaz-Villanueva W[2], Ivorra C[1]

1. SeqPlexing SL, Catedrático José Beltrán Martínez 2, 46980 Paterna, Valencia, Spain
2. Institute of Integrative Systems Biology (I2Sysbio), University of Valencia and CSIC, Catedrático Agustín Escardino Belloch, 46980 Paterna, Valencia, Spain
Presenter e-mail: rusu@seqplexing.com

Amplicon-based Next-Generation Sequencing (NGS) yields high coverage in an affordable way by specifically amplifying regions of interest in a multiplex PCR. The possibility of designing novel panels targeting new regions is appealing for researchers and clinicians alike, although optimizing the multiplex PCR is still a major challenge. Dimer structures between primer sequences and off-target PCR products are especially common and troublesome. In this work, we have developed a pipeline to predict primer dimer and off-target formation, based on DNA duplex thermodynamics and genome specificity. To assess performance, we have sequenced 40 samples using an 254-plex amplicon panel and classified short and unmapped reads as dimers or off-targets depending on length and sequence. Interestingly, primers that were in different pools during the first PCR but later pooled together in the nested PCR also displayed dimer and off-target formation. Compared to the actual number of reads identified as dimer or off-target sequences, our prediction pipeline demostrated a sensitivity of 0.91 and 0.72 and specificity of 0.78 and 0.74, for both types of sequence respectively. Altogether, these results show that the pipeline has the ability to *in silico* predict substandard primer designs and allow the researcher to re-do those regions, saving both costs and time.

**P-056: PHAGE-BACTERIA INTERACTIONS AFTER FECAL MICROBIOTA TRANSPLANTATION FOR UNDERSTANDING GUT-BRAIN AXIS IN AUTISM SPECTRUM DISORDER**

Paulína Belvončíková[1], Mária Džunková[2]

1. Institute of Molecular Biomedicine (IMBM), Faculty of Medicine, Comenius University, Sasinkova 4, 811 08 Bratislava, Slovakia
2. Institute for Integrative Systems Biology (I2SysBio), University of Valencia and CSIC, Calle Catedrático Agustín Escardino, 9, 46980 Paterna, Valencia, Spain
Presenter e-mail: paulina.belvoncik@gmail.com, paulina.belvoncikova@imbm.sk

The term autism spectrum disorder (ASD) refers to neurological and developmental disorders phenotypically manifested as an absence of social interaction and stereotypical behavior. Therapy targeting gut microbiota modulations, such as fecal microbiota transplantation (FMT), show to alleviate the asocial and stereotypical manifestations of behavior, suggesting that the gut microbiome induces the ASD behavioral changes via the “gut-brain axis”. The role of bacteriophages in the FMT treatment of ASD has not been investigated so far. The objective of our study is to analyze the transmission efficiency of FMT from the human donor to the animal model of SHANK3 mice, with an emphasis on bacteria-bacteriophage interactions. Representative stool samples of FMT human donors and mouse recipients were shotgun sequenced on the Illumina platform. Subsequently, the samples were assembled by the metaSPAdes pipeline, and binned using Metabat2. Viral contigs were mined using Virsorter and Virsorter2 tools. Afterward, the phage-bacteria links were obtained by comparing bacterial tRNA sequences and CRISPR spacers with the viral contigs. The obtained viral sequences found in recipient and donor stool samples will be used for the design of qPCR primers targeting bacteriophages, aiming to monitor abundances of bacteriophages and their hosts in the full longitudinal series of the mice stool samples. Knowledge of bacteriophages transport through the FMT could not only help us understand their FMT transport and the ability of viruses to reproduce in the FMT recipients' gut, but also could serve as a potentially interesting therapeutic tool for ASD in the future.

**P-057: COMPARISON OF ORAL, GUT AND TISSUE MICROBIOTA OF PATIENTS WITH HEAD AND NECK SQUAMOUS CELL CARCINOMA SHOWS THE MICROBIOME COMPOSITION HIGHER SIMILARITY BETWEEN SALIVA AND TISSUE SAMPLES**

XueQian Yin[1], Andrés Moya[1,2,3], Francesc-Xavier Avilés-Jurado[4], Susana Ruiz Ruiz[1,2], Vicente Pérez-Brocal [1,2]

1. Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO-Public Health), Avenida Catalunya 21, 46020 Valencia, Spain
2. Biomedical Research Networking Center for Epidemiology and Public Health (CIBEResp), Madrid, Spain
3. Institute for Integrative Systems Biology (I2SysBio), University of Valencia and Spanish National Research Council (CSIC), Carrer del Catedràtic Agustín Escardino Benlloch, 46980 Valencia, Spain
4. Hospital Clínic, C. de Villarroel, 08036 Barcelona, Spain
5. Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas (CIBERdem), C. de Melchor Fernández Almagro, 28029 Madrid, Spain
Presenter e-mail: xueyin@alumni.uv.es

Given the potential relationship between head and neck squamous cell carcinoma (HNSCC) and microbial dysbiosis, we used a metagenomic approach and next-generation sequencing to profile the microbiome from feces, saliva, and normal, peritumoral and tumoral tissue from patients diagnosed with HNSCC, before and after treatment with radiotherapy.

The purpose of this study was to characterize the taxonomic and functional composition of the microbiome from samples collected before treatment, to analyze differences in the microbiome from different types of samples, and to compare with samples collected from the same patient after treatment to find biomarkers of HNSCC so that we can ultimately evaluate the role of the microbiome as a predictor of radiation sensitivity.

We analyzed 142 saliva samples (containing 26 post-treatment samples), 139 feces samples (containing 25 post-treatment samples), and 239 tissue samples. We first described and compared the taxonomic and functional differences between the samples according to their origin, including alpha, beta-diversity and LEfSe analyses, focusing on those differentially present taxa and genes. Likewise, we analyzed differences in saliva and faeces between before and after treatment and its impact on bacterial composition, abundance, and diversity. This work provides a foundation for future studies aimed at understanding the role of the microbiome in HNSCC.

**P-058: MICROBIAL TRANSFORMATIONS OF MERCURY IN MARINE ENVIRONMENTS NEAR HYDROTHERMAL VENTS**

Ester-María López-García[1,2,3], Marisol Goñi-Urriza[1], Léa Cabrol[4], Lars-Eric Heimbürger-Boavida[4], Andrea G Bravo[3]

1. Université de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, Pau, France.
2. Universitat Autònoma de Barcelona, Facultat de Biociències, Departament de Genètica i Microbiologia, 08193 Bellaterra, Spain.
3. Institut de Ciències del Mar, ICM-CSIC, Departament de Biología Marina i Oceanografia, 08003, Barcelona, Catalunya, Spain.
4. Aix-Marseille Université, CNRS/INSU, University de Toulon, IRD, Mediterranean Institute of Oceanography (MIO), Bât Méditerrané, Campus de Luminy-Océanomed, 13009 Marseille, France
Presenter e-mail: estermlopezgarcia@gmail.com

Mercury (Hg) is one of the most dangerous, pervasive, and concerning pollutants. Sources of Hg can be anthropogenic and natural. Among the natural sources, hydrothermal vents are fissures on the sea floor that contribute to the release of Hg-containing geothermally heated fluids into seawater. In the ocean, several biological transformations of Hg determine the concentration of methylmercury (MeHg), a neurotoxin that bioaccumulates and biomagnifies in aquatic food webs and that might cause adverse health effects in severely exposed populations. The diversity and potential metabolic roles of microorganisms implicated in Hg transformations at hydrothermal vents are still largely unknown and even less when it comes to shallow hydrothermal vents (SHV), which are located above 200m depth. The current thesis project aims to determine the role of microorganisms in different Hg transformations and their contribution to MeHg production at diverse SHV sites. For this, we will quantify Hg species transformations and identify the microbial keyplayers and metabolic pathways involved in such processess, through metagenomic and metatranscriptomic analyses. The abundance and expression of genes involved in Hg methylation and demethylation (hgcAB gene cluster and mer genes operon, respectively) will be evaluated at 5 SHV locations: Panarea (Italy), Paleochori (Greece), Dom Joao de Castro Seamount (Portugal), Prony Bay (France) and Bahía Concepción (Mexico).

## P-059: DISTINCT TRANSCRIPTIONAL RESPONSES TO ACUTE AND CHRONIC OXIDATIVE STRESS IN SACCHAROMYCES CEREVISIAE

Alejandro Aguilar-Díaz[1], Beatriz Sabater-Muñoz[2], Christina Toft[1]

1. Institute for Integrative Systems Biology (I2SysBio, CSIC-UV), C/ Catedrático Agustín Escardino, 9, 46980 Paterna (Valencia), Spain
2. Department of Abiotic Stress, Institute of Molecular and Cellular Biology of Plants (IBMCP, CSIC), c/ Ingeniero Fausto Elio s/n. CPI 8E, Cubo Amarillo, 46022 Valencia, Spain
Presenter e-mail: Christina.Toft@csic.es

Oxidative stress is a common stress for many organisms as it can be generated as a by-product of aerobic respiration and different response strategies have therefore evolved. An organism responds to stress to maintain the redox homeostasis within the cell, hence, the cellular response to acute and chronic stress may not be the same. In this work, we set out to determine the transcriptional response that the yeast, Saccharomyces cerevisiae, has to this common stress of oxidation both during acute and chronic exposure. We determine the transcriptional response (RNAseq) and fitness effects of S. cerevisiae populations to oxidative stress (induced by hydrogen peroxide supplemented medium) compared to normal growth conditions (in YPD). The chronic response is obtained by evolving populations of S. cerevisiae under continuous exposure to oxidative stress for 66 generations (10 passages of 10% bottlenecks), whereas the acute response was determined under a short induction with hydrogen peroxide (24h, 6.6 generations). In previous studies, we have observed that anciently duplicated genes (small-scale and whole-genome duplicates) are particularly transcriptional plastic and are often at the centre of stress responses (osmotic and acidic stress, coupled with alternative non-fermentative carbon sources), showing a core of altered genes linked to this general oxidative stress response. We, therefore, have a special interest in duplicated genes. Overall, our analyses show that there is a clear difference in the transcriptional response to acute and chronic oxidative stress, with duplicated genes playing a bigger role in acute than in chronic response.

**P-060: KANAMYCIN EFECTS ON THE GUT MICROBIOTA FUNCTIONAL PROFILE OF BLATTELLA GERMANICA**

Jesús Marín-Miret[1,2], Rebeca Domínguez-Santos[1], Benjamí Pérez-Rocher[2], Ana Elena Pérez-Cobas[3], Andrés Moya[1,2], Amparo Latorre[1,2]

1. Institute for Integrative Systems Biology (I2SysBio), University of Valencia and Spanish Research Council, 46980 Paterna (Valencia), Spain.
2. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region, 46020 Valencia, Spain.
3. Department of Microbiology, Ramón y Cajal Institute for Health Research (IRYCIS).
Presenter e-mail: jesus.marin@uv.es

The bacterial composition of the gut microbiota can change due to different factors, such as diet, age, tissue, host genetics, and antibiotic treatment. These changes may foment a dysbiotic state of the microbiota that could affect the functional capabilities of the gut microbiota and, thus, affect the healthy status of the host. In this work, we used a metagenomic approach to analyze the gut microbiota of populations of Blattella germanica treated with kanamycin and the effect of treatment on the microbiota's functional profile. We analyzed amplicon sequencing data using QIIME2 and PICRUSt2 software to infer the metabolic pathways that the microbiota was performing. We also determined the robustness factors of the microbial populations over time. We found that kanamycin can disrupt the normal functional profile and the communication among the members of the microbiota and significantly increases the presence of defensive pathways such as ABC transporters. We also found that 13 taxa, mainly from the phyla Firmicutes and Proteobacteria, carried the kanamycin kinase enzyme, which is responsible for the degradation of aminoglycosides such as kanamycin. We demonstrated that kanamycin impacts the bacterial composition of the gut microbiota and its functional profile indirectly.

## P-061: INFERENCE OF METABOLIC NETWORKS AT MICROBIOTA LEVEL SHOWS DISCRETE STATES IN HEALTHY SUBJECTS

Benjamí Pérez-Rocher[1], Irene García[2], Pere Palmer-Roríguez[2], Andrés Moya[1], Mercè Llabrés[2]

1. I2SysBio, Campus Burjassot/Paterna Parc Cientific. C/ Catedrático José Beltrán, 2, 46980 Paterna Valencia
2. Universitat de les Illes Balears. Cra. de Valldemossa, km 7.5. 07122 Palma (Illes Balears)
Presenter e-mail:benjamin.perez@uv.es

Fecal samples were collected for 30 a priori healthy subjects every 30 days during a period of, approximately, 8 months. After elimination of samples of effectively non healthy subjects (illness discovered after sample collection) or perturbation (antibiotic intake), 157 samples from 29 individuals were considered. From metagenomic data, metabolic networks were obtained using KEGG data and computational reconstruction techniques, resulting 157 networks modelled as reaction graphs. Those networks were simplified to mDAGs (metabolic Directed Acyclic Graphs) using the methodology described by Alberich et al. (2017), which consist on a condensation of the reaction graph that preserves the relations between reactions, and yields a directed acyclic graph, i.e., a graph without directed cycles. A pairwise distance, as defined in (Alberich et al. (2017), was calculated among the 157 mDAGs. The comparison results, analysed through exploratory data analysis, shows discrete clusters not related with age groups, suggesting that metabolic networks oscillate over well defined states, at least in healthy people.

## P-062: RING1B IS PRESENT IN ACTIVE ENHANCERS ALONG WITH FUSION PROTEINS

Pablo Táboas[1,2], Enrique Blanco[3], María Sánchez-Jiménez[1,2], Carla Rios-Astorch[1,2], Elisabet Figuerola-Bou[1,2], Jaume Mora[1,2] and Sara Sánchez-Molina[1,2]

1. Developmental Tumor Biology Laboratory, Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat 08950 Barcelona, Spain
2. Pediatric Cancer Center Barcelona (PCCB), Hospital Sant Joan de Déu, Esplugues de Llobregat, 08950 Barcelona, Spain
3. Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, 08003 Barcelona, Spain
Presenter e-mail: pablo.taboas@sjd.es

Polycomb Repressive Complex 1 (PRC1) is a complex known to perform transcriptional repression along the genome in different cell types. RING1B is the core catalytic subunit of this complex, being its function the monoubiquitination of H2AK119 leading to transcriptional repression. Despite its repressive role in developmental regulated genes, recent publications have described its transcriptional activating function in some cellular contexts such as neurodevelopmental processes and cancer. Here, we demonstrate that RING1B is over-expressed in two fusion-driven sarcomas, Alveolar Rhabdomyosarcoma (ARMS) and Ewing Sarcoma (EwS). Alveolar Rhabdomyosarcoma emerges from a failure in muscle differentiation where a chromosomal translocation generates a fusion gene, PAX3-FOXO1 that maintains the cell in a proliferative and undifferentiated state. On the other hand, Ewing Sarcoma cell of origin is thought to be a mesenchymal stem cell and commonly appears in bone and soft tissues. It is characterized by the presence of the fusion gene EWSR1-FLI1 that acts as a transcription factor that leads to tumor formation. Both diseases predominantly affects adolescents and young adults. Using ChIP-seq technology, we demonstrate that RING1B co-localizes genome-wide with the fusion protein of each tumor both in cell lines (RH30 and A673) and in PDX mouse models (HSJD-ARMS-006). These findings establish that RING1B is a key modulator of chromatin remodeling induced by fusion proteins and its inhibition might open a new therapeutic strategy for these tumors.

## P-063: MATHEMATICAL MODELLING OF SIGNALLING PATHWAYS INVOLVED IN CHRONOLOGICAL AGING IN *SACCHAROMYCES CEREVISIAE*

Lorena Martínez-España[1,2], Agustín Aranda[3], Emilia Matallana[3], Irene Otero-Muras[2]

1. Master in Bioinformatics, University of Valencia, Valencia, Spain
2. Computational Synthetic Biology Lab, Institute for Integrative Systems Biology, I2SysBio, University of Valencia-CSIC, Paterna, Spain
3. Industrial Yeast Biotechnology Lab, Institute for Integrative Systems Biology, I2SysBio, University of Valencia-CSIC, Paterna, Spain
Presenter e-mail: lomares4@alumni.uv.es

The molecular mechanisms on which human aging is based have been extensively studied over the past decades, including metabolic and signalling pathways, and gene regulation. Furthermore, the mechanisms that induce or repress this phenomenon are also well known: cellular stress, nutritional restriction, radiation, etc. Saccharomyces cerevisiae, the budding yeast, has been widely used as a model organism to study, specifically, the signaling mechanisms involved in aging, being the most prominent ones the Ras/cAMP/PKA, TORC1/Sch9 and Snf1/AMPK pathways, due to the great homology between humans and yeast. In the present work, and based on the state of the art in the yeast signalling field, we develop a mathematical model based on ordinary differential equations (ODEs) implemented in Python and MatLab in order to predict cell behavior in its stationary phase under different nutritional regimens and radiation conditions. To calibrate the model, experimental data will be generated both in Valencia and in the Canfranc Underground Laboratory. The aim is contributing to elucidate the role of cosmic radiation in aging using an iterative method that combines in silico predictions with experiments to gain knowledge on the key molecular mechanisms involved (through the calibration, predictions and validation of the mathematical model). Then, we study the differential expression of genes at different times of aging using RNA-Seq experiments in order to see the evolution of gene expression in this phenomenon.

**P-064: THE GENE THAT WAS NOT THERE: LONG-READ ISO-SEQ TRANSCRIPTOMICS UNCOVERS HEMOCYANIN GENE STRUCTURE IN *CYPRIDEIS TOROSA***

Ximena-Aurora Altonar-Gómez[1], Markella Moraitou[1], Lidia López[1], Isa Schön[2,3], Raquel Ortells[1], Francesc Mesquita Joanes[1], Ferran Palero[1]

1. Institut Cavanilles de Biodiversitat i Biologia Evolutiva (Universitat de Valencia), Spain
2. Royal Belgian Institute of Natural Sciences, Brussels, Belgium
3. Centre of Environmental Sciences (University of Hasselt), Diepenbeek, Belgium
Presenter e-mail: Ximena.Altonar@uv.es

Cyprideis torosa (Jones, 1850) is a small ostracod crustacean found in brackish habitats (Meisch, 2000). When raised in low salinity conditions, its shell forms protruding features called nodes (Keyser & Aladin, 2004; van Harten, 2000). This physiological response has made C. torosa fossils a useful tool for reconstructing salinity in palaeoenvironmental analysis (Rosenfeld & Vesper, 1977). More recently, and beyond paleoecology, C. torosa has been presented as a model organism for all ostracods (de Deckker & Lord, 2017), thanks to its widespread Palearctic distribution and the broad range of salinity and oxygen concentrations it can tolerate. Yet, the genome of this species is little studied. The assembly of the first C. torosa genome (Tran Van et al., 2021) was a key step towards a better understanding of this species and ostracod genomes in general. However, due to methodological limitations, it was incomplete. Here, we produced long-fragment and high-quality RNA libraries of pooled individuals, which were sequenced using the PacBio Sequel II platform. RNA sequencing generated 2.2x106 HiFi reads (8.6 Gb data, with 3.8kb mean length) which resulted in a more accurate genome annotation and new information on the different isoform transcripts being expressed under extreme salinity conditions. In particular, the new data was used to identify the hemocyanin gene for the first time in Podocopida and analyze its gene structure.

**P-065: ROBUST CONTROL OF BIOCIRCUITS FOR SYNTHETIC BIOLOGY APPLICATIONS: THE TOGGLE-SWITCH STABILIZATION PROBLEM UNDER MOLECULAR NOISE**

Christian Fernández[1], Hamza Faquir[1], Manuel Pájaro[2], Irene Otero-Muras[1]

1. Computational Synthetic Biology Group, Institute for Integrative Systems Biology (UV-CSIC), University of València - Spanish National Research Council, 46980 Paterna, València.
2. Department of Mathematics and CITIC research center, University of A Coruña, Campus Elviña s/n, 15071 A Coruña
Presenter e-mail: christian.fernandez.perez@csic.es

One of the challenges towards real world implementation of synthetic biological circuits for cutting edge industrial, biomedical, or environmental applications is achieving robust control of the biocircuits behaviour for the correct functioning in presence of disturbances and molecular noise. Control of biocircuits is an active area of research, in which some significant advances have been recently achieved including the antithetic control by Briat et al., 2016, and Aoki et al., 2019 capable of perfect adaptation even in presence of stochastic noise, and the balancing of toggle switches at the level of single cell by Lugagne et al., 2017.

The inherent stochasticity of gene expression makes it very challenging to control gene regulatory networks. In this work, we exploit a recent very efficient stochastic modeling approach for gene regulation based on Integro-Differential Equations (PIDE) model developed by Pájaro et al., 2017. Combining this model with feedback control theory, we develop strategies to drive and keep the biocircuits under control at a desired stationary state. Here, we show how we achieve the closed-loop stabilization of an stochastic gene switch around the state of lowest probability. We successfully applied this integrative methodology to two different case studies of relevance in synthetic biology: a Toggle Switch based on Transcriptional Regulation, and a Toggle switch based on CRISPRi regulation (a hallmark of CRISPR interference mediated regulation systems developed in the group of Yolanda Schaerli by Santos-Moreno et al., 2020). We achieved the stabilization of the two switches around the desired reference state in silico and now we plan to implement our control strategy in vivo. Using the transcriptional regulation toggle switch, we will first calibrate our model with flow cytometry data and then implement the control with a computer-in-the-loop, using a microfluidics platform in combination with time-lapse microscopy.

Acknowledgments CF and IOM acknowledge funding from the GAIN Oportunius Grant Xunta de Galicia and MCIN/AEI/10.13039/501100011033 funding through grant PID2021-127888NA-I00 (COMPSYNBIO). HF and IOM acknowledge funding from e-MUSE MSCA-ITN-2020 European Training Network under Marie Sklodowska-Curie grant agreement No 956126. MP acknowledges support from grant FJC2019-041397-I funded by MCIN/AEI/10.13039/501100011033.

## P-066: SARS-COV-2 MUTATIONS ASSOCIATED TO BREAKTHROUGH INFECTIONS AND DIFFERENT CLINICAL OUTCOMES

Simón Rodríguez Santana[1*], Eugenia Ulzurrun[2,3*], Paula Ruiz-Rodriguez[4*], Miguel Álvarez-Herrera[4*], Roi Naveiro[1,8*], Daniel García Rasines[5*], Carmen Gil[2], Carlos Óscar Sorzano[3], David Ríos Insua[1], Nuria E. Campillo[1,2], Iñaki Comas[6,7], Fernando González-Candelas[4,7], Mireia Coscollá[4]

1. Institute of Mathematical Sciences, (ICMAT-CSIC)
2. Center for Biological Research Margarita Salas, (CIB-CSIC)
3. National Centre for Biotechnology, (CNB-CSIC)
4. Institute for Integrative Systems Biology, FISABIO Joint Unit, (I2SysBio CSIC-UV, FISABIO)
5. Department of Mathematics (Imperial College London)
6. Institute of Biomedicine of Valencia, (IBV-CSIC)
7. CIBER in Epidemiology and Public Health, (CIBERESP)
8. CUNEF Universidad

*. Contributed equally

Presenter e-mail: mireia.coscolla@uv.es

Determining the impact of SARS-CoV-2 mutations on immune evasion is essential for understanding the dynamics of the epidemic. Here we aim to identify SARS-CoV-2 mutations associated with breakthrough infections or hospitalization. We employed several sets of clinical data and SARS-CoV-2 genomes from Spain. We used over 25,000 SARS-CoV-2 sequences obtained from patients with information about the severity of the infection (hospitalization and ICU admission) and vaccination status. In addition, we studied the prevalence of certain viral mutations, pairs of mutations, and haplotypes based on the vaccination status of more than 5,000 patients using their respective SARS-CoV-2 sequences. We employed three different computational approaches to identify genetic determinants. Our first approach was based on genome-wide association studies. Second, we counted the frequencies of whole-genome haplotypes associated with a certain clinical outcome, testing the associations with a Chi-square test. Finally, we fit a hierarchical group lasso model that allows for the presence of individual effects and pairwise interactions between all the covariates employed, including not only the sequenced genomic data, but also patient's age, sex and sample date. This enables this last approach to control for all of these contributions, selecting functionally relevant gene products , including E, M, N, ORF8 and NSP12 (RdRP). Graphs built from the pairwise interactions selected by the model provided insight into the community structure for important mutations. We identified some of the most prevalent individual positions previously associated with immune escape, which serves as a good benchmark. Moreover, we found other relevant interactions and haplotypes that have not been reported in the literature yet and, in some cases, have stronger effects on the final outcome for the patient than individual mutations.

## P-067: CORRELATION BETWEEN THREE GENOME METRICS AND GENOME SIZE IN DIFFERENT PHYLA

Nuria De Frutos-Andicoechea[1], Pablo Román-Escrivá[1,2], Andrés Moya[1,2,3], Vicente Arnau[1], Wladimiro Díaz-Villanueva[1]

1. Institute for Integrative Systems Biology (I2SysBio). University of Valencia and Consejo Superior de Investigaciones Científicas (CSIC). 46980, Valencia, Spain.
2. Foundation for the Promotion of Sanitary and Biomedical Research of Valencian Community (FISABIO), 46020, Valencia, Spain.
3. Centro de Investigación Biomédica en Red de Epidemiología y Salud Púbilca (CIBEResp), Madrid, Spain.
Presenter e-mail: nudefru@alumni.uv.es

The tendency toward increasing complexity in biological evolution is a controversial issue in biology. Having a complexity measure can help with its resolution. We suggest appealing to genomes to measure complexity because they store information about the biotic and environmental interactions of species in their evolutionary history.

Here, we have determined the genome complexity of a large number of genomes by applying three complexity metrics: Biobit (BB), Genomic Signature (GS), and Sequence Compositional Complexity (SCC). Many species belonging to Archaea, Bacteria, and Eukarya were selected, and the genomes metrics were calculated and correlated with genome size.

The results were different according to metrics. BB appears to correlate positively with genome size at the phylum level. The other two metrics (GS and SCC) are more elusive. Although there seems to be a trend towards higher value with increasing genome sizes, at least among the three major domains (Archaea, Bacteria, and Eukarya), intra-phyla heterogeneity is very high for those metrics. It suggests that these metrics are greatly affected by the particular ecology and evolution of the species.

**P-068: 16S RRNA GENE DATABASE ENRICHMENT STRATEGY TO IMPROVE CLASSIFICATION OF NOVEL BACTERIAL TAXA IN NUDIBRANCH MICROBIOME**

Dafne Porcel Sanchis[1], Samuel Piquer-Esteban[1], Mária Džunková[1,2,3], Vicente Arnau[1], Wladimiro Díaz-Villanueva[1]

1. Institute for Integrative Systems Biology, University of Valencia and Consejo Superior de Investigaciones
Científicas (CSIC), Valencia, Spain
2. Department of Energy Joint Genome Institute, Berkeley, CA, USA
3. Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA,USA
Presenter e-mail: dafneposa@gmail.com

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 Tethybacteriales, 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 detected in nudibranch samples, we took as a baseline the SBDI Sativa curated 16S GTDB database (SBDI; 2021) and enriched it with Tethybacterales 16S rRNA gene sequences. We classified a dataset of 83 nudibranch microbiome samples using our enriched database and compare it with other two conventional databases: GTDB and SILVA SSU using the IDTAXA classifier of the DECIPHER
package. The enriched database managed to improve read classification capacity at the genus and family taxonomic levels, compared to the conventional databases. Our results highlight the importance of database enrichment strategies for analysing microbiomes containing a large portion of recently discovered novel bacterial taxa.

## P-069: EVOLUTION OF COMPLEXITY METRICS OF A GENOME UNDER THE EFFECT OF A SUBSTITUTION MODEL

Pablo Román-Escrivá[1,2], Vicente Arnau[1], Wladimiro Díaz-Villanueva[1], Andrés Moya[1,2,3]

1. Institute for Integrative Systems Biology (I2SysBio). University of Valencia and Consejo Superior de Investigaciones Científicas (CSIC). 46980, Valencia, Spain.
2. Foundation for the Promotion of Sanitary and Biomedical Research of Valencian Community (FISABIO), 46020, Valencia, Spain.
3. CIBER in Epidemiology and Public Health, 28029, Madrid, Spain.
Presenter e-mail: roespa@alumni.uv.es

In the search for the assessment of a genome's complexity and its comparison between organisms, several complexity metrics have been proposed by the literature in recent years. Genomic Signature1 (GS) is a k-mer-based metric, the value corresponding to the k-mer that maximizes the difference between observed and expected equifrequent classes of mers. This metric is based on the relative abundances of short oligonucleotides and chaos game representation applied to genomes. BioBit1 is also a k-mer-based metric based on the difference between the maximum entropy for a k-mer of a random genome of the same length as the genome under consideration and the entropy of that genome for such a k-mer. Finally, Sequence Compositional Complexity1 (SCC) is another metric that increases with the number of parts (i.e., compositional domains) and the length and compositional differences found in a genome sequence by a segmentation algorithm, paralleling the concept of 'pure complexity' of McShea and Brandon. In this study, we measured those metrics to an initial genome (E. coli) that evolved by a mutation substitution model with no selection using ALF2, a simulation framework for genome evolution. We have observed that the three metrics decrease with evolutionary time. We conclude that genome complexity also falls in the absence of natural selection.

Moya, A., Oliver, J.L., Verdú, M. et al. Driven progressive evolution of genome sequence complexity in Cyanobacteria. Sci Rep 10, 19073 (2020). https://doi.org/10.1038/s41598-020-76014-4 2.

Daniel A. Dalquen, Maria Anisimova, Gaston H. Gonnet, Christophe Dessimoz, ALF—A Simulation Framework for Genome Evolution, Molecular Biology and Evolution, Volume 29, Issue 4, April 2012, Pages 1115–1123, https://doi.org/10.1093/molbev/msr268

**P070: A SEQUENCE-DEPENDENT CARTESIAN COARSE-GRAINED MODEL OF DNA REPRODUCES WITH HIGH ACCURACY ATOMISTIC SIMULATIONS**

David Farré[1], Juan Pablo Arcón[1], Charlie A. Laughton[2] and Modesto Orozco[1,3]

1. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, E-08028 Barcelona, Spain.
2. Department of Pharmacy. Nottingham University
3. Department of Biochemistry and Biomedicine, University of Barcelona, E-08028 Barcelona, Spain.
Correspondence: modesto.orozco@irbbarcelona.org

We present a new coarse-grained model for molecular dynamics simulations of very long segments of B-DNA. The developed approach uses a one bead per nucleotide representation and a sequence-dependent Hamiltonian containing bonded and non-bonded contributions. The model is calibrated against a dataset of representative all-atom molecular dynamics simulations of DNA, considering both local and global descriptors. The simulation method is computationally very efficient allowing the dynamical representation of chromatin-scale segments of DNA at the nucleotide level with very high quality. Interestingly, the coupling of the coarse-grained simulations with a machine learning model trained with independent all-atom molecular dynamics data, allows to recover atomistic trajectories of a quality indistinguishable from those using all-atom representation in solution but with a computational efficiency enabling simulations of kilobase pair-long DNA segments.

## P071: PHAGES IN *VIBRIO VULNIFICUS*: THE UNTOLD STORY

Héctor Carmona-Salido[1], Carolin Charlotte Wendling[2], Carmen Amaro[1]

1. University Institute of Biotechnology and Biomedicine (Biotecmed), Universitat de València, Valencia, Spain.
2. Department of Environmental Systems Science, Institute of Integrative Biology, ETH Zürich, Zurich, Switzerland.
Presenter e-mail: hector.salido@uv.es

Vibrio genus is made up from more than one hundred different species that inhabit aquatic environments. Some of them are able to cause several diseases to both animals and humans. The incidence of these illnesses is arising due to climate change, which allows bacteria to colonise traditionally colder environments. This is especially true on *Vibrio vulnificus*, a zoonotic bacterium that causes vibriosis, with the highest fatality rate of any foodborne pathogen. There are several horizontal gene transfer mechanisms in bacteria that can turn non-virulent strains into harmful and dangerous strains that can cause disease. Recent studies have shown that fish farms act fuelling the genetic exchange among *V. vulnificus* and many other bacteria, producing new pathogenic variants on the species that are harmful in public health. One of those mechanisms that allow the genetic flow between bacteria is transduction. When phages assemble their DNA to produce new virions particles, it can sometimes carry bacterial genes along with them that are passed to other bacteria. This includes virulence and antimicrobial resistance (AMR) encoding genes. In this study, we performed a genomic analysis on 310 *V. vulnificus* strains coming from different sources, including environmental and pathogenic strains. By using software based on machine learning algorithms, we were able to identify and annotate prophage regions within the genomes. Interestingly, those new prophage regions harboured virulence and AMR genes. With this first step, we are close to understand how new pathogenic variants emerge *in V. vulnificus* and create effective strategies to control them.

**P072: AndorrAigua – AN ATLAS OF PYRENEAN FRESHWATER MICROBES THROUGH METAGENOMICS.**

Hannah Benisty[1], Manel Niell[2], Roderic Guigó[1,3]

1. Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona (BIST), Dr. Aiguader 88, 08003 Barcelona, Catalonia, Spain.
2. Andorra Recerca + Innovació, Andorra.
3. Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain.
Presenter e-mail: hannah.benisty@crg.eu

The monitoring of mountain lakes contributes to studying the impact of climate change, given their great vulnerability to climatic fluctuations and growing anthropic pressure. These factors alter the biomass and microbial activity that are crucial for the functioning of ecosystems. Therefore, mountain lakes are perfect sentinels for the study of microbial biodiversity in the context of climate change. The project "AndorrAigua" has as its main objective to capture the first image of the microbial biodiversity of emblematic lakes of the Principality of Andorra using a simple, fast and cost-effective workflow. The project will focus on the collection of samples and subsequent 16S rRNA gene sequencing with the portable nanopore technology. This procedure will allow us to identify the distinct taxonomic groups, as well as their relative concentrations in each of the altitude lakes. With Andorra as a small-scale living laboratory, we could contribute to biodiversity genomics and establish a valuable starting point for future efforts to study the adaptations and perils of mountainous freshwater microbiota.

## P073: GENETIC INTERACTION IN THE ASSOCIATION BETWEEN OXIDATIVE STRESS AND DIABETES IN SPANISH POPULATION

Rebeca Melero-Valverde[1]

1. Instituto de investigación Sanitaria (INCLIVA), Valencia, Spain.
Presenter e-mail: rebeca_melero_4@hotmail.com

Oxidative stress plays a pivotal role in the pathogenesis and progression of diabetes. Our aim is the study of the genetic interaction of genes related with type 2 diabetes mellitus (dyslipidemia, obesity, redox homeostasis, renin-angiotensin-aldosterone system and metal transporters) and oxidative stress in general population of Spain (Hortega Study). For this purpose, 1502 adults from area of the University Hospital Rio Hortega, were studied. It was analyzed 900 single nucleotide polymorphisms from 272 candidate genes. The oxidative stress biomarkers used were oxidized to reducedglutathione ratio, malondialdehyde and 8-oxo-7,8-dihydroguanine. Have been found significant associations between the oxidative stress levels and risk of developing type 2 diabetes mellitus in two polymorphisms: rs196904 (ERN1 gene) and rs2410718 (COX7C gene); and haplotypes of the following genes: SP2, HFF1A, ILI8R1, EIF2AK2, TXNRD2, PPARA, NDUFS2 and ERN1. COX7C and TXNRD2 encode enzymes that have a direct role in the elimination of reactive oxigen species; SP2, HFF1A, HFF1A, EIF2AK2, PPARA, ERN1, NDUFS2 are involved in pathways related to apoptosis caused by high levels of oxidative stress. These results could be of great interest to understand the connection between oxidative stress and the development of diabetes.

**P074: GENETIC DIVERSITY AND PHYLOGENY OF SPECIES OF THE LICHEN-FORMING FUNGAL GENUS *SQUAMARINA* POELT AND ITS PHYCOBIONTS IN THE MEDITERRANEAN AND MACARONESIAN REGIONS**

Tamara Pazos[1]; Garrido-Benavent, I[1]; Moya, P[1]; Chiva, S[1]; Barreno, E[1]

1. Institut Cavanilles de Biodiversitat i Biologia Evolutiva (ICBiBE) – Botánica & Departament de Botànica i Geologia (Fac. CC. Biológicas), Universitat de València, C/ Dr. Moliner 50, E-46100 Burjassot, València.
Presenter e-mail: tamara.pazosm@gmail.com

The genus *Squamarina* was introduced by Poelt in 1958 and includes species of lichenized ascomycetes distributed mainly within the northern hemisphere, which develop on calcareous and siliceous rocks or clay and gypsum soils. Most have squamulose, sometimes lobed, saxicolous or terricolous thalli, often with abundant and conspicuous apothecia. Although *Squamarina* species are common in Mediterranean ecosystems, comprehensive studies on the genetic and phylogenetic diversity of mycobionts and phycobionts involved in these symbioses have not yet been carried out in Europe. In the present study, thalli were collected from different localities in the northern and eastern Iberian Peninsula, Balearic Islands and several Macaronesian Archipelagos (Canary Islands, Madeira and Cape Verde). The identity and genetic diversity of mycobionts and phycobionts were studied by phylogenetic approaches using nrITS molecular data, and the identity of the microalgae was also corroborated with the observation of their ultrastructures by transmission electron microscopy. Additionally, interaction patterns between the different lineages of myco- and phycobionts were analyzed, and the relative importance of the genetic identity of each symbiont, macroclimate, geography and substrate in determining the variation in myco- and photobiont diversity was quantified by a variance partitioning study. The results revealed an unexpectedly high genotypic diversity in mycobionts of the genus *Squamarina* that associate with seven species of chlorophyte microalgae of the genera *Asterochloris* Puymaly, *Vulcanochloris* Tschermak-Woess and *Trebouxia* Vancurová, Peksa, Nemcová & Ŝkaloud. In symbiotic interactions, specificity patterns varied among the different species revealing, in most cases, a remarkable flexibility in phycobiont selection. However, the genetic identity of the myco- and phycobiont proved to be a key factor in the diversity and distribution of the *Squamarina* species, which would have allowed the colonization of bioclimatically diverse environments through the establishment of symbiosis with phycobiont species that might be more adapted to local conditions.

PROJECTS: PROMETEO/2021/005 (GVA to EB) and PID2021-127087NB-I00 (Ministerio de Ciencia e Innovación to IGB).

## P075: INFLUENCE OF MENTAL HEALTH MEDICATION ON MICROBIOTA IN THE ELDERLY POPULATION IN THE VALENCIAN REGION

Ana Barberá[1], Nicole Pesantes[1,2], Benjamí Pérez-Rocher[3], Alejandro Artacho[1], Andrés Moya[1,2,3] and Susana Ruiz-Ruiz[1,2]

1. Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO), València, Spain
2. Consorcio de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBEResp), Madrid, Spain
3. Instituto de Biología Integrativa y de Sistemas (I2Sysbio), Universitat de València y CSIC, València, Spain
Presenter e-mail: ana.barbera@fisabio.es

The improvement of economic, hygienic-sanitary conditions and health promotion has led to a progressive aging of the Spanish population, which comes with several health issues, including mental health disorders. The gut-brain axis is a bidirectional network linking the central nervous system with gastrointestinal tract. The presence of certain bacterial taxa and their associated metabolic functions could be influencing the normal functioning of the brain and triggering physiological disorders. We took a case-control approach to study the interplay between gut microbiota and mental health of elderly people. Fecal and saliva samples from 101 healthy volunteers over 65 were collected, 28 of which (EE|MH group) reported being treated for anxiety, depression, and insomnia. 16S rRNA gene sequencing and metagenomic sequencing were applied to determine the differences between intestinal and oral microbiota. Significant differences in genera were found, specifically eight in the gut microbiota, and five in the oral microbiota. Additional functional analysis of fecal samples showed differences in five orthologous genes related to tryptophan metabolism, the precursor of serotonin and melatonin, and in six categories related to serine metabolism, a precursor of tryptophan. Moreover, pathways regulating longevity, the dopaminergic synapse, the serotoninergic synapse, and two amino acids had significant inter-group differences

## P076: THE EPIGENETIC LOGIC OF GENE ACTIVATION

Beatrice Borsari[1], Amaya Abad[1], Cecilia C. Klein[1], Ramil Nurtdinov[1], Vasilis F. Ntasis[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]

1. Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona (BIST), Catalonia, Spain.
Presenter e-mail: silvia.perez@crg.cat

Histone modifications are widely accepted to play a causal role in the regulation of gene expression. This role has been recently challenged, however, by reports showing that gene expression may occur in the absence of histone modifications. To address this controversy, we have generated densely-spaced transcriptomic and epigenomic maps in a time-course cell homogeneous system that occurs with massive transcriptional changes. We found that the relationship between histone modifications and gene expression is weaker than previously reported, and that it can even run contrary to established assumptions, such as in the case of H3K9me3. Our data suggest a model that reconciles the seemingly contradictory observations in the field. According to this model, histone modifications are associated with expression only at the time of initial gene activation, when they are deposited in a dominant order at promoter regions, generally preceding deposition at enhancers. Further changes in gene expression, even larger than those occurring at gene activation, are essentially uncoupled from changes in histone modifications. Genes are in a very limited number of major chromatin states which mostly remain stable over time. Data available during mouse development largely recapitulates this model. Our work provides a first sketch of the epigenetic logic underlying gene activation in eukaryotic cells.

**P077: GOCOMPARE: AN R PACKAGE TO COMPARE FUNCTIONAL ENRICHMENT ANALYSIS BETWEEN TWO SPECIES.**

Chrystian C. Sosa[1,2], Diana Carolina Clavijo-Buriticá[1], Víctor Hugo García-Merchán[2], Nicolas López-Rozo[1], Camila Riccio-Rengifo[1], Maria Victoria Diaz[3], David Arango Londoño[1], Mauricio Alberto Quimbaya[1]

1. Department of Natural Sciences and Mathematics, Pontificia Universidad Javeriana, Cali, Cali, Colombia
2. Evolution, Ecology and Conservation Research Group EECO, Biology Program, Faculty of Basic Sciences and Technologies, Universidad del Quindío, Armenia, Colombia
3. International Center for Tropical Agriculture (CIAT), Cali, Colombia
Presenter e-mail: ccsosaa@javerianacali.edu.co

Functional enrichment analysis is a cornerstone in bioinformatics as it makes possible to identify functional information by using a gene list as source. Different tools are available to compare gene ontology (GO) terms, based on a directed acyclic graph structure or content-based algorithms which are time-consuming and require *a priori* information of GO terms. Nevertheless, quantitative procedures to compare GO terms among gene lists and species are not available. Here we present a computational procedure, implemented in R, to infer functional information derived from comparative strategies. GOCompare provides a framework for functional comparative genomics starting from comparable lists from GO terms. The program uses functional enrichment analysis (FEA) results and implement graph theory to identify statistically relevant GO terms for both, GO categories and analyzed species. Thus, GOCompare allows finding new functional information complementing current FEA approaches and extending their use to a comparative perspective. To test our approach GO terms were obtained for a list of aluminum tolerance-associated genes in *Oryza sativa* subsp. *japonica* and their orthologues in *Arabidopsis thaliana*. GOCompare was able to detect functional similarities for reactive oxygen species and ion binding capabilities which are common in plants as molecular mechanisms to tolerate aluminum toxicity. Consequently, the R package exhibited a good performance when implemented in complex datasets, allowing to establish hypothesis that might explain a biological process from a functional perspective, and narrowing down the possible landscapes to design wet lab experiments.

## P078: BUILDING THE FLOMICS BIOINFORMATICS INFRASTRUCTURE FROM THE GROUND UP

Julien Lagarde[1], André Lopes[1], Pedro Ferreira[1], Cristina Tuñí[1], Lluc Cabus[1], Joao Curado[1]

1. Flomics Biotech SL, Carrer del Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain
Presenter e-mail: cristina.tuni@flomics.com

Access to robust and scalable computing environments is nowadays a must for even the smallest bioinformatics research lab or startup company. Often, suitable on-premise infrastructures are unavailable or financially out of reach for such entities. For most, using cloud computing platforms such as Amazon Web Services (AWS) are the solution of choice to this problem. Integrating Nextflow pipelines in such environments provides obvious benefits in terms of scalability and computational reproducibility, but is not trivial to implement for the uninitiated. At Flomics (a biotech startup company operating in the liquid biopsies space and with a strong computational component), we have built our entire bioinformatics environment around Nextflow and nf-core pipelines in AWS. We will share our experience setting up our cost effective computing infrastructure, how we customized it to our needs using e.g. serverless functions, and share some of the lessons we learned in the process.

## P079: EXAMINING HUMAN FERTILITY AS A FITNESS COMPONENT IN THE CONTEXT OF THE ANTAGONISTIC PLEIOTROPY THEORY OF AGEING

Brigos E.[1], Vasallo C.[2], Morcillo-Suárez C.[1], Muntané G.[1,3,4], Navarro A.[1,5,6]

1. Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Barcelona, Catalonia 08003, Spain
2. European Genome-phenome Archive (EGA) in the Centre for Genomic Regulation (CRG), the Barcelona Institute of Science and Technology Dr. Aiguader 88, Barcelona, 08003 Spain.
3. Research Department, Hospital Universitari Institut Pere Mata, Institut d'Investigació Sanitària Pere Virgili (IISPV), Reus, Spain
4. Centro de Investigación Biomédica en Red en Salud Mental (CIBERSAM), Reus, Spain
5. Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain
6. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
Presenter email: eva.brigos@upf.edu

Fertility and ageing are key phenotypes to understand fitness trade-offs that have shaped human biology. As Antagonistic Pleiotropy predicts, genetic variants that are risk factors for disease can be tolerated or even selected if they also determine an advantage in early-life fertility. Contemporary human populations are characterized by reduced fertility rates and increased prevalence of complex diseases. However, it is still unclear how recent demographic transitions are shaping patterns of natural selection. In the present study we use genomic data to evaluate the shared pleiotropic landscape between fertility and complex diseases in an evolutionary perspective. We show that positive genetic correlations are more common in early-onset diseases. Also, positive and negative pleiotropies have different accumulation curves with positive pleiotropies increasing at earlier disease onset. When evaluating signals of positive selection results suggest differential selective pressures acting over positive and negative pleiotropies. Overall, our findings suggest that the role of fertility in shaping modern disease patterns is consistent with the Antagonistic Pleiotropy theory of human ageing.

## P080: MASSIVELY PARALLEL MUTAGENESIS TO CHARACTERIZE VARIABILITY IN NONSENSE-MEDIATED RNA DECAY (NMD) ACROSS MUTATIONS

Ignasi Toledano[1,2], Guillermo Palou[1], Ben Lehner[2,3,4,5] and Fran Supek[1,3]

1. Genome Data Science, Institute for Research in Biomedicine (IRB), Barcelona, Spain.
2. Center for Genomic Regulation (CRG), Barcelona, Spain.
3. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.
4. Universitat Pompeu Fabra (UPF), Barcelona, Spain.
5. Wellcome Sanger Institute, London, UK.
Presenter e-mail:ignasi.toledano@irbbarcelona.org

Ten percent of heritable diseases are caused by nonsense mutations. Nonsense mutations are single nucleotide alterations that generate premature termination codons (PTCs). These mutated genes are considered to behave as null alleles, since they produce truncated proteins and the transcript is targeted by nonsense-mediated RNA decay (NMD). This pathway recognizes and degrades mRNA species with a PTC, in order to protect the cell from toxic truncated proteins. Although the biological relevance of NMD is evident, there is still a gap in understanding the genetic determinants that govern NMD and the underlying molecular mechanisms. The efficiency of NMD differs strongly across nonsense mutations, with many mutations being fully resistant to NMD and sometimes able to translate partially functional truncated proteins, questioning the classical assumption that all nonsense mutations behave as null alleles. This variability is partially explained by previously described NMD mechanistic models, but still there are many cases in which these models fail to predict the efficiency of NMD. Our objective in this project is to better understand how the sequence landscape modulates the effect of the PTCs on transcript expression. We performed systematic deep mutational scanning (DMS) experiments in human cell lines to validate, refine and learn new NMD rules, to improve our ability to predict the extent to which different stop codons in the human genome do or do not trigger transcript degradation. For instance, we have seen that the position of the PTC within the exon correlates with NMD efficiency, where those PTCs placed in the 5' end display higher NMD resistance than those near the 3'end. In addition, we have also been able to elucidate an accurate quantitative map of the 5'end-mediated NMD protection along the 5'end of a transcript, and break down such effects into two causal mechanisms: a methionine-dependent (driven by translation reinitiation) and a methionine independent mechanism of 5' NMD evasion (with less clear mechanistic basis). Also, by randomizing the nucleotides surrounding the PTC, we have been able to capture the effect of sequence context on NMD. To ascertain relevance of our experimental results to genetic variation observed in human cell lines, we are currently assessing how the observations obtained in these experiments explain NMD variability in RNA-seq data from the TCGA tumors and the GTex healthy human tissues, as well as trying to derive improved NMD predictive models. Together, these results expand our knowledge of ways in which nonsense mutations regulate transcript expression in human cells with all its potential applications downstream.

**P081: MUTATING THE DOMAINOME**

Toni Beltran[1] and Ben Lehner[1,2,3,4]

1. Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology Barcelona, Spain.
2. University Pompeu Fabra (UPF), Barcelona, Spain.
3. Institució Catalana de Recerca i estudis Avançats (ICREA), Barcelona, Spain.
4. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
Presenter e-mail: toni.beltran@crg.eu

The vast majority of variants on human protein-coding genes – including the most studied disease genes – are variants of uncertain significance, meaning that their phenotypic effects and their contribution to human disease are unknown. In order to carry out their functions, proteins need to fold into stable spatial conformations: their native structures. Accordingly, reduced protein fold stability or solubility is the most common mechanism by which missense mutations can lead to disease.

Recently, deep learning methods have leveraged the large amounts of structural data for a wide range of proteins generated throughout decades of work in the structural biology community to accurately predict protein structure from sequence alone. This type of approach could be readily applied to many other traits of proteins; however, large datasets covering a wide range of proteins and biophysical traits other than structure are scarce.

We are to filling this gap by generating a large dataset of protein stability of all single amino acid substitutions in a diverse set of ~1000 human protein domains enriched in disease-associated genes. To do this, we are using a yeast in vivo protein abundance assay in which the growth of yeast is coupled to the fold stability and solubility of a target protein. We will present our progress so far on the selection of target domains for mutagenesis, and our ongoing efforts that currently stand at ~300,000 amino acid substitution measurements in ~450 domains covering >100 unique protein families.

We believe that our datasets will be useful to better understand, predict and engineer the fine balance of forces that govern the folding of proteins, implications for clinical genetics and biotechnological applications.

## P082: CHARACTERISATION OF ALTERNATIVE POLYADENYLATION AT SINGLE CELL RESOLUTION IN ALZHEIMER DISEASE

Franz Ake[1,2], Ana Gutierrez-Franco[1,2], Sandra Maria Fernandez-Moya[1,2], Mireya Plass-Portulas[1,2,3]

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 (AD). The rapid development of 3'tag-based single-cell RNA sequencing (scRNAseq) has enabled the study of gene expression at the individual cell level and the implementation of methods for describing APA sites at single cell resolution. Here we present SCALPEL, a tool for characterising APA sites at single cell resolution using 10X Genomics or Dropseq scRNA-seq dataset. SCALPEL allows quantifying RNA expression at isoform level and single-cell resolution and identifying changes in isoform usage across cell populations and conditions. We used SCALPEL to study the changes in APA during the differentiation of 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 project to use SCALPEL to investigate the role of APA in neural differentiation and its role in the development of AD and how APA changes during neural differentiation and how these changes are altered in AD.