IX Jornada de Bioinformàtica i Genòmica

Organitzada per:
Secció de Bioinformàtica i Genòmica de la SCB
Associació Bioinformatics Barcelona – BIB
Institut de Recerca contra la Leucèmia Josep Carreras

Patrocinada per:

Virtual platform

Slack: https://tinyurl.com/JBG2021

16 de desembre de 2021

Comitè organitzador:

Tanya Vavouri (IJC)
Lorenzo Pasquali (UPF)
Mario Cáceres (ICREA, UAB)
Roderic Guigó (CRG-UPF)

Suport:

MariàngelsGallego (SCB)
Maite Sánchez (SCB)
Romina Garrido (CRG)

PROGRAM

9:15 – 9:30 Welcome and opening of the symposium

SESSION I.

Chair: Tanya Vavouri (IJC)

9:30 - 10:15 **Invited Lecture: Mark Robinson** (University of Zurich, Switzerland). Flexible differential analyses of single cell data.
10:15 - 10:30 **Aida Ripoll Cladellas** (BSC) Single-cell transcriptomics reveals insights into telomere shortening with aging.
10:30 - 10:45 **Ramil Nurtdinov** (CRG) Genome-wide Alu-mediated weak CEBPA binding is associated with slower B cell transdifferentiation in human.

10:45 - 11:15 Break

SESSION II.

Chair: Eduard Porta (IJC)

11:15 - 11:30 **Xavier Farré Ramon** (IDIBELL) The shared genetic architecture of cancer types and blood cell traits.
11:30 - 11:45 **Uciel Chorostecki** (BSC) Structural characterization of NORAD, a human lncRNA dysregulated in cancer.
11:45 - 12:00 **Paula Torren Peraire** (IRB) Using bioactivity signatures to predict drug-target interactions.
12:00 - 12:45 **Flash talks** (3 mins, 1 slide)
Mireya Plass (IDIBELL) Single-cell transcriptomics of iPSC-derived neurons reveals functional changes in Alzheimer's Disease.
Núria Olvera Ocaña (IDIBAPS) Lung Tissue Multi-layer Network in Chronic Obstructive Pulmonary Disease.
Miriam Magallón Lorenz (IGTP) Genomic characterization of eight established MPNST cell lines: a resource for precision medicine.
Nerea Moreno (UPF) Assessing the digenic model in primary immunodeficiencies using population whole-genome sequencing data.
Arnau Comajuncosa Creus (IRB) Novel binding site descriptors built upon inverse virtual screening.
Altair Hernandez (UPF) Integrative Modelling to explore functionality of cellular complexes.
Marta Sanvicente García (UPF) CRISPR Analytics: a versatile and precise genome editing simulation and analysis tool.
Luca Cozzuto (CRG) Analysis of nanopore data using Master of Pores 2.
Ferriol Calvet Riera (CRG) Efficient and accurate protein-coding gene prediction.
Q&A to flash talk presenters

12:45 - 14:00 Break

SESSION III.

Chair: Elisabetta Mereu (IJC)

14:00 - 14:15 **Sponsor Talk: Evan Floden (Sequera)** Building the foundations of a tech-enabled biology economy.

14:15 - 14:30 **Beatrice Borsari** (CRG) Multi-tissue integrative analysis of personal epigenomes.

14:30 - 14:45 **Natalia Blay** (IGTP) Genome Structural Variants analysis in chronic diseases through SV imputation using the GCAT Panel, the first haplotype-based reference panel from Iberian Population.

14:45 - 15:15 **Flash talks** (3 mins, 1 slide)

Miquel Àngel Schikora Tamarit (BSC) perSVade: personalized Structural Variation detection in your species of interest.

Aina Colomeri Vilaplana (UAB) PopHumanVar: an interactive application for the functional characterization and prioritization of adaptive genomic variants in humans

Jose Francisco Sanchez Herrero (IGTP) BacterialTyper: a general purpose suite for comprehensive analysis of bacterial whole genome sequencing data for clinical and epidemiological applications.

Oriol Bárcenas (UAB) SolupHred: a server to predict the pH-dependent aggregation of intrinsically disordered proteins.

Javier Garcia Pardo (UAB) A3D Database: Structure-based Protein Aggregation Predictions for the Human Proteome.

Q&A to flash talk presenters

15:15 - 15:45 Break

SESSION IV.

Chair: Lorenzo Pasquali (UPF)

15:45 - 16:30 **Invited Lecture: Meritxell Oliva** (University of Chicago, USA). TBD.

16:30 - 16:45 **Ignasi Moran Castany** (BSC) The gene expression regulatory variation landscape of human pancreatic islets.

16:45 - 17:00 **Alejandro Valenzuela** (IBE CSIC-UPF) Leveraging comparative genomics across primates to decipher the genomic architecture of complex traits.

17:00 - 17:30 Break

17:30 - 18:00 *Genes* award for the best oral and poster presentation and closing of the symposium.

Single-cell transcriptomics reveals insights into telomere shortening with aging

Aida Ripoll-Cladellas[1]*, Sergio Andreu-Sánchez[2,3]*, Geraldine Aubert[4,5]*, Sandra Henkelman[6]*, Daria V. Zhernakova[2,7#], Trishla Sinha[2], Alexander Kurilshikov[2], Maria Carmen Cenit[2,8], Marc Jan Bonder[2,9,10], Lifelines cohort study, Lude Franke[2], Cisca Wijmenga[2], Jingyuan Fu[2,3], Peter Lansdorp[3,6,11], Alexandra Zhernakova[2], Monique G.P. van der Wijst[2#], Marta Melé[1$#].

[1]Life Sciences Department, Barcelona Supercomputing Center, 08034, Barcelona, Catalonia, Spain
[2]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[3]Department of Pediatrics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[4]Terry Fox Laboratory, British Columbia Cancer Research Center, Vancouver, BC, Canada
[5]Repeat Diagnostics Inc., Vancouver, BC, Canada
[6]European Research Institute for the Biology of Ageing, University of Groningen, Groningen, the Netherlands
[7]Laboratory of Genomic Diversity, Center for Computer Technologies, ITMO University, St. Petersburg, 97101, Russia
[8]Microbial Ecology, Nutrition, and Health Research Unit. Institute of Agrochemistry and Food Technology (IATA-CSIC), 46980 Paterna-Valencia, Spain
[9]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[10]European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany
[11]Departments of Hematology and Medical Genetics, University of British Columbia, Vancouver, BC, Canada
* Shared first author
# Second author
$ Last author
# Co-corresponding authors

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

Although aging is a universal process affecting all tissues, the underlying cellular and molecular transcriptional mechanisms remain largely unknown. Telomere length has been proposed as one of the main hallmarks of aging. The average length of telomere repeats declines with age in cells of most self-renewing tissues. However, the interconnectedness between telomere length and its contribution to aging at the cell-type resolution level is not yet well understood. Here, we studied the relationship between telomere length changes and gene expression at single-cell resolution in 62 donors from the Netherlands Lifelines Deep cohort. We coupled clinically validated flow-FISH measurements of telomere length in six blood cell types to the expression level of corresponding cell types using single-cell RNA-sequencing data. This revealed 97 genes whose expression level varied with changes in telomere length in T cells. Three of them *(DNAJA1, EEF1A1, RPL29)* were previously reported as telomere binding proteins, indicating that our approach captures genes directly involved in telomere length dynamics. Moreover, the genes negatively associated with telomere length were enriched for pathways related to translation and nonsense-mediated decay, which might have important physiological consequences. Even though some of the telomere length-associated genes were located near the telomere ends, in general, our differential expression findings could not be explained by previously described mechanisms (telomere position effect (TPE) or TPE over long distances (TPE-OLD)). It suggests that T cells are sensitive to telomere length-induced expression changes that can act through both short- and longer-range interactions, but also that cell expression may directly influence telomere dynamics. Altogether, this study reveals the importance of further exploring the context-specificity of telomere shortening-induced changes that occur with aging or specific treatments.

Genome-wide Alu-mediated weak CEBPA binding is associated with slower B cell transdifferentiation in human

Ramil Nurtdinov[1], María Sanz[1], Amaya Abad[1], Alexandre Esteban[1], Sebastian Ullrich[1], Carme Arnan[1], Rory Johnson[1], Sílvia Pérez-Lluch[1], Roderic Guigó[1,2]

1. Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology; Barcelona, Catalonia, Spain
2. Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain

Presenter e-mail: ramil.nurtdinov@crg.eu

Many developmental and differentiation processes take substantially longer in human than in mouse. To investigate the molecular mechanisms underlying this phenomenon, here we have specifically focused on the transdifferentiation from B cells to macrophages. The process is triggered by exactly the same molecular mechanism, the induction by the transcription factor CEBPA, but takes three days in mouse and seven in human (Bussmann et al., 2009, Rapino et al., 2013). In mouse, the speed of this process is known to be associated with *Myc* expression (Francesconi et al., 2019). We found that in this species, CEBPA binds strongly to the *Myc* promoter, efficiently down-regulating *Myc*. In human, in contrast, CEBPA does not bind this promoter, and *MYC* is indirectly and more slowly down-regulated. Attenuation of CEBPA binding is not specific to the *MYC* promoter, but a general trait of the human genome across multiple biological conditions. We traced back weak CEBPA binding to the primate-specific Alu repeat expansion. Many Alu repeats carry strong CEBPA binding motifs, which sequester CEBPA, and attenuate CEBPA binding genome-wide. We observed similar CEBPA and *MYC* dynamics in natural processes regulated by CEBPA, suggesting that CEBPA attenuation could underlie the longer duration in human processes controlled by this factor. Our work highlights the highly complex mode in which biological information is encoded in genome sequences, evolutionarily connecting, in an unexpected way, lineage-specific transposable element expansions to species-specific changes in developmental tempos.

The shared genetic architecture of cancer types and blood cell traits

Xavier Farré[1], Miguel Angel Pardo[2], Roderic Espín[2], Rafael de Cid[1] and Miquel Angel Pujana[2]

1. Genomes for Life-GCAT Lab Group - Health Research Institute Germans Trias i Pujol (IGTP), Badalona, Spain IGTP Badalona, Spain
2. ProCURE, Catalan Institute of Oncology, Oncobell, Bellvitge Institute for Biomedical Research (IDIBELL), L'Hospitalet del Llobregat, Barcelona, Catalonia, Spain.
Presenter e-mail: xfarrer@igtp.cat

The immune system plays a key role in cancer protection, with immunosurveillance providing the first defence against tumour cells through innate and adaptive immune response. Furthermore, there is extensive evidence of the relevance of immune cell homeostasis as an important prognostic outcome determinant in patients with cancer (Le Cornet et al., 2020). Although, this interplay between immune cell homeostasis and cancer is well described, little is known about the shared genetic basis between basic immune cell counts and cancer risk. To this end, we analysed the shared genetic architecture between 27 immune cell counts and 22 cancer types, using GWAS summary statistics and state-of-the art methods to detect genetic correlations, as well as to unveil the pleiotropic loci that influence both phenotypes. This allowed us to unearth new loci associated with several cancer types and get a better understanding of its pathophysiology.

STRUCTURAL CHARACTERIZATION OF NORAD, A HUMAN LNCRNA DYSREGULATED IN CANCER

Uciel Chorostecki[1,2], Ester Saus[1,2] , and Toni Gabaldón[1,2,3]

1. Barcelona Supercomputing Centre (BSC-CNS). Jordi Girona, 29. 08034. Barcelona, Spain.
2. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain
3. Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
Presenter e-mail: ucielp@gmail.com

The ENCODE project has reported that around 90% of the human genome is biologically active, and possibly more than 50% of the human genome codes for non-coding RNAs (ncRNAs). Long non-coding RNAs (lncRNAs) represent a heterogeneous group of ncRNAs, longer than 200 nucleotides, and it remains unclear the function of most of them. Several studies have demonstrated that the misregulation of lncRNAs may be among the causes of many complex human diseases, including cancer, neurological disorders, inflammatory response and roles in immunity. Furthermore, recent studies have shown that the functions of lncRNAS are mediated mainly by their structures.

We present a structural characterization of the long non-coding RNA activated by DNA damage (NORAD). NORAD is a conserved and highly expressed lncRNA that regulates genomic stability by interacting with proteins and microRNAs. It has been implicated in cancer and other disease processes. Previous characterizations of NORAD were only based on sequence-based computational inferences and have identified a modular organization of NORAD composed of several NORAD repeat units (NRUs). These units comprise the protein-binding elements and are separated by regular spacers of unknown function.

Here, we experimentally determined for the first time the secondary structure of NORAD using in-vitro enzymatic probing coupled to Illumina sequencing (nextPARS approach) to provide a thorough characterization of NORAD structure. Our results suggest that the spacer regions provide structural stability to NRUs. Furthermore, we uncover two previously-unreported NRUs and determine the core structural motifs conserved across NRUs. Given the current interest in understanding the structure, function and evolution of lncRNAs, and the clinical relevance of NORAD, we think that eventually, this may open the door to design novel therapeutic strategies in the fight against cancer.

Using bioactivity signatures to predict drug-target interactions

Paula Torren-Peraire[1], Miquel Duran-Frigola[1], Adrià Fernández-Torras[1], Patrick Aloy[1,2]

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

Drug-target interaction (DTI) prediction has become a paramount step in the drug discovery pipeline. When enough data is available, the chemical similarity of active compounds can be used to predict new interactions for a target of interest. With bioactivity data becoming more available, bioactivity signatures have been developed to complement chemical data with biological data. Making use of bioactivity signatures we have pre-trained hundreds of models, using human proteins from ChEMBL, that predict new interactions for these targets. Furthermore, acknowledging the importance of prediction reliability, we implement a measure of confidence that provides calibrated probabilities for each individual prediction. We further validate these predictions using time-series data to test them in real-world scenarios and provide a package to develop models for any in-house protein of interest.

Single-cell transcriptomics of iPSC-derived neurons reveals functional changes in Alzheimer's Disease

Ana Gutiérrez-Franco[1,2], Franz Arnold Ake[1,2], Sandra Fernández-Moya[1,2] and Mireya Plass[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: mplass@idibell.cat

Alzheimer's disease (AD) is the most common age-related neurodegenerative disease that heavily burdens healthcare systems worldwide. Most AD patients are sporadic and despite all the efforts, we still do not know the molecular mechanisms triggering the development of AD. One of the main problems to study AD and other neurodegenerative diseases is the lack of good experimental models that recapitulate the pathological features of the disease. In that context, induced pluripotent stem cell (iPSC) technology has provided an excellent tool to model disease pathogenesis considering the patients' genetic background.

In this project, we have used single-cell transcriptomics (scRNA-seq) to study the molecular changes that happen during the differentiation of iPSCs derived from sporadic AD patients to neurons. Preliminary results show that AD-derived neural progenitor cells already show changes in the expression of genes previously associated with AD or related to neuronal differentiation and RNA processing. These results demonstrate that neurons from sporadic AD patients show transcriptomic differences before the onset of the disease and thus can be used as a relevant model to study the molecular networks driving AD. Future work will be directed to validate these findings and assess its impact in the development of AD.

Taken together, our results that the combination of iPSC technology and scRNA-seq is a potent tool for the study of the molecular mechanisms triggering the development of neurodegenerative diseases such as AD.

LUNG TISSUE MULTI-LAYER NETWORK IN COPD

Nuria Olvera[1,2,3], Guillaume Noell[1,3], Jon Sánchez-Valle[2], Iker Núñez[2], Sandra Casas-Recasens[1,3], Alejandra Lopez-Giraldo[1,3,4], Angela Guirao[1,3,4], Rosalba Lepore[2], Davide Cirillo[2], Alvar Agustí[1,3,4], Alfonso Valencia[2], Rosa Faner[1,3]

1. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain.
2. Barcelona Supercomputing Center (BSC), Barcelona, Spain.
3. Centro de Investigación Biomédica en Red de Enfermedades Respiratorias M.P. (CIBERES), University of Barcelona Biomedicine Department, Barcelona, Spain
4. Respiratory Institute, Hospital Clinic, University of Barcelona, Barcelona, Spain.

Presenter email: olvera@clinic.cat

Chronic Obstructive Pulmonary Disease (COPD) is a prevalent disease characterized by airflow limitation and persistent respiratory symptoms. It is a highly heterogeneous condition as patients with the same level of airflow limitation (defined by FEV_1 % ref.) can present different levels of symptoms, comorbidities and biomarkers (i.e. overweight, cardiovascular disease and blood eosinophils). However, the biological mechanisms (endotypes) underlying this heterogeneity are still largely undefined.

We hypothesize that profiling lung tissue with different omics, and integrating the results using a suitable graph-based approach has the potential to tackle the biological heterogeneity.

Accordingly, in this study for the first time we integrate three omics levels (mRNA, miRNA and DNA methylation) determined in 135 lung tissue samples of ex-smokers with COPD. A patient network, identifying the molecular similarities between individuals, was built for each omics level resulting into a multi-layer network. Finally, communities of patients were detected within the multi-layer network.

Using this methodology, we identified four multi-omics communities, that presented significant differences in clinical features: (1) two of them included patients with higher FEV_1 % ref. and the other two presented lower FEV_1 % ref (p.val= 0.0031) , (2) comparing to the community with higher lung function, one of the severe groups was characterized by individuals with lower body mass index (p.val= 0.011) and the other by lower concentration of blood eosinophils (p.val= 0.012). Indeed, both communities were very dissimilar at the molecular level.

Specifically, at mRNA level the severe group with lower eosinophils had an increased inflammatory and damage response, as well as cilia dysfunction, which was mediated by the strongly downregulation of miR-34/449 family (required in ciliogenesis) at miRNA level.

In conclusion, we show for the first time that the use of a multi-layer network based on the lung tissue patient similarities, uncovered communities that differ in clinical COPD characteristics and shed light on the biological mechanisms underlying the heterogeneity of the disease.

Genomic characterization of eight established MPNST cell lines: a resource for precision medicine

Miriam Magallón-Lorenz[1], Ernest Terribas[1], Marco Fernández[2], Gerard Requena[2], Imma Rosas[3,4], Helena Mazuelas[1], Itziar Uriarte[1], Alex Negro[3,4], Elisabeth Castellanos[3,4], Juana Fernández Rodriguez[5,6], Conxi Lázaro[5,6], Meritxell Carrió[1], Bernat Gel[1,7], Eduard Serra[1,6]

1) Hereditary Cancer Group, Germans Trias i Pujol Research Institute (IGTP); Can Ruti Campus, Badalona, Barcelona, 08916; Spain
2) Cytometry Core Facility, Germans Trias i Pujol Research Institute (IGTP), Badalona, Barcelona, Spain
3) Clinical Genomics Research Unit, Germans Trias i Pujol Research Institute (IGTP); Can Ruti Campus, Badalona, Barcelona, 08916; Spain
4) Clinical Genomics Unit, Clinical Genetics Service, Northern Metropolitan Clinical Laboratory, Germans Trias i Pujol University Hospital (HGTP), Can Ruti Campus, Badalona, Barcelona, 08916; Spain
5) Hereditary Cancer Program, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona, 08098; Spain
6) Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Spain
7) Departament de Fonaments Clínics, Universitat de Barcelona, 08036 Barcelona, Spain

Presenter e-mail: mmagallon@igtp.cat

Malignant peripheral nerve sheath tumors (MPNST) are aggressive soft tissue sarcomas with a poor prognosis. Half of the tumors develop in the context of the genetic disease Neurofibromatosis type 1 (NF1), and the rest constitute sporadic sarcomas. There is a lack of therapeutic options beyond timely surgery. Thus, different cell lines have been established to facilitate their development. However, in many cases these cell lines have not been properly identified and comprehensively characterized at a genomic level, hampering their full use in precision medicine strategies.

To fill this gap, we present a complete characterization of 8 MPNST cell lines: 5 NF1 related and 3 sporadic cell lines. We encountered a misidentified cell line by STR profile analysis. We characterized the global ploidy, the complete copy number, structural rearrangements, small nucleotide variants and mutational signatures by using flow cytometry, SNP-array, WES and WGS. We also made a summary of the status of a set of MPNST-related genes, many of which inactivated by structural variants, showing the need of WGS for the analysis of this type of tumors. We compiled all this data in a summary sheet for each cell line.

Genomic characterization evidenced a significant degree of variability, specifically regarding the sporadic cell lines tested. Further methylome and marker characterization questioned the MPNST identity of a few cell lines, uncovering the need for a better diagnosis and classification of MPNSTs by systematic and complete characterization of both cell lines and tumors. This characterization might help NF1 and MPNST community to perform pharmacogenomic analyses and test new therapeutic strategies based on precision medicine.

Assessing the digenic model in primary immunodeficiencies using population whole-genome sequencing data

Nerea Moreno-Ruiz [1,2]; Genomics England Research Consortium, Oscar Lao [3,4,5], Juan Ignacio Aróstegui [6,7]; Hafid Laayouni [2,8]* and Ferran Casals[1,9]* .

1. Servei de Genòmica, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, Spain.
2. Institut de Biologia Evolutiva (UPF-CSIC), Departament de Ciències Experimentals I de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, Spain.
3. CNAG-CRG, Centre for Genomic Regulation, C/ Baldiri i Reixach 4, 08028, Barcelona, Spain
4. Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
5. Universitat Pompeu Fabra (UPF), Barcelona, Spain
6. Department of Immunology, Hospital Clínic - Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain
7. School of Medicine, Universitat de Barcelona, Barcelona, Spain
8. Bioinformatics Studies, ESCI-UPF, Barcelona, Spain.
9. Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona, Barcelona , Spain

*To whom correspondence should be addressed (ferran.casals@upf.edu, hafid.laayouni@upf.edu).

Abstract

An important fraction of patients with rare disorders remains with no clear genetic diagnostic, even after whole-exome or whole-genome sequencing. This poses a difficulty in giving adequate treatments and genetic counseling. The analysis of genomic data in rare disorders mostly considers the presence of single gene variants in coding regions that follow a concrete monogenic mode of inheritance. A digenic inheritance, with variants in two functionally-related genes in the same individual, is a plausible alternative that might explain the genetic basis of the disease in some cases. If this is the case, digenic disease combinations should be absent or underrepresented in healthy individuals. In this work, we develop a method to evaluate the significance of reported digenic combinations and detect new associations by interrogating whole-genome data from the Genomics England 100,000 Genomes Project cohort. In particular, we applied the method to detect new candidate digenic combinations in primary immunodeficiencies, a group of disorders in which the clinical heterogeneity and complex etiology, along with the existence of a large fraction of undiagnosed cases suggests a more complex scenario than that of monogenic diseases. Beyond the identification of some possible pathogenic associations, we also suggest that this approach will be relevant with the advent of new sequencing efforts in projects including hundreds of thousands of samples.

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project; http://www.genomicsengland.co.uk.

Novel binding site descriptors built upon inverse virtual screening

Arnau Comajuncosa-Creus[1], Miquel Duran-Frigola[1,2], Xavier Barril[2,4] and Patrick Aloy[1,4]

1. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
2. Ersilia Open Source Initiative, Cambridge, United Kingdom
3. Facultat de Farmàcia and Institut de Biomedicina, Universitat de Barcelona, Barcelona, Catalonia, Spain
4. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

Presenter e-mail: arnau.comajuncosa@irbbarcelona.org

Pocket descriptors embed relevant features of protein binding sites in the shape of numerical vectors. Unlike small molecule fingerprints, strategies to derive binding site descriptors are scarce and usually exhibit limited applicability. We herein present PocketVec, a novel strategy to derive comprehensible binding site descriptors based on the assumption that similar pockets bind similar ligands. By first docking an ordered set of molecules against the protein binding site of interest, we then store the corresponding molecular rankings into a numerical vector to build the pocket descriptor, which ends up representing the behaviour of the binding site against the set of docked molecules. In this way, we are able to provide a unique, meaningful and handy descriptor for each binding site, in a similar way molecular fingerprints do for small molecules. We then benchmark PocketVec descriptors in several pocket similarity exercises, showing remarkable great performances when addressing the sensitivity to the binding site flexibility and sensibility and detecting similar pockets in unrelated proteins derived from published literature. Indeed, PocketVec ranks as the 2nd and the 3rd best pocket comparison strategy when being compared using global metrics against other pocket descriptors (5) and alternative binding site comparison strategies (11), respectively, while overcoming most of their typical limitations and drawbacks. Finally, we combine PocketVec descriptors with molecular signatures in order to predict protein-ligand interactions from a proteome-wide perspective. In this way, we prove that PocketVec descriptors exhibit a significant early enrichment in such predictions.

Title: Integrative Modelling to explore functionality of cellular complexes

Author list: Altair C. Hernandez[1], Baldo Oliva[1], Damien P Devos[2] and Oriol Gallego[1]

[1] Department of Experimental and Health Science (DCEXS), Universitat Pompeu Fabra (UPF), Barcelona, 08003, Spain.
[2] Centro Andaluz de Biología del Desarrollo (CABD), Universidad Pablo de Olavide-CSIC, Carretera de Utrera km1, 41013 Sevilla, Spain

Abstract

The exocyst is an hetero-octamer responsible for tethering secretory vesicles to the plasma membrane during exocytosis. The exocyst and its interplay with the rest of the exocytic machinery (SNARE complex and GTPases) is essential for all eukaryotic cells. However, the complexity and dynamism of this protein machinery has maintained the molecular mechanism mediating the exocyst function unknown. Recently, the development of integrative approaches combining *in vitro* and *in situ* structural information opened up possibilities to study the molecular bases of cellular functions. Integrative modelling offers a unique opportunity to resolve complex and dynamic molecular systems, allowing high-resolution observations in a near-physiological context. We are now developing a method to integrate data *in vitro* (high-resolution structures obtained by cryo-EM) and *in situ* (functional structural data by live-cell imaging). We use the Integrative Modelling Platform (IMP) to set the representation of the system, transform the input data into spatial restraints and sample the configurational space of solutions using the Monte Carlo method. This approach allows determining the functional structure of the exocytic machinery including its complexity and structural dynamism. As a proof of concept, we have modelled the functional architecture of the exocyst complex. Our analysis provides information on the structural dynamics that mediate the activity of exocyst during vesicle tethering. This approach could be fundamental to resolve the interplay between the exocyst and the rest of the exocytic machinery.

CRISPR Analytics: a versatile and precise genome editing simulation and analysis tool

Marta Sanvicente-García[1], Socayna Joude[1], Albert Garcia-Valiente[1], Marc Güell[1]

1. Research Program on Translational Synthetic Biology, Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

Presenter-email: marta.sanvicente@upf.edu

Gene editing characterization with currently available tools leads to many false-positive results and does not give precise relative proportions among the different kinds of indels present in an edited bulk of cells. We have developed CRISPR-Analytics. CRISPR-A is a comprehensive and versatile genome editing web application tool (https://synbio.upf.edu/crispr-a/), as well as a nextflow pipeline, which gives support to gene editing experiments from design to analysis. Design can be assessed by the simulation of gene editing results. In addition, these results can be analyzed using the same tool. Therefore, this tool analyses multiple kinds of experiments: single cut experiments, base editing, primer editing, HDR... without the need of specifying the used experimental approach.

We have benchmarked the performance of aligning methods and edit calling systems together with the current gene editing variant calling tools (CRISPR-GA, Crispresso2, CrispRVariants, CRISPRpic, and cris.py). In addition, we propose the use of uni-molecular identifiers (UMIs) or spikes to get a more accurate quantification of the multiple alleles resulting from gene editing experiments. Sequenced negative controls can also be processed and used to do an empirical error subtraction. Thanks to all these new capabilities and parameters optimization, we have achieved higher accuracy than obtained with previous tools.

The best alignment algorithm and parameters have been defined using more than 100 simulated data sets, in addition to more than 400 sequenced samples from several gene editing experiments. Moreover, we have been able to correct biases in the amplification and sequencing processes. These biases are related to the deletion size and can be corrected with the use of UMIs in the first amplification of just two cycles. Regardless of these improvements, long deletions can still not be correctly characterized. The characterization of long deletions could be done with the sequencing of longer reads with sequencing platforms such as Nanopore instead of Illumina short-read sequencing platform.

Title: Analysis of nanopore data using Master of Pores 2

Authors: Luca Cozzuto[1], Anna Delgado1, Toni Hermoso1, Eva Novoa1,2, Julia Ponomarenko1,2

1. Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain

2. Universitat Pompeu Fabra, Barcelona, Spain

Master Of Pores (MoP) [1] is a suite of Nextflow-based pipelines for fully reproducible processing and analysis of the Nanopore sequencing data, both cDNA and RNA, with a focus on in-depth analysis of direct RNA-seq data. The first module of the software is for pre-processing raw fast5 files applying base calling (using either CPU or GPU) and read demultiplexing, filtering, aligning and assembling, producing a comprehensive report. The second module is used for predicting modified RNA bases applying four different approaches that can be chosen independently, or the consensus result can be called. The third module was developed for estimating the length of polyA tails using two independent approaches. We have developed a new version of the MoP software, MoP2, implemented using the new Nextflow DSL2 syntax. MoP2 is completely modular and scalable up to the amount of data produced by a PromethION flowcell. For small datasets, the integrated pipelines can be run on a laptop, and, for large datasets, in different HPC or cloud environments.

We will present the MoP2 software, its technical implementation and applications to studies of SARS-Cov2.

MoP is available on GitHub at https://github.com/biocorecrg/MOP2 <https://github.com/biocorecrg/MOP2>.

1 - Cozzuto L, Liu H, Pryszcz LP, Pulido TH, Delgado-Tejedor A, Ponomarenko J, Novoa EM. MasterOfPores: A Workflow for the Analysis of Oxford Nanopore Direct RNA Sequencing Datasets. Front Genet. 2020 Mar 17;11:211. doi: 10.3389/fgene.2020.00211. PMID: 32256520; PMCID: PMC7089958.

EFFICIENT AND ACCURATE PROTEIN-CODING GENE PREDICTION

Francisco Camara[1], Ferriol Calvet[1], Roderic Guigó[1,2]

1. Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Catalonia, Spain
2. Universitat Pompeu Fabra (UPF), Barcelona 08002, Catalonia, Spain

Identifying the encoded genes is essential to relate the genome sequence to the biology of species. This task is more complex than often assumed: twenty years after the sequencing of the human genome, and despite the substantial effort of the scientific community, the human gene catalogue has not yet been finalized. The Earth BioGenome Project will face the unprecedented challenge of efficiently and accurately assembling and annotating close to two million genome sequences in the next decade. Most computational methods to annotate genes and transcripts combine heterogeneous information: RNA sequencing data, statistical biases in the genome sequence, similarity to known coding sequences or to proteins in other species, etc. This produces accurate annotations, but it is computationally very expensive and may require additional sequencing data with its associated cost. Based on a simplified definition of a gene, we developed Geneid+BLASTx, a very efficient light-wise first-pass pipeline that accurately predicts protein-coding genes in eukaryotic genomes. We have benchmarked it on vertebrate genomes, which can be typically analyzed in times ranging from 15 minutes to two hours. Our method is based on combining the *ab initio* predictions produced by Geneid, with information from the matches between the genome to annotate and the human proteins annotated by GENCODE. The matches are found using DIAMOND BLASTx, which is a faster implementation of the BLASTx algorithm, and they are incorporated by Geneid when scoring the predicted exons. Comparing the performance of Geneid+BLASTx with other *ab initio* gene prediction tools, only Augustus is similar in terms of accuracy. However, in terms of time, our method is between 1.5 and 2 orders of magnitude faster. Using Geneid+BLASTx for producing an initial annotation of the vertebrates' genomes would shorten the time between the sequencing and the downstream analyses for many new species.

Building the foundations of a tech-enabled biology economy

Dr. Evan Floden
CEO & Co-founder of Seqera Labs

Biology — enabled by technology — is undergoing an industrial shift. This transformation stretches beyond traditional biotech drug development and into all aspects of health, agriculture, manufacturing and energy. We are still in the installation phase, akin to mobile in the 90’s or the internet prior to Google or Amazon. This phase necessitates building out the data infrastructure to engineer biology at scale. Seqera partners with customers like 23andMe to build this future, allowing them to organize and scale their critical data operations. We are on a mission to make data, infrastructure and collaboration seamless. The pandemic has highlighted just how important a task this is. It must be seen as our generation's call to arms and the spark for the industrial bio economy that will make the world a safer, cleaner and healthier place.

Biography

Evan Floden is CEO and co-founder of Seqera Labs and the open-source project Nextflow. He holds a Doctorate in Biomedicine from Universitat Pompeu Fabra (ES) for the large-scale deployment of analyses and is the author of 14 peer-reviewed articles. Prior to his Ph.D Evan obtained a BSc in Biotechnology from Victoria University (NZ) developing the tissue bioscaffold platform at Aroa Biosurgery which has been used in over 1 million patient procedures to date. Combining a passion for computing and biology, he completed an MSc in Bioinformatics at the University of Bologna (IT) before joining the Sanger Institute (UK). During his doctoral studies at the Centre for Genomic Regulation, he began working on Nextflow and co-founded Seqera Labs - the leading provider of scientific workflow orchestration software.

MULTI-TISSUE INTEGRATIVE ANALYSIS OF PERSONAL EPIGENOMES

Joel Rozowsky[1,2$], Jorg Drenkow[3$], Yucheng T Yang[1,2$], Gamze Gursoy[1,2$], Timur Galeev[1,2$], Beatrice Borsari[4$], Charles B Epstein[5$], Kun Xiong[1,2$], Jinrui Xu[1,2$], Jiahao Gao[1,2$], The EN-TEx consortium, Michael C Schatz[6,7#], Roderic Guigó[4,8#], Bradley E Bernstein[5,9#], Thomas R Gingeras[3#], Mark Gerstein[1,2#]

$ - co-first authors
# - co-senior authors
1 - Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
2 - Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA
3 - Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
4 - Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
5 - Broad Institute of MIT and Harvard, Cambridge, MA, USA
6 - Departments of Computer Science and Biology, Johns Hopkins University, Baltimore, MD, USA
7 - Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
8 - Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
9 - Department of Pathology and Center for Cancer Research, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

Presenter e-mail: beatrice.borsari@crg.eu

Evaluating the impact of genetic variants on transcriptional regulation is a central goal in biological science that has been constrained by reliance on a single reference genome. To address this, we constructed phased, diploid genomes for four cadaveric donors using long-read sequencing, and systematically charted noncoding regulatory elements and transcriptional activity across more than 25 tissues from these donors. Integrative analysis revealed over a million variants with allele-specific activity, coordinated, locus-scale allelic imbalances, and structural variants impacting proximal chromatin structure. We relate the personal genome analysis to the ENCODE encyclopedia, annotating allele- and tissue-specific elements that are strongly enriched for variants impacting expression and disease phenotypes. These experimental and statistical approaches, and the corresponding EN-TEx resource, provide a framework for personalized functional genomics.

Genome Structural Variants analysis in chronic diseases through SV imputation using the GCAT|Panel, the first haplotype-based reference panel from Iberian Population

Natalia Blay[1], Xavier Farre[1], Jordi Valls-Margarit[2], Iván Galván-Femenía[1,3], Daniel Matías-Sánchez[2], Montserrat Puiggròs[2], Anna Carreras[1], Cecilia Salvoro[2], Beatriz Cortés[1], Ramon Amela[2], Jon LergaJaso[4], Marta Puig[4], Jose Francisco Sánchez-Herrero[5], Victor Moreno[6,7,8,9], Manuel Perucho[10,11], Lauro Sumoy[5], Lluís Armengol[12], Olivier Delaneau[13,14], Mario Cáceres[4,15], David Torrents[2,15] & Rafael de Cid[1]

1. Genomes for Life-GCAT lab Group, Institute for Health Science Research Germans Trias i Pujol (IGTP), Badalona, 08916, Spain.
2. Life sciences dept, Barcelona Supercomputing Center (BSC), Barcelona, 08034, Spain.
3. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028, Barcelona, Spain (current affiliation).
4. Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, 08193, Spain.
5. High Content Genomics and Bioinformatics Unit, Institute for Health Science Research Germans Trias i Pujol (IGTP), 08916, Badalona, Spain.
6. Catalan Institute of Oncology, Hospitalet del Llobregat, 08908, Spain.
7. Bellvitge Biomedical Research Institute (IDIBELL), Hospitalet del Llobregat, 08908, Spain.
8. CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, 28029, Spain.
9. Universitat de Barcelona (UB), Barcelona, 08007, Spain.
10. Sanford Burnham Prebys Medical Discovery Institute (SBP), La Jolla, CA 92037, USA.
11. Cancer Genetics and Epigenetics, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Health Science Research Institute Germans Trias i Pujol (IGTP), Badalona, 08916, Spain.
12. Quantitative Genomic Medicine Laboratories (qGenomics), Esplugues del Llobregat, 08950, Spain.
13. Department of Computational Biology, University of Lausanne, Génopode, 1015 Lausanne, Switzerland.
14. Swiss Institute of Bioinformatics (SIB), University of Lausanne, Quartier Sorge – Batiment Amphipole, 1015 Lausanne, Switzerland.
15. ICREA, Barcelona, 08010, Spain.

Presenter e-mail: nblaym@igtp.cat

The combined analysis of genetic and phenotypic data is the common first approach to explore the genetic architecture of human diseases. Most of the genetic studies are mainly based on single nucleotide variants (SNV) and small insertions and deletions (indels), not considering structural variants (SV) such as large insertions and deletions, inversions, translocations or transposable elements, leaving an important part of the genetic variation unexplored. This, weighs down the conclusions, since the functional role expected from most of these variants.

Currently, high performance whole genome sequencing (WGS) with long reads promises the comprehensive analysis of SV, but the costs of the analysis, especially when large cohorts are required is a bottleneck for SV analysis generalisation. To fill this important gap, GCAT|Genomes for Life (http://www.gcatbiobank.org/), created together with BSC, a new panel for SV imputation: the GCAT|Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing, the first imputation panel focused on SV.

We present here this new GCAT|Panel resource, and its application in the analysis of 70 chronic conditions. We used the resources of the GCAT|Genomes for Life project, that includes information from an adulthood cohort of 20,000 individuals (Obón-Santacana et al., 2018) and longitudinal linked Electronic Health Records (EHR) (2010-2020) to explore the role of this comprehensive characterized genomic profiles, including SNV and SV.

This study (1) puts the spotlight on the role of this type of variants in large genome studies and (2) the importance of a panel focused on SV, to systematically include SV into genome-wide genetic studies.

perSVade: personalized Structural Variation detection in your species of interest

Miquel Àngel Schikora-Tamarit[1,2] and Toni Gabaldón[1,2,3]

[1] *Barcelona Supercomputing Centre (BSC-CNS). Jordi Girona, 29. 08034. Barcelona, Spain.*

[2] *Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.*

[3] *Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.*

Structural variants (SV) such as translocations, inversions, deletions, and other genomic rearrangements contribute significantly to genetic and phenotypic variability. The role of SV has been traditionally overlooked due to technical limitations for SV detection and interpretation from short-read sequencing datasets. Most available algorithms yield low recall when tested on humans, but few studies have investigated performance in non-human genomes. Similarly, despite remarkable differences across species' genomes, most approaches use standard parameters, generally optimized for humans. In order to fill this gap and enable tailored approaches for each species, we have developed perSVade (personalized Structural Variation Detection), a pipeline that identifies and annotates SVs in a way that is optimized for any input sample. Starting from a set of paired-end whole-genome sequencing reads, perSVade uses simulations on the reference genome to choose the best SV calling parameters. The output includes the optimally-called SVs and a report of the accuracy, useful to assess the confidence in the results. In addition, perSVade allows the calling of small variants and copy-number variation. In summary, perSVade identifies several types of genomic variation from short reads and using sample-optimized parameters. We validated that perSVade increases the SV calling accuracy on both simulated and real variants for six diverse eukaryotic organisms. Importantly, we find that there is no universal set of "optimal" parameters, which makes our method essential to yield accurate variant calls. We consider that this tool will help to understand how SVs generate phenotypes across non-human organisms.

Keywords: structural variants, genomics, parameter optimization

PopHumanVar: an interactive application for the functional characterization and prioritization of adaptive genomic variants in humans

Aina Colomer-Vilaplana[1], Jesus Murga-Moreno[1,2], Aleix Canalda-Baltrons[1], Clara Inserte[2], Daniel Soto[1], Marta Coronado-Zamora[1,2,3], Antonio Barbadilla[1,2] and Sònia Casillas[1,2]

[1]Department of Genetics and Microbiology, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain; [2]Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain; [3]Present address: Institute of Evolutionary Biology, CSIC-Universitat Pompeu Fabra, Barcelona 08003, Spain.

Adaptive challenges that humans faced as they expanded across the globe left specific molecular footprints that can be decoded in our today's genomes. Different sets of metrics are used to identify genomic regions that have undergone selection. However, there are fewer methods capable of pinpointing the allele ultimately responsible for this selection. Here, we present PopHumanVar, an interactive online application that is designed to facilitate the exploration and thorough analysis of candidate genomic regions by integrating both functional and population genomics data currently available. PopHumanVar generates useful summary reports of prioritized variants that are putatively causal of recent selective sweeps. It compiles data and graphically represents different layers of information, including natural selection statistics, as well as functional annotations and genealogical estimations of variant age, for biallelic single nucleotide variants (SNVs) of the 1000 Genomes Project phase 3. Specifically, PopHumanVar amasses SNV-based information from GEVA, SnpEFF, GWAS Catalog, ClinVar, RegulomeDB and DisGeNET databases, as well as accurate estimations of iHS, nS_L and iSAFE statistics. The utility of PopHumanVar has been tested on frequently reported candidate genomic regions in genome-wide scans for positive selection in humans, including regions close to the genes *EDAR* (chr2:109450405-109606617; GRCh37/hg19), which is associated to hair follicle thickness and straightness and shovel shaped incisors in East-Asians; *LCT* (chr2:135792491-136822774; GRCh37/hg19), which is associated to lactase persistence in several human populations; and *ACKR1* (*DARC*, chr1:159174665-159176290; GRCh37/hg19), which is associated with resistance to malaria in Africans. In all cases, PopHumanVar is able to identify the causal variant reported in previous studies (rs3827760, rs4988235/rs182549/rs145946881, and rs2814778, respectively). PopHumanVar is open and freely available at https://pophumanvar.uab.cat.

BacterialTyper: a general purpose suite for comprehensive analysis of bacterial whole genome sequencing data for clinical and epidemiological applications

Sanchez Herrero FJ1,4, Pluvinet R1,4, Lacoma A2,3,4, Molina B3,4, Gimenez M2, Casañ C2, Antuori A2, Blanco I5, Matas L2, Cardona PJ2,3,4, Saludes V2,3,4, Martró E2,3,4, Prat Aymerich C2,3,4,6,7,Sumoy L1,4

1 High Content Genomics and Bioinformatics, Institut Germans Trias i Pujol (IGTP),Badalona, Spain

2 Microbiology Department, Northern Metropolitan Clinical Laboratory, Hospital Universitari Germans Trias i Pujol (HUGTP), Badalona, Spain

3 CIBER in Epidemiology and Public Health (CIBERESP), Madrid, Spain.

4 Institut d'Investigació en Ciències de la Salut Germans Trias i Pujol (IIS-IGTP), Badalona, Spain

5 Northern Metropolitan Clinical Laboratory, Hospital Universitari Germans Trias i Pujol (HUGTP), Badalona, Spain

6 Universitat Autònoma de Barcelona (UAB), Badalona, Spain.

7 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.

Microbiological research has benefited from the development of NGS technology enormously in recent years by gaining access to genotypic information allowing rapid diagnosis. The COVID 19 pandemic has demonstrated the power of sequencing for tracking outbreak and identifying emergent new variants. Many tools have been developed but too often they are focused on a single microorganism. We have developed a pipeline, named BacterialTyper (https://github.com/HCGB-IGTP/BacterialTyper), which can process WGS data at different levels allowing for both rapid bacterial sample typing and in depth analysis of whole genomes with or without a reference. The tool uses publicly available k-mer based methods for fast matching to reference genomes or genes which enables closest matching to any reference database. This can be used in turn for reference mapping or for guiding whole genome assembly. Antibiotic resistance and virulence gene profiles are also generated during processing of samples. Detection of extra-chromosomal elements such as plasmids and transposable elements, or phage insertions or copy number variation is also built in the tool. Variant identification relative to a common reference and one to one comparison allows constructing distance matrices for tree reconstruction with applications in phylogenetic and outbreak analysis. Genome annotation as well as functional interpretation of detected variants is enabled as well. The tool has already been used for the study of S. aureus outbreak analysis (Lacoma et al, 2021), and tested on other genera such as Mycobacterium or Klebsiella. The pipeline may be adapted for each pathogen of interest, taking into account their specific genomic characteristics.

Funded by core funding to HCGB-IGTP to LS (recognized by Generalitat de Catalunya, 2017 SGR 484) and by ISCIII to CP (PI17/01139) integrated in the National R + D + I and funded by the ISCIII and the European Regional Development Fund. JFSH was beneficiary of a contract partially funded by ISCIII through the Acción Estratégica en Salud 2018 (Co-funded by the European Regional Development Fund/European Social Fund; "A way to make Europe"/"Investing in your future") CA18/00019 (contrato de técnico bioinformático de apoyo a la investigación en los IIS acreditados del SNS).

SolupHred: a server to predict the pH-dependent aggregation of intrinsically disordered proteins

Oriol Bárcenas, Carlos Pintado, Jaime Santos, Valentín Iglesias and Salvador Ventura

Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra (Barcelona) 08193, Spain

Presenter e-mail: oriol.barcenas@autonoma.cat

Polypeptides are exposed to changing environmental conditions that modulate their intrinsic aggregation propensities. Intrinsically disordered proteins (IDPs) constitutively expose their aggregation determinants to the solvent, thus being especially sensitive to its fluctuations. However, solvent conditions are often disregarded in computational aggregation predictors. We recently developed a phenomenological model to predict IDPs' solubility as a function of the solution pH, which is based on the assumption that both protein lipophilicity and charge depend on this parameter. The model anticipated solubility changes in different IDPs accurately. In this application note, we present SolupHred, a web-based interface that implements the aforementioned theoretical framework into a predictive tool able to compute IDPs aggregation propensities as a function of pH. SolupHred is the first dedicated software for the prediction of pH-dependent protein aggregation.

A3D Database: Structure-based Protein Aggregation Predictions for the Human Proteome

Javier Garcia-Pardo[1], Aleksandra E. Badaczewska-Dawid[2], Aleksander Kuriata[2], Jordi Pujols[1], Sebastian Kmiecik[2] and Salvador Ventura[1].

[1]Institut de Biotecnologia i de Biomedicina (IBB) and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.

[2] Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland.

Protein aggregation is associated with highly debilitating human disorders and constitutes a major bottleneck for producing therapeutic proteins. Our knowledge of the human protein structures repertoire has dramatically increased with the recent development of the AlphaFold (AF) deep-learning method. This structural information can be used to understand better protein aggregation properties and the rational design of protein solubility. This article uses the Aggrescan3D (A3D) tool to compute the structure-based aggregation predictions for the human proteome and make the predictions available in a database form.

Here, we present the A3D Database, in which we analyze the AF-predicted human protein structures (for over 17 thousand non-membrane proteins) in terms of their aggregation properties using the A3D tool. Each entry of the A3D Database provides a detailed analysis of the structure-based aggregation propensity computed with A3D. The A3D Database implements simple but useful graphical tools for visualizing and interpreting protein structure datasets. We discuss case studies illustrating how the database could be used to analyze physiologically relevant proteins. Furthermore, the database enables testing the influence of user-selected mutations on protein solubility and stability, all integrated into a user-friendly interface.

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] Sívia 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-Valéry Turatsinze,[2] Jason M. Torres,[8,9] Vibe Nylander,[10] Ji Chen,[11] Lena Eliasson,[7] Matthieu Defrance,[2] Ramon Amela,[1] MAGIC Consortium, 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,[1,19,20,21,24,#] Miriam Cnop,[2,22,24,25,#] and David Torrents[1,23,24,#]

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. Senior author
25. Lead contact

* These authors contributed equally

Correspondence: mercader@broadinstitute.org (J.M.M.), mcnop@ulb.ac.be (M.C.), david.torrents@bsc.es (D.T.)

Presenter e-mail: ignasi.moran@bsc.es

Genome-wide association studies (GWASs) identified hundreds of signals associated with type 2 diabetes (T2D). To gain insight into their underlying molecular mechanisms, we have created the translational human pancreatic islet genotype tissue-expression resource (TIGER), aggregating >500 human islet genomic datasets from five cohorts in the Horizon 2020 consortium T2DSystems. We impute genotypes using four reference panels and meta-analyze cohorts to improve the coverage of expression quantitative trait loci (eQTL) and develop a method to combine allele-specific expression across samples (cASE). We identify >1 million islet eQTLs, 53 of which colocalize with T2D signals. Among them, a low-frequency allele that reduces T2D risk by half increases CCND2 expression. We identify eight cASE colocalizations, among which we found a T2D-associated SLC30A8 variant. We make all data available through the TIGER portal (http://tiger.bsc.es), which represents a comprehensive human islet genomic data resource to elucidate how genetic variation affects islet function and translates into therapeutic insight and precision medicine for T2D.

Leveraging comparative genomics across primates to decipher the genomic architecture of complex traits

Alejandro Valenzuela[1], Lukas F.K. Kuderna[1], Joseph D. Orkin[1], Fabio Barteri[1], Borja Esteve-Altava[1], Claudia Vasallo[2], Carlos Morcillo[1], Mareike C. Janiak[3], Amanda D. Melin[4,5,6], Robin M. D. Beck[7], Jean P. Boubli[7], Kyle KaiHow Farh[8], Jeffrey Rogers[9], Gerard Muntané[1,10], Tomàs Marquès-Bonet[1,2,11,12], David Juan[1], Arcadi Navarro[1,2,11], Primate Sequencing Conservation Initiative

[1]Institute of Evolutionary Biology (UPF-CSIC), PRBB, Dr. Aiguader 88, 08003 Barcelona, Spain
[2]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
[3]School of Science, Engineering & Environment, University of Salford, Salford, UK
[4]Department of Anthropology and Archaeology, University of Calgary, Alberta, Canada
[5]Department of Medical Genetics, University of Calgary, Alberta, Canada
[6]Alberta Children's Hospital Research Institute, University of Calgary, Alberta, Canada
[7]School of Environmental and Life Sciences, University of Salford, Salford M5 4WT, UK
[8]Artificial Intelligence Lab, Illumina Inc, San Diego, California, USA
[9]Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA
[10]Hospital Universitari Institut Pere Mata, IISPV, Universitat Rovira i Virgili, Biomedical Network Research Centre on Mental Health (CIBERSAM), Reus, Spain
[11]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain
[12]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain
Presenter e-mail: alejandro.valenzuela@upf.edu

The study of the genetic architecture of complex traits and diseases has focused on Genome-Wide Association Studies (GWAS), the search for associations between genetic variation and of intra-specific genetic differences in such traits. Large-scale comparative genomics affords the opportunity of focusing on the traits themselves –including their presence and absence– uncovering relevant and actionable pathways, genes and even individual mutations that have remained undetectable so far. We catalogue 265 complex primate trait measurements, covering morphology, behavior, ecology, physiology and life-history strategies. We use this resource to perform genome-phenome analysis focusing on the protein-coding genes of 230 different primate species. We track genomic changes building up relevant phenotypes along the primate phylogeny, focusing on convergent amino acid substitutions across primate families and shifts in the rates of protein evolution of the genes under study as genomic sources of variation. This approach allows us to recover thousands of phylogenetic gene-trait associations, both at the level of genes and of individual amino acid changes, improving our knowledge of the genomic basis for traits of evolutionary and biomedical relevance for primates in general and humans in particular.

XICRA: a pipeline for integrated analysis of small RNA sequencing data with error correction from paired end reads

Sanchez Herrero FJ, Lopez Balastegui M, Luna de Haro A, R. Pluvinet, Sumoy L
High Content Genomics and Bioinformatics, Institut Germans Trias i Pujol (IGTP), Badalona, Spain

NGS methods applied to Small RNA sequencing allows characterization of small non coding RNAs (including miRNA, tRNA, snoRNA, piRNA among others). Full read coverage of these short RNA molecules is akin to single molecule profiling and offers allows to fully discriminate between different isoforms, such as isomiRs in the case of miRNAs. However, true definition of these isoforms from single read data is confounded by intrinsic technical errors in Illumina sequencing by synthesis technology. We have developed a pipeline, named XICRA (https://github.com/HCGB-IGTP/XICRA), that focuses on exploiting paired end reads to detect and correct some of these errors to provide more precise profiles. XICRA has different modules that can integrate different short read alignment algorithms and makes use of latest annotation consensus for unequivocal identification of individual isoforms. It processes samples from raw sequencing data. It also enables differential expression analysis at different levels (such as miRNA, canonical miRNA, isomiR class, or individual isomiR) . We have already applied this tool to show that single read datasets may have an overrepresentation of SNP isoforms (Sanchez Herrero et al , 2021) and we have shown the benefits of using paired end reads. This has implications for the interpretation of isomiR variation as these are the isoforms that confer changes in the seed region that are hypothesized to result in major functional consequences due to shifts in mRNA target specificity. Ongoing work is aimed at expanding the tool for creating a comparison of samples regarding RNA biotype profile and specific analysis application to other common ncRNA such as tRNA derived fragments (tRFs) and piRNAs.

Funded by ISCIII (PI10-01154) to LS and IGTP core funding to HCGB-IGTP (recognized by Generalitat de Catalunya, 2017 SGR 484). FJSH was a recipient of a ISCIII contrato de técnico bioinformatico de IIS (AES02o18 CA18.00019) JFSH was beneficiary of a contract partially funded by ISCIII through the Acción Estratégica en Salud 2018 (Co-funded by the European Regional Development Fund/European Social Fund; “A way to make Europe”/“Investing in your future”) CA18/00019 (contrato de técnico bioinformático de apoyo a la investigación en los IIS acreditados del SNS).

DispHScan: A Multi-Sequence Web Tool for Predicting Protein Disorder as a Function of pH

Carlos Pintado-Grima[1], Valentín Iglesias[1], Jaime Santos[1], Vladimir N. Uversky[2] and Salvador Ventura[1]

[1] Institut de Biotecnologia i Biomedicina, Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain.

[2] Department of Molecular Medicine, USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA.

Proteins are exposed to fluctuating environmental conditions in their cellular context and during their biotechnological production. Disordered regions are susceptible to these fluctuations and may experience solvent-dependent conformational switches that affect their local dynamism and activity. In a recent study, we modeled the influence of pH in the conformational state of IDPs by exploiting a charge–hydrophobicity diagram that considered the effect of solution pH on both variables. However, it was not possible to predict context-dependent transitions for multiple sequences, precluding proteome-wide analysis or the screening of collections of mutants. In this article, we present DispHScan, the first computational tool dedicated to predicting pH-induced disorder–order transitions in large protein datasets. The DispHScan web server allows the users to run pH-dependent disorder predictions of multiple sequences and identify context-dependent conformational transitions. It might provide new insights on the role of pH-modulated conditional disorder in the physiology and pathology of different organisms. The DispHScan web server is freely available for academic users at *http://disphscan.ppmclab.com*, it is platform-independent and does not require previous registration

The Bioteque, a comprehensive repository of biological embeddings

Biomedical data are accumulating at an unprecedented rate and integrating them in a unified framework is a major challenge of the post-genomics era. We have created a gigantic heterogeneous network (more than 550k nodes and 30M edges) that harmonizes and connects data points from over >200 data sources. Overall, 12 types of biological entities (e.g. genes, diseases, drugs) were linked by 70 types of relationships between them (e.g. drug treats disease, gene interacts with gene). This network constitutes the basis to explore more complex rationales, such as the functional relationship between two disease-associated genes that operate within a certain pathway, or a drug repurposing opportunity discovered by virtue of multiple mechanistic connections between two apparently unrelated drugs. Furthermore, in order to properly exploit the gathered knowledge, we systematically encoded these connections as numerical vectors (aka embeddings) creating the Bioteque, a resource of biological network embeddings of unprecedented size and scope. We prove this concise representation of the data retains the meaningful information contained within the biological network, can be plugged to machine learning implementations and show how it can be used to characterize a given set of experimental observations.

Long read nanopore RNA sequencing improves planarian transcriptome annotation and differential gene expression analysis approaches.

Maria Rosselló[1,2], Teresa Adell[1,2], Emili Saló[1,2], Josep F Abril[1,2]

1. Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia (UB), Avinguda diagonal 643, 08028, Barcelona, Catalonia, Spain.
2. Computational Genomics Lab; Institut de Biomedicina (IBUB), Universitat de Barcelona, Barcelona, Catalonia, Spain.

Presenter email: mariarossello@ub.edu

Planarians have become a model organisms in regeneration and adult tissue renewal research. Despite its relevance in that field, *Schmidtea mediterranea* genomic and transcriptomic diversity is not completely understood yet. In this work we propose a new approach to obtain full length transcripts using a single-molecule long-read sequencing method based on Oxford Nanopore technologies (ONT) for the first time in this species. Using nanopore sequencing data we could successfully assemble a transcriptome and describe some alternative splicing isoform events. We further implemented some improvements on the computational protocol to correct the errors derived from the sequencing methodology to retrieve the proper ORFs translations. From the multiplexed samples of our sequencing run, we demonstrated that data derived from nanopore sequencing can be applied to differential gene expression (DGE) analyses too. The results from the comparison among different physiological states in the planarian led us to undercover key genetic pathways, which play a role on the molecular mechanisms that control regeneration and body size.

In conclusion, in this work we established that single-molecule long-read sequencing can be a reliable method for *de novo* transcriptome assembly as well as DGE analyses in *S. mediterranea*.

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 PLAPA, a tool for characterising APA sites at single cell resolution using 10X Genomics or Dropseq scRNA-seq dataset. PLAPA allows quantifying RNA expression at isoform level and single-cell resolution and identifying changes in isoform usage across cell populations and conditions. We used PLAPA 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 PLAPA 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.

3D chromatin remodeling in the germ line modulates genome evolutionary plasticity

Lucía Álvarez-González[1,2,#], Frances Burden[3,#], Dadakhalandar Doddamani[3,#], Roberto Malinverni[4], Cristina Marín-García[1,2], Laia Marin-Gual[1,2], Albert Gubern[1], Covadonga Vara[1,2], Andreu Paytuví-Gallart[1,2,5], Marcus Buschbeck[5,6], Peter Ellis[3], Marta Farré[3], Aurora Ruiz-Herrera[1,2]

[1]Departament de Biologia Cel·lular, Fisiologia i Immunologia, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain. [2]Genome Integrity and Instability Group, Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain. [3]School of Bioscience, University of Kent, UK. [4]Cancer and Leukemia Epigenetics and Biology Program, Josep Carreras Leukaemia Research Institute (IJC), Campus ICO-GTP-UAB, Badalona, Spain. [5]Sequentia Biotech, Cerdanyola del Vallès, 08193, Spain. [6]Program for Predictive and Personalized Medicine of Cancer, Germans Trias i Pujol Research Institute (PMPPC-IGTP), Badalona, Spain. [#] Equally contribution

The spatial folding of chromosomes and their organization in the nucleus has profound regulatory impacts on gene expression and genome architecture, whose evolutionary consequences are far from being understood. Here we explore the evolutionary plasticity of the 3D chromatin remodelling in the germ line given its pivotal role in the transmission of genetic information. Using a comprehensive integrative computational analysis, we (i) reconstruct ancestral rodent genomes analyzing whole-genome sequences of 14 rodent species representatives of the major phylogroups, (ii) detect lineage-specific chromosome rearrangements and (iii) identify the dynamics of the structural and epigenetic properties of evolutionary breakpoint regions throughout mouse spermatogenesis by applying integrative computational analyses. Our results show that evolutionary breakpoint regions are devoid of programmed meiotic DSBs and meiotic cohesins in primary spermatocytes but associated with functional long-range interaction regions and sites of DNA damage in post-meiotic cells. Moreover, we detect the presence of long-range interactions in spermatids that recapitulate ancestral chromosomal configurations. Overall, we propose a model, which integrates evolutionary genome reshuffling with DNA damage response mechanisms and the dynamic spatial genome organization of germ cells.

Introducing FastCAAS, a tool for detection and statistical validation of Convergent Amino-Acid Substitutions in orthologous protein alignments.

Fabio Barteri.

ABSTRACT

Background.

Convergent evolution represents an opportunity to understand the genetic bases of biological diversity. The identification of amino-acid substitutions that are consistent with phenotypic convergence allows highlighting the genes that are likely to be associated with a specific trait and understanding which genetic changes are responsible for phenotype determination. Convergent Amino Acid Substitutions (CAAS) between two groups of species can be identified by scanning Multiple Sequence Alignments (MSA) of orthologous proteins. Identified CAAS are validated through a bootstrap analysis that consists in repeating the CAAS discovery with randomised groups. CAAS workflow implementation on a genome scale is made challenging by the high computational costs associated with a large set of alignments and/or species.

Results.

Here I introduce FastCAAS, a computational tool that detects and validates CAAS in orthologous proteins. FastCAAS allows CAAS discovery and validation at a large scale. It is designed to work on distributed computer systems and through parallel execution. FastCAAS represents a solid and fast software solution for CAAS analysis.

Availability and implementation

FastCAAS source code is going to be available in GitHub, along with documentation and examples. FastCAAS is written in Python v. 3.8+. The tool imports Brownian motion-based randomisation from R package RERconverge (https://github.com/nclark-lab/RERconverge) through a script written in R v. 4.0+. The multiple alignments manipulation relies on Biopython v. 1.7+.

Urinary metabolite quantitative trait loci in children and their interaction with dietary factors

Beatriz Calvo-Serra[1,2,3], Léa Maitre[1,2,3], Chung H Lau[4], Alexandros P Siskos[4], Kristine B Gützhow[5], Sandra Andrusaityte[6], Maribel Casas[1,2,3], Leda Chatzi[7], Juan R González[1,2,3], Regina Grazuleviciene[6], Rosie McEachan[8], Remy Slama[9], John Wright[8], Murieann Coen[4,10], Matine Vrijheid[1,2,3], Hector Keun[4], Geòrgia Escaramís[11,3¶], Mariona Bustamante[1,2,3¶*]

[1] ISGlobal, Institute for Global Health, Barcleona, Spain
[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain
[3] CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[4] Cancer Metabolism & Systems Toxicology Group, Division of Cancer, Department of Surgery and Cancer & Division of Systems Medicine, Department of Metabolism, Digestion & Reproduction, Imperial College London, London, United Kingdom
[5] Department of Environmental Health, Norwegian Institute of Public Health, Oslo, Norway
[6] Department of Environmental Science, Vytautas Magnus University, Kaunas, Lithuania
[7] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA
[8] Bradford Institute for Health Research, Bradford, United Kingdom
[9] University Grenoble Alpes, Inserm, CNRS, Team of Environmental Epidemiology Applied to Reproduction and Respiratory Health, IAB, Grenoble, France
[10] Oncology Safety, Clinical Pharmacology and Safety Sciences, R&D Biopharmaceuticals, AstraZeneca, 1 Francis Crick Avenue, Cambridge, United Kingdom
[11] Universitat de Barcelona (UB), Barcelona, Spain

¶These authors contributed equally to this work.
* Corresponding author

Presenter email: beatriz.calvo@upf.edu

Human metabolism is influenced by genetic and environmental factors. Previous studies have identified over 23 single nucleotide polymorphisms (SNPs) associated with more than 26 urine metabolites levels in adults (urinary metabolite quantitative loci – metabQTLs), but there are no published data pertaining to children. The aim of the present study is the identification of urinary metabQTLs in children and their interaction with dietary patterns.

Association between genome-wide genotyping data and 44 urine metabolite levels measured by ^{1}H NMR was tested in 996 children from the Human Early Life Exposome (HELIX) project. Twelve statistically significant urine metabQTLs were identified, involving 11 unique loci and 10 different metabolites. Comparison with findings in adults revealed that six metabQTLs were already known, one had been described in serum and

three involved the same locus but different urinary metabolites. The remaining two metabQTLs represent novel urine metabolite-locus associations, reported for the first time in this study (SNP rs12575496 for taurine, and the missense SNP rs2274870 for 3-hydroxyisobutyrate). For the 10 known loci, functional annotation revealed the same potential causal gene as the one reported before in the literature. The novel locus associated with 3-hydroxyisobutyrate was annotated to *Nipsnap Homolog 3A (NIPSNAP3A)*, which codes for a protein highly expressed in the kidney that participates in vesicular transport, while our annotation strategy did not provide any candidate gene for the taurine association. We also found that urinary taurine levels were affected by the combined action of genetic variation and dietary patterns of meat frequency intake.

Overall, we identified 12 urinary metabQTLs in children, two of them being novel. A substantial part of the identified loci affected urinary metabolite levels both in children and in adults. The metabQTL for taurine seemed to be specific to children and interacted with dietary patterns.

GENOME-WIDE POSTNATAL CHANGES IN IMMUNITY FOLLOWING FETAL INFLAMMATORY RESPONSE

Daniel Costa[1,2], Núria Bonet[3], Amanda Solé[2,4,5], José Manuel González de Aledo-Castillo[6], Eduard Sabidó[2,4,5], Ferran Casals[3], Carlota Rovira[7], Alfons Nadal[8], Jose Luis Marin[9], Teresa Cobo[9], Robert Castelo[2,10]

1. Department of Pediatrics, Hospital de Figueres, Spain
2. Department of Experim ental and Health Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain
3. Genomics Core Facility, Department of Experimental and Health Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain
4. Proteomics Unit, Centre de Regulacio Genomica (CRG), Barcelona, Spain
5. Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
6. Inborn Errors of Metabolism Section, Laboratory of Biochemistry and Molecular Genetics, Hospital Clınic, Barcelona, Spain
7. Hospital Sant Joan de Deu, Barcelona, Spain
8. Department of Pathology, Hospital Clınic, Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), University of Barcelona, Spain
9. Hospital Clınic, Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), Centre for Biomedical Research on Rare Diseases (CIBER-ER), University of Barcelona, Spain
10. Research Programme on Biomedical Informatics, Institut Hospital del Mar d'Investigacions Mediques (IMIM), Barcelona, Spain
Presenter e-mail: robert.castelo@upf.edu

The fetal inflammatory response (FIR) increases the risk of perinatal brain injury, particularly in extremely low gestational age newborns (ELGANs, < 28 weeks of gestation). One of the mechanisms contributing to such a risk is a postnatal intermittent or sustained systemic inflammation (ISSI) following FIR. The link between prenatal and postnatal systemic inflammation is supported by the presence of well-established inflammatory biomarkers in the umbilical cord and peripheral blood. However, the extent of molecular changes contributing to this association is unknown. Using RNA sequencing and mass spectrometry proteomics, we profiled the transcriptome and proteome of archived neonatal dried blood spot (DBS) specimens from 21 ELGANs. Comparing FIR-affected and unaffected ELGANs, we identified 782 gene and 27 protein expression changes of 50% magnitude or more, and an experiment-wide significance level below 5% false discovery rate. These expression changes confirm the robust postnatal activation of the innate immune system in FIR-affected ELGANs and reveal for the first time an impairment of their adaptive immunity. In turn, the altered pathways provide clues about the molecular mechanisms triggering ISSI after FIR, and the onset of perinatal brain injury.

Reference:

Costa D, Bonet N, Solé A, González de Aledo-Castillo JM, Sabidó E, Casals F, Rovira C, Nadal A, Marin JL, Cobo T, Castelo R. Genome-wide postnatal changes in immunity following fetal inflammatory response. The FEBS journal. 2021 Apr;288(7):2311-31.

Shared Genetics for inflammatory conditions in a adulthood cohort (GCAT)

Beatriz Cortés[1], Xavier Farré[1], Natalia Blay[1], Anna Carreras[1], Gemma Castaño-Vinyals[2], Manolis Kogevinas[2], Rafael de Cid[1].
1 Health Research Institute Germans Trias i Pujol (IGTP), Badalona, Spain. Genomes for Life-GCAT Lab Group - Program of Predictive and Personalized Medicine of Cancer (PMPPC), Badalona, Spain
2 Instituto de Salud Global de Barcelona (ISGlobal), 08036 Barcelona, Spain.

Presenter e-mail: bcortes@igtp.cat

Inflammation, as the central component of innate immunity, has been recognized as a crucial part of the host defense, but when it persists, normally in adulthood, it becomes chronic into what is known as low-grade inflammation. This basal inflammation is associated with a wide range of chronic diseases and conditions, been the combination of genomics, personal characteristics, life style and exposure, important determinants. Since highly redundant and pleiotropic risk factors are present in this conditions, we want to determine the impact of the internal (genetics, lifestyle) and external exposome (environment) on the definition of a low-grade persistent inflammation, its role in the evolution of chronic diseases, and the definition of an objective measure.

We will use the resources of the *GCAT | Genomes for life* project, a database that includes information from 20,000 adulthood general population individuals. The cohort has been mapped through a geographic information system (GIS) generating a personalized exposure map (e.g. air pollution, green spaces, etc…); longitudinal linked Electronic Health Records (EHR) (2010-2020); personal life-style habits recorded thought self-report questionnaires; and genome-wide profiles.

Firstly, a list of conditions with an inflammation component was selected based on systematic bibliographic research. Analyzing 20,000,000 imputed single nucleotide variants (SNV) of 5,000 individuals, we conducted a multiple phenotype Genome-Wide Association study (GWAS), to derive inflammatory genetic profiling. The analysis was performed using a novel machine learning method (REGENIE) that have been shown more efficient that previously standardized approaches. Post-GWAS, with significant loci were extensive functional annotated is presented here, as first step to explore the shared genetic correlation and overlapping signals driving inflammatory conditions.

IDENTIFICATION OF NOVEL CANDIDATE MICROPEPTIDES IN CANCER FROM DE NOVO TRANSCRIPTOME ASSEMBLY TO PERFORM MS ANALYSIS.

Ana Dueso-Barroso [1], Marion Martínez [2], Pilar Ximénez de Embún [3], Javier Muñoz [3], Hector Peinado [4], Maria Abad [2], David Torrents [1, 5].

1. Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain.
2. Vall d'Hebron Institute of Oncology (VHIO), Barcelona, 08035, Spain.
3. Proteomics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
4. Microenvironment and Metastasis Laboratory, Molecular Oncology Programme, Spanish National Cancer Research Center (CNIO), Madrid, Spain.
5. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain.
Presenting author: Ana Dueso-Barroso (ana.dueso@bsc.es)

Micropeptides (mp) are proteins shorter than 100 aa. They result from translating an open-reading frame (ORF) on an mRNA transcript from a coding sequence (CDS) of fewer than 300 nucleotides. These small proteins have critical roles in many biological processes. Computational and experimental approaches have been developed and implemented to infer protein-coding potential, analyze a given region's transcriptional and translational state, and detect the putative protein product generated from translation.

Together with collaborators, we aim to understand the role of micropeptides in metastatic processes in pancreatic adenocarcinoma (PACA) by using mass spectrometry (MS). A list of protein sequences is needed to compare the results from this technique. Although diverse datasets of sORFs are publicly available, most of these aa sequences have been identified from annotated human transcripts. For this reason, our goal within this project is to design a set of novel and tissue-specific candidate micropeptides using transcriptomic data from PACA samples.

We have used six STAR-aligned (hg19) paired-end samples from the PCAWG project corresponding to PACA patients. We performed a *de novo* transcriptome assembly for all these samples to identify transcribed regions and, therefore, transcripts. Results were merged, and a consensus list of 28.040 transcripts was obtained. A *in silico* 3-frames translation was done for all the coding sequences using the AUG start codon and considering the five most abundant non-canonical ones. After this, we could identify 6.366.689 aa sequences. We then filter these sequences taking into account expression levels and overlapping with annotated coding sequences, getting a list of 1.211.051 candidate micropeptides. Preliminary MS results obtained from analyzing PACA samples and compared with this set of novel candidates could detect around 69 tryptic peptides matching 234 micropeptides. At this time, more RNA-seq samples are being analyzed. Micropeptide's location and conservation and transcript expression in non-cancer samples will also be studied.

EPITOPE CONSERVATION IN VIRULENCE MOONLIGHTING PROTEIN, CANCER AND AUTOIMMUNE DISEASES

David Sánchez-Redondo[1], Luis Franco-Serrano[1], Sergio Hernández[1], Isaac Amela[1], Josep Antoni Perez-Pons[1], Jaume Piñol[1], Angel Mozo-Villarias[1], Juan Cedano[1] and Enrique Querol[1]

1. Institut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona. Cerdanyola del Vallés, Barcelona. 08193, Spain

Presenter e-mail: David Sanchez-Redondo; davsanred@uoc.edu

Moonlighting and multitasking proteins refer to proteins with two or more functions performed by a single polypeptide chain. Surprisingly, 25% of the moonlighting functions of our Multitasking Proteins Database (MultitaskProtDB-II) are related with pathogen virulence activity. Moreover, they usually have a canonical function belonging to highly conserved ancestral key functions, and their moonlighting functions are often involved in inducing Extracellular Matrix Proteins (ECM) remodeling. Often the antigenic regions of these proteins are highly conserved between pathogenic microorganisms and hosts. This can explain why evolution have selected these key enzymes to acquire new functions related to virulence, such as adhesion to host or tissue remodeling. Also, a deep analysis of the Gene Ontology codes of moonlighting proteins involved in virulence and human proteins that participate in cancer metastasis, has revealed that in both might share the same molecular mechanisms, especially those related to binding and remodeling of ECM.

DNA METHYLATION PROFILING OF HIGH RISK NEUROBLASTOMA: POTENTIAL EPIGENETIC BIOMARKERS FOR MOLECULAR RISK STRATIFICATION.

Authors: Alícia Garrido-Garcia (1), Sara Pérez-Jaume (1), Marta Garcia (1), Laura Garcia-Gerique (1), Isadora Lemos (1), Eva Rodríguez (2), Óscar Muñoz (1), Mariona Suñol (2), Helena Carén (3), Kai-Oliver Henrich (4), Frank Westermann (4), Jaume Mora (1, 5), Soledad Gómez-González (1) and Cinzia Lavarino (1).

(1) Developmental Tumor Biology Laboratory, Fundació Sant Joan de Déu, Institut de Recerca Sant Joan de Déu, Barcelona, Spain.
(2) Department of Pathology, Hospital Sant Joan de Déu, Barcelona, Spain.
(3) Sahlgrenska Cancer Center, Department of Pathology, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Sweden.
(4) Neuroblastoma Genomics, German Cancer Research Center (DKFZ), Heidelberg, Germany.
(5) Department of Pediatric Oncology, Hospital Sant Joan de Déu, Barcelona, Spain.

Email: alicia.garrido@sjd.es

Neuroblastoma (NB), the most frequent extracranial pediatric solid tumor, accounts for 15% of cancer-related deaths in children. High-risk NB (HR-NB) represents a heterogeneous group, whereby patients can display response to treatment and long-term outcome or develop early progressive, chemoresistant disease with poor outcome. To date, HR-NB patients are generally treated uniformly with no further stratification, as established in routinely used risk stratification systems. A revised molecular risk stratification has been proposed based on the analysis of telomere maintenance mechanisms, and RAS or TP53 pathway mutations. However, to date, risk-stratification of HR-NB tumors is still challenging.

Here, we investigated the genome-wide DNA methylation profile of HR-NB and identified distinct methylation patterns which defined two subgroups of patients with diverse clinical evolution. By using Cox-regression models and Machine Learning analysis, we identified a reduced set of differentially methylated CpG sites that enabled us to discriminate a subgroup of highly aggressive, chemo-refractory tumors (herein ultra high-risk; UHR-NB) within HR-NB patients. Our results were consistent with the recently proposed molecular risk stratification based on telomere maintenance mechanisms. Correlation of the identified methylation patterns with microarray gene expression using matching patient data, revealed differences in pathways related to cellular metabolism, purine biosynthesis and AKT/mTOR cell signaling. Validation of the most significant results was performed by pyrosequencing, RT-qPCR and immunohistochemistry. Our findings suggested that the heterogeneous clinical behavior of HR-NB is mediated, in part, by the metabolism of purines and the activation of oncogenic pathways such as AKT/mTOR, potentially contributing to tumor survival and progression profiles.

Taken together, we have identified a reduced set of methylation-based biomarkers that enabled stratification of patients with HR-NB tumors at the moment of diagnosis. Our findings also provide further understanding of the biology underlying the heterogeneous behavior of HR-NB, and reveal altered biological pathways of interest for potential therapeutic options.

Understanding the link between mutant NPM1 and HOX gene expression in AML

Maria Cadefau[1,2], Lucía Lorenzi[1] and Sergi Cuartero[1,2]

1. Josep Carreras Leukaemia Research Institute (IJC), Campus Can Ruti, 08916 Badalona, Spain
2. Germans Trias i Pujol Research Institute (IGTP), Campus Can Ruti, 08916 Badalona, Spain
Presenter email: llorenzi@carrerasresearch.org

Nucleophosmin 1 (NPM1) is a nucleolar protein involved in several cellular functions, including ribosome biogenesis, chromatin remodeling and genome stability. The gene encoding NPM1 is one of the most frequently mutated genes in AML. NPM1 mutations consist of small insertions resulting in the generation of a nuclear export signal (NES) that leads to an aberrant cytoplasmic localization. A direct correlation between the cytoplasmic mutant isoform (NPM1c) and HOX gene expression has been described; however, the molecular mechanisms underlying this relationship remain unclear. To understand the link between NPM1 and HOX gene expression, we have examined the chromatin landscape of AML cells with and without NPM1c. For this, we have used an acute degradation system in AML cells with NPM1 mutations that allows the rapid degradation of the mutant isoform within hours of addition of dTag13. To identify the genomic regulatory elements that are dependent on the cytoplasmic localization of NPM1, we have performed Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) at 15 minutes and 2 hours of dTag13 treatment. As the differences in ATAC-seq profiles between samples are small, we have compared different strategies to call peaks and to identify differentially accessible peaks. This has led to the identification of a set of regions with decreased accessibility after NPM1 degradation. These peaks overlap genes in the HOX locus and might represent direct targets of NPM1c. To understand the functional role of the identified regions, we are investigating the transcription factors (TF) that bind them. For this, we are performing motif discovery analysis and studying the overlap with TF binding peaks from multiple public ChIP-seq datasets. With these approaches, we aim to uncover the regulatory mechanisms that link NPM1c and HOX expression. Importantly, this knowledge may possibly reveal new therapeutic angles for treatment of AML patients carrying this common NPM1 mutation.

Mathematical modeling of strigolactone production in maize

Abstract

Maize (Zea mays) is a food staple to the world's poorest regions in Africa, Asia, and Latin America, and there is an expected surge in demand for the next decade. Though, it remains vulnerable to drought and parasites, which can cause up to 50% and 100% yield losses, respectively. It was elucidated that the plant hormone strigolactone (SL) is responsible for germinating these dormant *Striga* seeds, which is vital in preventing this host-parasite interaction. Aside from that, SL triggers symbiosis with arbuscular mycorrhizal (AM) fungi which helps the plant to gain resistance to environmental stresses. However, little knowledge is known about SL production, especially the enzymes that mediate each biosynthetic step. Hence the need to build a mathematical model of the biosynthetic pathway of SL that allows *in silico* analysis. We proposed models that most likely describe two alternative pathways of SL and interrogated these models to determine the differences in dynamic behavior between these two models. This approach initiates a different perspective in wet-lab experiments and sets the stage for future experiments to elucidate enzymes responsible for SL biosynthesis.

genoMatriXeR: a R package for association analysis and visualization of genomic multiple regions based on permutation test

Roberto Malinverni [1], Marcus Buschbeck [1,2]

[1] Cancer and Leukemia Epigenetics and Biology Program, Josep Carreras Leukaemia Research Institute (IJC), Campus ICO-GTP-UAB, Badalona, Spain

[2] Program for Predictive and Personalized Medicine of Cancer, Germans Trias i Pujol Research Institute (PMPPC-IGTP), Badalona, Spain

Presenter e-mali: rmalinverni@carrerasresearch.org

The meaningful interpretation of overlaps between binding profiles of multiple chromatin regulators is a major challenge in epigenomics. To address this, we published on BioconductoR in 2015 the package regioneR that we developed for statistically assessing the association between genomic regions sets. These regions sets can be those gained from ChIP-seq and Cut&Run peak calling algorithms or any other annotation of the genome. Until now regioneR has been cited in about 250 publications. The permutation framework applied in regioneR package randomizes genomic data taking into account the complexity of the genome such as masked areas and chromosome structure. RegioneR has two technical limitations. First, only a limited number of regions sets can be simultaneously computed. Second and most importantly, different pair-wise associations cannot be directly compared as the z-score scales with the size and other characteristics of the individual regions sets.
Here, we now present the genoMatriXeR, an R package that is the natural evolution of regioneR and allows to calculate the statistical association between multiple regions sets at the same time. As its predecessor, its randomization-based approaches implicitly take into account the complexity of the genome without the need to assume an underlying statistical model. In addition, genoMatriXeR is also designed to work with multiple regions sets associations at the same time. To compare z-scores coming from multiples analysis, different strategies have been introduced to normalize the z-score and to improve the p-value calculations. Furthermore, we have given great importance to the visualization of the data. Different clustering approaches can be applied to the output matrices in order to easily extract the most relevant results of the analyzes. genoMatriXeR is designed to work with associations complex region sets, implementing a series of randomization and evaluation strategies that address the most common use cases leaving the possibility to completely customize them. Taken together, genoMatriXeR aims to be a novel and precious addition to NGS tools and for whole genome analysis.

RAD21L depletion alters the 3D genome architecture in the male germ line

Laia Marín-Gual[1,2], Covadonga Vara[1,2], Natalia Felipe-Medina[3], Yasmina Cuartero[4,5], Lucía Álvarez[1,2], François Le Dily[4,5], Francisca Garcia[6], Elena Llano[3], Marc A. Marti-Renom[4,5,7,8], Alberto M. Pendás[3], Aurora Ruiz-Herrera[1,2]

[1]Departament de Biologia Cel·lular, Fisiologia i Immunologia, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, 08193, Spain. [2]Genome Integrity and Instability Group, Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, 08193, Spain.[3]Molecular Mechanisms Program, Centro de Investigación del Cáncer and Instituto de Biología Molecular y Celular del Cáncer (CSIC-Universidad de Salamanca), Salamanca, Spain.[4]Centre for Genomic Regulation, The Barcelona Institute for Science and Technology, Carrer del Doctor Aiguader 88, Barcelona, 08003, Spain. [5]CNAG-CRG, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Baldiri Reixac 4, Barcelona, 08028, Spain. [6]Unitat de Cultius Cel.lulars, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, 08193, Spain. [7]Pompeu Fabra University, Doctor Aiguader 88, Barcelona, 08003, Spain. [8]ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

During meiotic prophase I, cohesin complexes participate in synapsis and recombination of homologous chromosomes by keeping the two sister chromatids together and holding chromatin loops to the synaptonemal complex (SC). RAD21L is a meiotic-specific cohesin subunit essential for synapsis and male fertility but its implications on the spatial folding of chromosomes during meiosis remain unclear. Here, we study the impact of RAD21L depletion on the 3D genome architecture in the male germ line by combining fluorescence activated cell sorting (FACS) and the chromosome conformation capture technique (Hi-C). We demonstrate that the loss of RAD21L prevents proper chromatin condensation during meiosis, with changes in the inter- and intra-chromosomal interactions ratio and the A/B compartmentalization in pre-meiotic (spermatogonia) and meiotic (primary spermatocytes) cells. We detected defects in the bouquet formation and an increase in telomeric interactions between heterologous chromosomes in primary spermatocytes, resembling telomere aberrations detected in other cohesin deficient models. Overall, our results show how the three-dimensional genomic structure is affected in the absence of the meiotic cohesin subunit RAD21L during mouse spermatogenesis.

Gene expression profiles of visceral and subcutaneous adipose tissues in children with overweight or obesity: the KIDADIPOSEQ project

Mireia Bustos-Aibar[1,2], Augusto Anguita-Ruiz[1,2,3,4,5], Álvaro Torres-Martos[1,2], Jesús Alcalá-Fdez[6], Francisco Javier Ruiz-Ojeda[1,2,3,4], Marjorie Reyes Farias[7,9], Andrea Soria-Gondek[10], Laura Herrero[4,7], David Sánchez-Infantes[4,8,9], Concepción María Aguilera[1,2,3,4]

1. Department of Biochemistry and Molecular Biology II, School of Pharmacy, University of Granada, 18071 Granada, Spain. E-mail: mireiabustos@correo.ugr.es (M.B.-A); alvarotorres@correo.ugr.es (A.T.-M.); fjrojeda@gmail.com (F.J.R.-O.); caguiler@ugr.es (C.M.A.)
2. Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, Avda. del Conocimiento s/n., 18016 Granada, Spain
3. Instituto de Investigación Biosanitaria IBS.GRANADA, Complejo Hospitalario Universitario de Granada, 18014 Granada, Spain
4. CIBEROBN (CIBER Physiopathology of Obesity and Nutrition), Instituto de Salud Carlos III, 28029 Madrid, Spain
5. Barcelona Institute for Global Health (ISGlobal), Doctor Aiguader 88, 08003 Barcelona, Spain. E-mail: augusto.anguita@isglobal.org (A.A.-R.)
6. Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071 Granada, Spain. E-mail: jalcala@decsai.ugr.es (J.A.-F.)
7. Department of Biochemistry and Physiology, School of Pharmacy and Food Sciences, Institute of Biomedicine of the University of Barcelona (IBUB), University of Barcelona, E-08028 Barcelona, Spain. E-mail: marjorie.reyesfarias@ub.edu (M.R.F.); lherrero@ub.edu (L.H.)
8. Department of Health Sciences, Campus Alcorcón, University Rey Juan Carlos (URJC), E-28922 Madrid, Spain
9. Department of Endocrinology and Nutrition, Germans Trias i Pujol Research Institute, 08916, Badalona, Spain. E-mail: dsanchez@igtp.cat (D.S.-I.)
10. Pediatric Surgery Department, Hospital Universitari Germans Trias i Pujol, 08916, Badalona, Spain. E-mail: andreasoriagondek@gmail.com (A.S.-G.)

Abstract

Childhood obesity is a multifactorial disease influencing the development of a range of metabolic disorders, where adipose tissue has been proved to be fundamental. The adipose tissue can be distributed throughout the body as visceral adipose tissue (VAT)

and subcutaneous adipose tissue (SAT) and there are considerable anatomical differences between both adipose tissues in the body. Importantly, VAT is associated with low-grade systemic inflammation and insulin resistance, which are key factors underlying metabolic alterations associated with childhood obesity. This study aimed to identify the molecular signatures underlying obesity and overweight in children, differentiating between shared and individual signatures in VAT and SAT. Both tissue samples were collected from 18 children (11 girls) aged 0.54 to 16.63 years and hospitalized for abdominal surgery, of which only 6 children (2 girls) had overweight or obesity. RNAseq analysis was performed to identify gene expression patterns associated with obesity and overweight in each tissue. In VAT there were 759 genes showing statistically significant differential expression between groups (nominal p-value < 0.05), from which 48 passed an FDR threshold of 0.05. In SAT there were 945 genes showing statistically significant differential expression, from which 28 passed the FDR threshold. Among significantly associated genes, there were 126 common genes, associated with overweight and obesity in both tissues. Results were validated with genes whose influence on obesity has been previously described (*e.g., LEP*) and we identified new molecular targets for the pathology, highlighting the results of VAT (*i.e., XIST, PRKY* and *TTTY10*). Therefore, our approach identified independent and common gene expression patters in VAT and SAT associated with overweight and obesity in children. Understanding the molecular architecture of obesity with approaches like these is crucial for the identification of powerful molecular targets and developing of effective precision medicine therapies.

The Bioinformatics Unit at IJC provides both IJC and external researchers with high-quality computational analysis services covering all project aspects related to clinical and biological data. This includes experimental design and data analysis for microarray experiments and Next Generation Sequencing, statistical consulting, data integration, interpretation and reporting, as well as software development and scientific database management.

Differential Regulation of Insulin Signaling by Monomeric and Oligomeric Amyloid Beta-Peptide

Rubén Molina-Fernández,[1,†] Pol Picón-Pagès,[2,†] Alejandro Barranco-Almohalla,[2] Giulia Crepin,[2] Víctor Herrera-Fernández,[2] Anna García-Elías,[2] Hugo Fanlo-Ucar,[2] Xavier Fernàndez-Busquets, [3,4,5] Jordi García-Ojalvo,[6] Baldomero Oliva,[1,*] and Francisco J. Muñoz[2,*]

[†]**These authors contributed equally to this work.**

*** Co-corresponding senior authors.**

Abstract

Alzheimer's disease and diabetes type 2 (T2D) are pathological processes associated to aging. Moreover, there are evidence supporting a mechanistic link between Alzheimer's disease and insulin resistance (one of the first hallmarks of T2D). Regarding Alzheimer's disease, amyloid β-peptide (Aβ) aggregation into β-sheets is the main hallmark of Alzheimer's disease. At monomeric state (mAβ) is not toxic but its function in brain, if any, is unknown.

Here we show, by in-silico study, that mAβ1−40 shares the tertiary structure with insulin and is thereby able to bind and activate insulin receptor (IR). We validated this prediction experimentally by treating human neuroblastoma cells with increasing concentrations of mAβ. Our results confirm that Aβ1−40 activates IR autophosphorylation,triggering downstream Akt phosphorylation and GLUT4 translocation to the membrane.

On the other hand, neuronal insulin resistance is known to be associated to Alzheimer's disease since early stages. We thus modelled the docking of oligomeric Aβ1−40

(oAβ1−40) to IR. We found that oAβ1−40 blocks IR, impairing its activation. It was confirmed in vitro observing the lack of IR autophosphorylation, and also the impairment of insulin-induced Akt activation and GLUT4 translocation to the membrane. By biological system analysis, we have carried out a mathematical model recapitulating the process that turns Aβ-IR binding from the physiological to the pathophysiological regime. Our results suggest that mAβ1−40 contributes to mimic insulin effects in the brain, which could be good when neurons have an extra requirement of energy beside the well-known protective effects on insulin intracellular signaling, while its accumulation and subsequent oligomerization blocks the IR producing insulin resistance and compromising neuronal metabolism and protective pathways.

SEMI-AUTOMATION OF VARIANT CLASSIFICATION IN HEREDITARY CANCER

Elisabet Munté, Laura Arnaldo, Marta Pineda, Conxi Lázaro, Lidia Feliubadaló

INTRODUCTION

The use of new generation sequencing has increased the detection of pathogenic variants but also the detection of variants of unknown significance. Variant classification represents a huge challenge and the main bottleneck in daily diagnostics practice, but only a correct classification allows proper genetic counselling and personalized risk estimation. Classifying a variant is a manual complex and long process that combines information of distinct nature, such as type of variant, population frequencies, in silico predictors, etc. and must follow published updated guidelines. Moreover, it is also an iterative task because the appearance of new information enforces the periodic revision of variant classification.

AIM: The aim of the project is to automatize as much as possible the process of variant classification to streamline the work of biologists and avoid possible manual error.

METHODS:

1. Analysis of the different DBs used in diagnostics.
2. Design and implementation of an R pipeline that integrates information from Mutalyzer, Ensembl-VEP, ClinVar, gnomAD and in silico predictors with a gene-specific decision tree.
3. Validation with 326 manually classified variants.

RESULTS

An alpha version of the tool was developed and demonstrate its capacity to classify a variant in 30 to 100 seconds, which will save a lot of time to users. It implements the updated international general guidelines plus the gene-specific guidelines for *BRCA1, BRCA2, ATM, MLH1, MSH2, MSH6* and *PALB2*. It automates criteria related to mutation type, population frequencies and in *silico* predictors, and provides information from clinical databases to help calculate the remaining criteria and it returns an excel file. As for the validation, the tool matched with manual classified variants in 94% of the cases. Discrepancies were due to exceptions added by users while classifying, change of cut-offs or predictors used, to criterion that are not fully automated or to manual error. For the latter, 40 criteria were corrected demonstrating we have accomplished one of our main goals.

DISCUSSION

Nevertheless, of course, manual curation is necessary to check criteria that cannot be automated or can only be partially automated and any guides exception that could lead to a change of the variant's final classification.

Improving drug-induced transcriptional descriptors and their biological connectivity

Elena Pareja-Lorente[1] , Adrià Fernández-Torras[1] and Patrick Aloy [1,2,*]

[1] Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
[2] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
[*] To whom correspondence should be addressed.

Abstract
Compound bioactivity signatures allow the description of small molecules according to the biological effects that they exert, providing complementary opportunities to current drug discovery strategies. The Chemical Checker (CC) resource provides a rich collection of bioactivity signatures, including drug-induced transcriptional changes, enabling the assessment of functional similarities between small molecules. Here, we present a novel characterization of small molecules based on the gene expression changes that they induce in different cell lines. By recomputing the gene expression signatures of the compounds and filtering the unrobust ones, we show how our bioactivity signatures better characterise and preserve the biological coherence of the raw data, improving those in the Chemical Checker repository. Finally, we train a neural network to integrate any novel differential gene expression experiment with the corpus of available drug-induced gene expression signatures, connecting biology and chemistry.

Title: Automating structural information retrieval for Integrative Modelling of cellular complexes

Ferran Pegenaute[1], Baldo Oliva[1], and Oriol Gallego[1]

[1] Department of Experimental and Health Science (DCEXS), Universitat Pompeu Fabra (UPF), Barcelona, 08003, Spain.

Abstract

The exocytic machinery is essential for all eukaryotic cells. However, its complexity and dynamism prevents its functional structural characterization. Recently, the development of integrative approaches combining different sources of structural information enabled cellular structural biologists to study complex and dynamic molecular systems, achieving high-resolution observations in a near-physiological context. We are now developing a tool to retrieve and generate structural data to input into the Integrative Modelling Platform (IMP) for modeling the exocytic machinery. More specifically, for a protein known to interact with the complex of interest, it performs a search of its sequence on the Protein Data Bank to look for already experimentally solved structures. Then, uses Deep Learning (DL) tools to infer a structure prediction, and uses the confidence scores of the output to provide information about potentially rigid and flexible regions for IMP to model. This will allow for an additional type of data to contribute to the models, apart from the *in vitro* (cryo-EM) and *in situ* (live-cell imaging) information being used at the moment in the hosting lab, enhancing the functional structural analysis of the exocytosis machinery.

Noncoding regulatory functions in β-cell derived neuroendocrine tumors

Ramos-Rodríguez M[1], Norris R[1], Subirana-Granés M[1], Sordi V[2], Raurell-Vila H[1], Beatriz Pérez-González[1], Pellegrini S[2], Falconi M[3], Piemonti L[2], Pasquali L[1]

[1] Endocrine Regulatory Genomics, Department of Experimental & Health Sciences, University Pompeu Fabra, 08003, Barcelona, Spain.
[2] Diabetes Research Institute (DRI) - IRCCS San Raffaele Scientific Institute, Milan, Italy.
[3] Chirurgia del Pancreas, Ospedale San Raffaele IRCCS, Università Vita e Salute, Milano, Italy.

Insulinomas are rare neuroendocrine tumours that arise from pancreatic β-cells. While retaining the ability to produce insulin, insulinomas feature aberrant proliferation and altered hormone secretion resulting in failure to maintain glucose homeostasis.

The role of *cis*-elements and their aberrations to the development of these tumors is currently unexplored. We have now generated insulinoma regulatory maps by profiling gene expression and H3K27ac deposition in a large set of human pancreatic neuroendocrine tumors. We observed widespread chromatin activation of ~8,500 H3K27ac enriched sites in the tumoral tissue but not in untransformed human pancreatic islets. These regions are mostly distal to gene TSS, evolutionarily conserved and contribute to the transcriptomic aberrations of nearby genes. Of note, we show that ~20% of the differential H3K27ac activated regions are H3K27me3 repressed in unaffected β-cells (mean z-score 27.96), suggesting that tumoral transition is coupled with derepression of β-cell polycomb targeted domains. By coupling epigenetic profiling with whole genome sequencing we now aim in uncovering genetic variation implicated in deregulation of noncoding functions

Our epigenomic profiling provides a compendium of aberrant *cis*-regulatory elements that alter β-cell function and fate in their progression to pancreatic neuroendocrine tumors and a framework to identify coding and noncoding driver mutations.

For decades, chemoinformatic methods have employed vector descriptors of compound structure as an input for prediction tasks. However, from a biochemical perspective, chemical similarity is not always the best metric to compare a set of molecules or drugs. That is due to the fact that two different molecules can have similar bioactivities despite being chemically dissimilar, and also two very chemically similar molecules can exhibit different bioactivity profiles. That is why novel approaches are arising in order to generate molecular descriptors including other information apart from chemical properties, such as information about protein binding, clinical outcomes of drugs or biological processes in which molecules are involved [Figure 1, bottom].

In the same way, biological data is being accumulated at an unprecedented rate, and different efforts have been made in order to integrate and harmonize all this information in the form of knowledge graphs. This kind of network representation of biological knowledge can be further exploited thanks to network embedding algorithms that allow us to represent the biological entities from the graphs (e.g., genes, cells) as numerical vectors that encode their relationships (e.g. “is associated to”, “has mutation”) with other entities [Figure 1 top].
These types of bioactivity and biological descriptors are being used in a wide variety of prediction tasks, with impressive results.

Further than that, there has been a rising interest in generative models, which consist of deep learning models that are able to learn from different types of compound representations, such as SMILES or molecular graphs, and then generate novel, potentially synthesizable structures. These kinds of deep learning models have demonstrated their potential in the generation of new chemical entities, both for exploration of the chemical space and for the design of molecules with desired properties [Figure 2].

Our aim is to combine these 3 tools (bioactivity descriptors, biological embeddings and generative models) in order to generate new chemical entities with desired bioactivities.

Here we present a case study in which we used property predictors based on bioactivity and biological descriptors based in genomics to bias the generation of new compounds towards potential desired clinical outcomes.

Figure 1

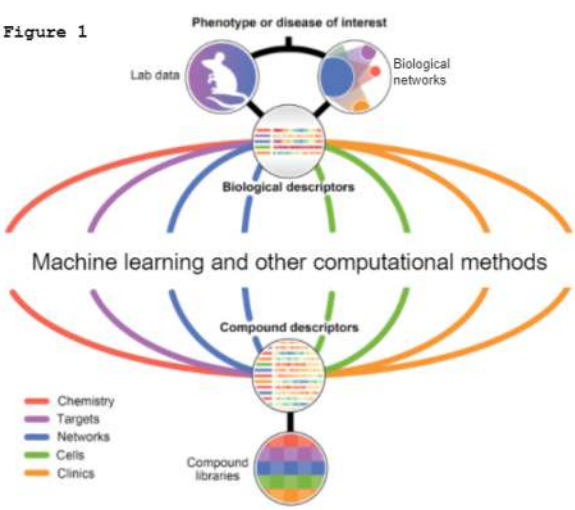
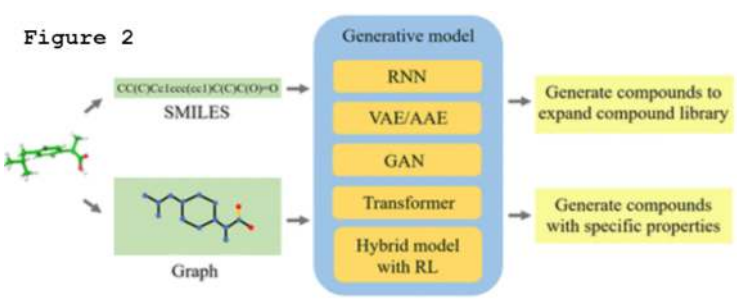


Figure 2



Tong, X., Liu, X., Tan, X., Li, X., Jiang, J., & Xiong, Z. et al. (2021). Generative Models for De Novo Drug Design. *Journal Of Medicinal Chemistry*

Long read nanopore RNA sequencing improves planarian transcriptome annotation and differential gene expression analysis approaches.

Maria Rosselló[1,2], Teresa Adell[1,2], Emili Saló[1,2], Josep F Abril[1,2]

1. Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia (UB), Avinguda diagonal 643, 08028, Barcelona, Catalonia, Spain.
2. Computational Genomics Lab; Institut de Biomedicina (IBUB), Universitat de Barcelona, Barcelona, Catalonia, Spain.
Presenter email: mariarossello@ub.edu

Planarians have become a model organisms in regeneration and adult tissue renewal research. Despite its relevance in that field, *Schmidtea mediterranea* genomic and transcriptomic diversity is not completely understood yet. In this work we propose a new approach to obtain full length transcripts using a single-molecule long-read sequencing method based on Oxford Nanopore technologies (ONT) for the first time in this species. Using nanopore sequencing data we could successfully assemble a transcriptome and describe some alternative splicing isoform events. We further implemented some improvements on the computational protocol to correct the errors derived from the sequencing methodology to retrieve the proper ORFs translations. From the multiplexed samples of our sequencing run, we demonstrated that data derived from nanopore sequencing can be applied to differential gene expression (DGE) analyses too. The results from the comparison among different physiological states in the planarian led us to undercover key genetic pathways, which play a role on the molecular mechanisms that control regeneration and body size.

In conclusion, in this work we established that single-molecule long-read sequencing can be a reliable method for *de novo* transcriptome assembly as well as DGE analyses in *S. mediterranea*.

SUCCESSES AND CHALLENGES IN MULTISCALE MODELLING OF ARTIFICIAL METALLOENZYMES: THE CASE STUDY OF POP-RH_2 CYCLOPROPANASE

José-Emilio Sánchez-Aparicio[1], Giuseppe Sciortino[1], Eric Mates-Torres[1], Agustí Lledós[1], Jean-Didier Maréchal[1]

1. Insilichem, Department of Chemistry, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain.
Presenter e-mail: joseemilio.sanchez@uab.cat

Molecular modelling applications in metalloenzyme design are still scarce due to a series of challenges. In top of that, the simulations of metal-mediated binding and the identification of catalytic competent geometries requiring both large conformational exploration and simulation of fine electronic properties. Here, we demonstrate how the incorporation of new tools in multiscale strategies, namely substrate diffusion exploration, allow taking a step further. As a showcase the enantioselective profiles of the most outstanding variants of an artificial Rh_2-based cyclopropanase (GSH, HFF and RFY) developed by Lewis and co-workers have been rationalized. DFT calculations on the free-cofactor-mediated process identify the carbene insertion and the cyclopropanoid formation as crucial events, being the latter the enantiodetermining step, which displays up to 8 competitive orientations easily altered by the protein environment. The key intermediates of the reaction were docked into the protein scaffold showing that some mutated residues have direct interaction with the cofactor and/or the co-substrate. These interactions take form of a direct coordination of Rh in GSH and HFF and a strong hydrophobic patch with the carbene moiety in RFY. Posterior molecular dynamics sustain that the cofactor induces global re-arrangements of the protein. Finally, massive exploration of substrate diffusion, based on the GPathFinder approach, defines this event as the origin of the enantioselectivity in GSH and RFY. For HFF, fine molecular dockings suggest that it is likely related to local interactions upon diffusion. This work shows how modelling of long-range mutations on the catalytic profiles of metalloenzymes may be unavoidable and software simulating substrate diffusion should be applied.

Insulinoma genetic and epigenetic profiling

Marc Subirana-Granés[1], Mireia Ramos-Rodríguez[1], Lorenzo Pasquali[1]

1. Endocrine Regulatory Genomics, Department of Experimental & Health Sciences, University Pompeu Fabra, 08003, Barcelona, Spain.

Neuroendocrine tumors are neoplasms arising from cells that have hormone-producing endocrine cells and nerve cells traits. The main type of functional PNETs is insulinoma, which is derived from pancreatic β-cells and secrete insulin independent of glucose and cause hypoglycemia. These neoplasms are very rare with an incidence in the general population of approximately one to four per million per year. Moreover, the major genetic and epigenetic alterations in sporadic insulinomas are still unknown. The non-coding genome plays a critical role in regulating gene expression and in maintaining a cell phenotype. Thus, non-coding landscape can elucidate the genetic basis of insulinoma tumors, which remains unexplained based on protein-coding driver genes. Understanding this insulinoma malignant transformation can shed light on inducing human beta cells to regenerate and diabetes treatment.

Here we defined the genetic insulinoma profile by analysing somatic whole genome mutations. We established for the first time the mutational burden of insulinomas and established the mutational signatures that drive insulinomas. We also established the major coding drivers of this cohort. In order to unmask driver non-coding somatic mutations in insulinomas, we took advantage of H3K27ac ChIP-seq assays. We reconstructed the cis-regulatory landscape in insulinoma and untransformed human pancreatic islets. We interrogated this cis-regulatory landscape implementing oncodriveCLUSTL approach on our dataset of somatic mutations in insulinoma. This approach unraveled a set of genomic regions bearing a significant clustering of variants. These mutated genomic regions may provide insight into the mechanisms of malignant insulinoma transformation defining novel driver mutations and disrupted pathways.

These findings can elucidate the impact of somatic genomic variants to the loss of β-cell fate and the development of insulinomas, identifying the contribution of coding and non-coding variants in this transition.

Low input promoter capture Hi-C method enables to decipher the molecular mechanisms underlying genetically complex diseases

Laureano Tomás-Daza[1,2], Paula López-Marti [1,2], Alfonso Valencia [2], Biola M. Javierre [1]

1. Josep Carreras Leukaemia Research Institute (IJC)
2. Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS)

Presenter e-mail: ltomas@carrerasresearch.org

3D chromatin organization is key for the transcriptional regulation of the cell, bringing in physical proximity distal regulatory regions with their target genes, and its alteration promotes disease. To study in detail the promoter interactome of the cells, we implemented promoter capture Hi-C (PCHi-C) method. It allows systematic identification of genomic regions in proximity with more than 22,000 gene promoters, independently of the activity status of both regions. Since PCHi-C just relies on sequence capture technology but not in antibody immunoprecipitation as other methods do, it allows robust comparison between conditions or cell types, and the customization to interrogate any specific interactome. Using this method, we and others have associated non-coding variants to their distal target promoters, identifying hundreds of potential new disease-candidate genes and/or pathways. However, PCHi-C, together with other widely used methods, typically requires from 20 to 50 million cells per biological replicate, which prohibits the analysis of rare cell populations such as those commonly relevant in clinical settings.

Here, to circumvent this shortcoming, we present low input promoter capture Hi-C (liCHi-C) method that allows the generation of high-quality genome-wide promoter interactomes using very low amounts of cells. We validate our new method by comparing promoter interactomes for a controlled cell titration of primary human cells against the highest resolution PCHi-C datasets until the date, demonstrating that the interactomes are robust down to 50,000 cells of starting material. In addition, this method captures the cell type and lineage specificity of the genome organization, similarly as PCHi-C maps produced with 40 million cells. Besides, we demonstrate the applicability of low input PCHi-C to associate disease-related non-coding alterations with potential target genes and gene pathways in sparse primary cell types. Finally, using blasts from acute lymphoblastic leukemia (ALL) from pediatric biopsies we show the power of low input PCHi-C to diagnose patient specific chromosomal rearrangements and identify cancer-specific topological features.

Use of Machine learning methods to improve the early-life predictive programs of obesity childhood

Álvaro Torres-Martos[1], Augusto Anguita-Ruiz[1,2,3], Mireia Bustos-Aibar[1], Rafael Alcalá[4], Concepción M. Aguilera[1,3] and Jesús Alcalá-Fdez[4]

1. Department of Biochemistry and Molecular Biology II, Institute of Nutrition and Food Technology "José Mataix", Center of Biomedical Research, University of Granada, Avda. del Conocimiento s/n. Armilla, 18016 Granada, Spain; alvarotorres@correo.ugr.es (A.T.-M); augusto.anguita@isglobal.org (A.A.-R.); mireiabustos@correo.ugr.es (M,B,-A); caguiler@ugr.es (C.M.A.)

2. ISGlobal, Doctor Aiguader 88, 08003 Barcelona, Spain; augusto.anguita@isglobal.org (A.A.-R.)

3. CIBEROBN (Physiopathology of Obesity and Nutrition Network CB12/03/30038), Institute of Health Carlos III (ISCIII), 28029 Madrid, Spain; augusto.anguita@isglobal.org (A.A.-R.); caguiler@ugr.es (C.M.A.)

4. Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071 Granada, Spain; alcala@decsai.ugr.es (R.A.); jalcala@decsai.ugr.es (J.A.-F)

Overweight and obesity in children are important risk factors for a number of chronic cardiometabolic alterations during adulthood, which considerably increase population morbimortality. To address this problem effectively, it is an urgent need to implement early-life predictive programs able to deal with the problem from the origin, when there is still time for clinical actions. Insulin resistance is one of the metabolic comorbidities of obesity that shows an earliest appearance in life, and therefore, it has become a cornerstone in preventing obesity-associated comorbidities. In the present work, we use several machine learning algorithms for the construction of predictive models of pubertal insulin resistance in children with obesity. For that purpose, we employ information from the prepubertal stage of children (ages 4-12) consisting of three layers of data: Anthropometry, cardiometabolic and inflammatory biomarkers; genetics variants; and DNA methylation measures. Among the machine learning models tested, we have selected a battery of tree-based algorithms that provide models with high predictive capability, such as Random Forest, eXtreme Gradient Boost, etc. The best model was provided by the algorithm Random Forest whose accuracy, sensibility, specificity and kappa coefficient was of 0.79, 0.84, 0.73 and 0.54, respectively. Interestingly, it is pointed out that the Adiponectin and leptin ratio could be a great biomarker to predict insulin resistance. Moreover, it has been found that methylation patterns in *CTBP2*, *HDAC4*, *PTPRN2*, *RASGRF1* and *TMEM30C* genes could be useful for the prediction of insulin resistance. The use of these predictive tools from early ages could improve the healthcare and knowledge of obesity children who have a high risk to develop cardiometabolic alterations during adulthood.

DETECTION OF ONCOGENIC AND CLINICALLY ACTIONABLE MUTATIONS IN CANCER GENOMES CRITICALLY DEPENDS ON VARIANT CALLING TOOLS

Carlos A. García-Prieto[1,2], Francisco Martínez-Jiménez[3], Alfonso Valencia[2,4*], Eduard Porta-Pardo[1,2*]

1.- Josep Carreras Leukaemia Research Institute (IJC), Badalona, Spain
2.- Barcelona Supercomputing Center (BSC), Barcelona, Spain
3.- Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain.
4.- Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
Presenter e-mail: cgarcia@carrerasresearch.org

The analysis of cancer genomes is critical to understand its aetiology, oncogenesis and potential treatments. While researchers and clinicians are often only interested in the identification of oncogenic mutations or actionable variants, the first crucial step in the analysis of any tumor genome is the identification of somatic variants in cancer cells (i.e. those that have been acquired during their evolution). While there have been some efforts to benchmark somatic variant calling tools and strategies, the extent to which variant calling decisions impact the results of downstream analyses of tumor genomes remains unknown.

Here we quantify the impact of variant calling decisions by comparing the results obtained in three important analyses of cancer genomics data (identification of cancer driver genes, quantification of mutational signatures and detection of clinically actionable variants) when changing the somatic variant caller (MuSE, MuTect2, SomaticSniper, VarScan2) or the strategy to combine them (Consensus of two, Consensus of three and Union). Our results show that variant calling decisions have a significant impact on these analyses, creating important differences that could even impact treatment decisions for some patients. Moreover, the Consensus of three calling strategy to combine the output of multiple variant calling tools, a very widely used strategy by the research community, can lead to the loss of some cancer driver genes and actionable mutations. Overall, our results point to important differences in critical analyses of tumor sequencing data depending on variant calling and highlight the limitations of widespread practices within the cancer genomics community.