XIV Symposium on Bioinformatics
Granada 14th-16th November
JBI 2018

ORGANIZATION

Organizing Committees

Chairs

Ignacio Rojas-Ruiz
Pedro Carmona-Sáez
Salvador Capella
Alfonso Valencia

Local Committee

Coral del Val
Michael Hackenberg
Fuencisla Matesanz
Eduardo León

Scientific Committee

Patrick Aloy
Fátima Al-Shahrour Núñez
Eduardo Andrés León
Sergi Beltran
Mattia Bosio
Mónica Campillos
Salvador Capella-Gutiérrez
Jose María Carazo
Pedro Carmona-Sáez
Uciel Chorostecki
Ana Conesa
Marta Coronado-Zamora
Piotr Wojtek Dabrowski
Joaquín Dopazo
Josep Ll Gelpí
Dietlind Gerloff
Gonzalo Gómez
Juergen Haas
Jaime Huerta-Cepas
Alba Jene
Marina Marcet Houben
Pier Luigi Martelli
Mariana Neves
Cedric Notredame
Cinta Pegueroles
María Peña-Chilet
Javier Pérez Florido
Jordi Rambla
Daniel Rico
Juan Rodriguez-Rivas
Ana María Rojas
Ignacio Rojas
Marianne Rooman
Ferran Sanz
Joan Segura
Sonia Tarazona
Silvio Tosatto
Oswaldo Trelles
Michael Tress

JBI2018 Short Program

Wednesday, 14th/Nov		
09.00 - 10:00	REGISTRATION	
10.00 - 10.30	Coffee Break	
10.30 - 11.30	**Opening Keynote** **Julio Sáez-Rodríguez** **"Computational personalised medicine in cancer: from in vitro to in vivo"**	
11.30 - 12.10	**Highlights Session #1**	**Genome Denmark: Sequencing and de novo assembly of the Danish reference genome** José Mg Izarzugaza, Søren Brunak and Danish Pangenome Consortium. Genome Denmark
12.10 - 12.50		**Scutoids, a geometrical solution to three-dimensional packing of epithelia** Luis M Escudero
12.50 - 13.00	COMPANIES	
13.00 - 14.30	Lunch Break & Posters (odd submission numbers)	
14.30 - 14.50	**NGS Applications**	**Precise sample classification of the MetaSUB environmental microbiota using functional biomarkers** Carlos S. Casimiro-Soriguer, Javier Pérez Florido, Daniel López López Carlos Loucera and Joaquín Dopazo
14.50 - 15.10		**Identification of RNA-processing prognostic and therapeutic markers in MLL-rearranged infant acute leukemia** Adria Closa, Antonio Agraz-Doblas, Ignacio Varela, Pablo Menéndez and Eduardo Eyras
15.10 - 15.30		**Implementation of a Copy Number Variant detection workflow within the URD-Cat pilot project on Personalised Medicine** Gemma Bullich, Steven Laurie, Jordi Morata, David Ovelleiro, Sandra Redó, Ricky Joshi, Cristina Luengo, Leslie Matalonga, Genís Parra, Raul Tonda and Sergi Beltran
15.30 - 15.50		**Cohesin variants SA1 and SA2 play distinct roles in chromatin organization and gene expression required for embryonic stem cell identity** Daniel Gimenez, Aleksandar Kojic, Marc A. Marti-Renom, François Le Dily, Ana Cuadrado and Ana Losada
15.50 - 16.10		**Improving RNA-Seq germline variant calling within RD-connect consortium** Mattia Bosio, Alfonso Valencia and Salvador Capella-Gutiérrez
16.10 - 16.30		**The regulatory genome of drosophila regeneration** Cecilia Coimbra Klein, Elena Vizcaya-Molina, Florenci Serras, Roderic Guigó and Montserrat Corominas
16.30 - 17.00	Coffee Break & Posters (odd submission numbers)	
17.00 - 17.20	**Integrative Bioinformatics**	**Proteome-wide discovery of degrons and analysis of their role in tumorigenesis across cancer types** Francisco Martínez-Jiménez, Abel Gonzalez-Perez and Núria López-Bigas
17.20 - 17.40		**Analysis of Transcription Factor activity patterns in Systemic Lupus Erythematosus** Raúl López-Domínguez, Daniel Toro-Domínguez, Jordi Martorell-Marugan, Christian Holland, Guillermo Barturen, Julio Sáez-Rodríguez, Marta Alarcón-Riquelme and Pedro Carmona-Sáez
17.40 - 18.00	**Phylogeny & Evolution**	**Intronic CNVs cause gene expression variation in human populations** Maria Rigau, David Juan, Alfonso Valencia and Daniel Rico
18.00 - 18.20		**Recent evolution of the epigenetic regulatory landscape in human and other primates** Raquel García-Pérez, Gloria Mas-Martín, Martín Kuhlwilm, Meritxell Riera, Antoine Blancher, Marc Marti-Renom, Luciano Di Croce, José Luis Gómez-Skarmeta, Tomás Marques-Bonet and David Juan
18.20 - 18.30	REMARKS	

Thursday, 15th/Nov		
09.00 - 10:00	**Keynote** **María Rodríguez** **"Artificial Intelligence approaches for personalized medicine"**	
10.00 - 10.30	Coffee Break	
10.30 - 10.50	**Translational Bioinformatics**	**vulcanSpot: a tool to prioritize therapeutic vulnerabilities in cancer** Javier Perales-Patón, Tomás Di Domenico, Coral Fustero-Torre, Elena Piñeiro-Yáñez, Carlos Carretero-Puche, Héctor Tejero, Alfonso Valencia, Gonzalo Gómez-López and Fátima Al-Shahrour
10.50 - 11.10		**Muscle Invasive Bladder Cancer stratification by genomic architecture** Sonia González-Alvaredo, David Juan, Miguel Vázquez, Alfonso Valencia, Francisco X. Real and Enrique Carrillo-De Santa Pau
11.10 - 11.30		**The immunogenic impacts of splicing alterations in small cell lung cancer** Juan Luis Trincado Alonso, Marina Reixachs, Eduardo Eyras and Jun Yokota
11.30 - 11.50		**Network, Transcriptomic and Genomic Features Differentiate Genes Relevant for Drug Response** Janet Piñero, Abel Gonzalez-Perez, Emre Guney, Joaquim Aguirre-Plans, Ferran Sanz, Baldo Oliva and Laura I. Furlong
11.50 - 12.10		**Training IBM Watson with MelanomaMine** Davide Cirillo, Andres Cañada, Javier Omar Corvi, José M. Fernández, Jose Antonio Lopez-Martin, Salvador Capella-Gutiérrez, Martin Krallinger and Alfonso Valencia
12.10 - 12.30	**Special Session: ELIXIR-ES**	**European Genome-phenome Archive (EGA) - Granular solutions for the next 10 years** Audald Lloret-Villas and Jordi Rambla
12.30 - 12.50		**OpenEBench. The ELIXIR platform for benchmarking** Salvador Capella-Gutiérrez, Juergen Haas, Vicky Sundesha, Dmitry Repchevski, Javier Garrayo, Víctor Fernández-Rodríguez, Miguel Madrid, Laia Codo, José M. Fernández, Anália Lourenço, J.L. Gelpí and Alfonso Valencia
12.50 - 13.00	COMPANIES	
13.00 - 14.30	Lunch Break & Posters (even submission numbers)	
14.30 - 15.10	**Highlights Session #2**	**SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification** Lorena de La Fuente Lorente, Ana Conesa, Manuel Tardaguila, Cristina Marti, Héctor Del Risco, Victoria Moreno, Cécile Pereira, Francisco Pardo-Palacios, Ali Mortazavi, Iakes Ezkurdia, Marc Ferrel, Marissa Macchietto, Lennart Martens, Maravillas Mellado, Susana Rodríguez, Michael Tress and Jesús Vázquez
15.10 - 15.50		**The Module of Personalized Medicine: the pioneer experience of including genomic data into the Andalusian Health System (SAS)** José L. Fernández-Rueda, Antonio Rueda, Javier López, Ignacio Medina, Javier Pérez Florido and Joaquín Dopazo
15.50 - 16.30	**Oncothon Results**	
16.30 - 17.00	Coffee Break & Posters (even submission numbers)	
17.00 - 18.20	**Round Table: "Bioinformatics Challenges for Personalized Medicine"** **Participants: Alfonso Valencia** (BSC & INB), **Joaquín Dopazo** (FPS), **Fátima Al-Shahrour** (CNIO) and **José Antonio López Escázmez** (IBS.GRANADA, GENyO) **Moderator: Rafael Solana Lara,** General Secretary of Research, Development and Innovation in Health, Andalusian Regional	
18.20 - 18.30	REMARKS	
18:30-19:30	**Satellite: "Towards a National Professional Bioinformatics Society"**	
21:00	**Gala Dinner at "CARMEN DE LOS CHAPITELES "**	

Friday, 16th/Nov		
09.00 - 09.20	**Structural Bioinformatics and Function**	**Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes** Oriol Pich, Ferran Muinos, Radhakrishnan Sabarinatahan, Iker Reyes-Salazar, Abel González-Pérez and Nuria López-Bigas
09.20 - 09.40		**Adopting structural flexibility to fill the gap between structure and function in lncRNAs** Uciel Chorostecki and Toni Gabaldón
09.40 - 10.00	**Omics Technologies**	**PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data** Rafael Hernández-de-Diego, Tarazona Sonia, Carlos Martínez-Mira, Leandro Balzano-Nogueira, Pedro Furió-Tarí, Georgios Pappas and Ana Conesa
10.00 - 10.30	Coffee Break	
10.30 - 11.30	**Closing Keynote** **Janet Kelso** **"The impact and fate of Neandertal DNA in modern humans"**	
11.30 - 12.10	**Highlights Session #3**	**A novel prokaryotic MMR pathway** Ana María Rojas and Jesús Blazquez
12.10 - 12.50		**Loose ends: almost one in five human genes still have unresolved coding status** Federico Abascal, David Juan, Laura Martínez Gomez, José Manuel Rodriguez, Jesús Vázquez, Irwin Jungreis, María Rigau and Michael Tress
12.50 - 13.00	FINAL REMARKS	
13.00 - 14.30	Lunch	

Satellite Meetings

STUDENT SYMPOSIUM

Like in previous editions, the VI Bioinformatics Student Symposium is meant to be a meeting point for all students and young postdocs working on the field of Bioinformatics in Spain. The Symposium will include research communications presented by participating students, social activities and a special career advice talk by leading expert.

When: 13th November 2018 from 14:00 to 18:00

Where: Centre for Genomics and Oncological Research (GENyO), Avenida de la Ilustración 114

More info:http://jbi2018.ugr.es/stuSympo.html

SECOND MEETING OF TRANSBIONET (TRANSLATIONAL BIOINFORMATICS NETWORK)

When: 13th November 2018 from 15:00 to 18:30

Where: Centre for Genomics and Oncological Research (GENyO), Avenida de la Ilustración 114

ONCOTHON

Oncothon will be a 2-day event that will bring the opportunity for professional from different areas such as Oncology, Bioinformatics, Engineer or Biomedical Research, from academic and enterprise settings, to know how cancer genomics is opening new pathways for cancer diagnosis and treatment. Attendees will have the opportunity to attend outstanding talks about how to exploit cancer genomics data and participate in collaborative sessions with interdisciplinary teams to collaborate and brainstorm about innovative solutions.This Oncothon arises in the framework of the ONCONET SUDOE European project

When: 12th-13th November 2018

Where: Parque Tecnológico de Ciencias de la Salud Avenida del Conocimiento 33

More info: http://oncothon.ptsgranada.com

PRESENTATION: TOWARDS A NATIONAL PROFESSIONAL BIOINFORMATICS SOCIETY

The advances in the life sciences that our society has experienced, together with the great generation of data produced by massive technologies, has made Bioinformatics a key element for generation of knowledge and analysis from exprerimental data. Bioinformatician is currently a high demanded profile and currently, there are positions in many laboratories worldwide. However, this demand has not been accompanied by a professional management of the sector.

There are many challenges that the bioinformatics community faces: professionalization of its profiles and competencies planning, career management in clinical environments, dialogue with different agents of the society etc...So, a National Professional Society of Bioinformatics should play a role in a medium term to progress in these issues

When: 15th Thursday 2018 from 18:30 to 19:30

Where: Paraninfo (JBI2018 Conference Center)

Keynotes

Dr. Julio Saez-Rodriguez is Professor of Medical Bioinformatics and Data Analysis at the Faculty of Medicine of the University of Heidelberg. He is an affiliated member of Sage-Bionetworks and a director of the DREAM challenges to catalyze the development of methods in systems biology, and the scientific coordinator of the EU-H2020 PrECISE (Personalized engine for Cancer integrative study and evaluation) consortium.
http://saezlab.org

Computational personalised medicine in cancer: from in vitro to in vivo.
In this talk I will revise our work analysing large pharmaco-genomic screenings in cell lines, that provide rich information about alterations in tumours that confer drug sensitivity or resistance. Integration of this data with prior knowledge on signaling pathways and transcription factors provides biomarkers and offer hypotheses for novel combination therapies. Our own analysis as well as the results of a crowdsourcing effort (as part of a DREAM challenge) reveals that prediction of drug efficacy is far from accurate, implying important limitations for personalised medicine. An important aspect that deserves further attention is the dynamics of signaling networks and how they response to perturbations such as drug treatment. I will show how cell-specific logic models, trained with measurements upon perturbations, can provide new biomarkers and treatment opportunities not noticeable by static molecular characterisation. Finally, I will show how, using novel microfluidics-based technologies, this approach can also be applied directly to tumor biopsies**.**

Dr. María Rodriguez Martinez is member of the Systems Biology group at IBM Research – Zürich, and an associated member of the Department of Biology at ETH since 2014. Her current research focuses on the development of computational and statistical approaches to unravel cancer molecular mechanisms using high-throughput multi-omics datasets and single-cell molecular data.
https://researcher.watson.ibm.com/researcher/view.php?person=zurich-MRM

Artificial Intelligence approaches for personalized medicine.
In recent years, deep learning has become one of most active fields in machine learning with astounding performances in a broad area of applications such as computer vision, speech recognition and natural language processing. In computational biology, the recent availability of large amounts of data generated by word-wide consortia together with technical developments facilitating the implementation and training of more performant models have made possible the broad application of deep learning to a vast set of problems. In this talk, I will present current activities at the Computational Systems Biology group in IBM Research, Zurich, that illustrate the application of AI approaches to integrate disparate data types with the goal of unraveling disease mechanisms and develop personalized patient models. Specifically, I will show two examples. First, I will demonstrate how state-of-the-art text ingestion and analysis can be used to automatically extract knowledge from text sources and obtain comprehensive maps of molecular interactions. Second, I will explain how deep learning can be used to characterize tumor heterogeneity in single cell data and identify correlations that are predictive of patient prognosis and other important clinical endpoints.

Dr. Janet Kelso is head of the Bioinformatics research group at the Max-Planck Institute for Evolutionary Anthropology in Leipzig, Germany. Her research focuses on comparative primate genomics and ancient DNA with a particular emphasis on analysis of the genomes of archaic humans.
https://www.eva.mpg.de/genetics/staff/janet-kelso/cv.html

The impact and fate of Neandertal DNA in modern humans

Recent technological advances have made it possible to recover genome sequences from a number of archaic and early modern humans. Analyses of these genomes have provided direct evidence for interbreeding between early modern and archaic humans. As a result all present-day people outside of Africa carry approximately 2% Neandertal DNA, and some populations, largely in Oceania, also carry DNA from Denisovans. This introgressed DNA has been shown to have both positive and negative outcomes for present-day carriers: underlying apparently adaptive phenotypes as well as influencing disease risk. In recent work we have identified Neandertal haplotypes that are likely of archaic origin and determined the likely functional consequences of these haplotypes using public genome, gene expression, and phenotype datasets. We have also used simulations, as well as the distribution of Neandertal DNA in ancient modern humans, to understand how selection has acted on Neandertal introgressed sequences over the last 45,000 years.

JBI-2018 Accepted Abstracts for Oral Presentations

Highlights

GENOME DENMARK: SEQUENCING AND DE NOVO ASSEMBLY OF THE DANISH REFERENCE GENOME

José Mg Izarzugaza[1], Søren Brunak[2] and Danish Pangenome Consortium[2]

[1]*DTU Bioinformatics, Denmark*
[2]*NNF Center for Protein Research, Denmark*

Abstract: Hundreds of thousands of human genomes are now being sequenced to characterize genetic variation and use this information to augment association mapping studies of complex disorders and other phenotypic traits Genetic variation is identified mainly by mapping short reads to the reference genome or by performing local assembly. However, these approaches are biased against discovery of structural variants and variation in the more complex parts of the genome. Hence, large-scale de novo assembly is needed. Here we show that it is possible to construct excellent de novo assemblies from high-coverage sequencing with mate-pair libraries extending up to 20 kilobases. We report de novo assemblies of 150 individuals (50 trios) from the GenomeDenmark project. The quality of these assemblies is similar to those obtained using the more expensive long-read technology. We use the assemblies to identify a rich set of structural variants including many novel insertions and demonstrate how this variant catalogue enables further deciphering of known association mapping signals. We leverage the assemblies to provide 100 completely resolved major histocompatibility complex haplotypes and to resolve major parts of the Y chromosome. Our study provides a regional reference genome that we expect will improve the power of future association mapping studies and hence pave the way for precision medicine initiatives, which now are being launched in many countries including Denmark.

PUBLICATION: Maretty et al., Nature 2017. https://www.nature.com/articles/nature23264

SCUTOIDS, A GEOMETRICAL SOLUTION TO THREE-DIMENSIONAL PACKING OF EPITHELIA

Luis M Escudero[1]

[1]*Instituto de Biomedicina de Sevilla. Universidad de Sevilla, Spain*

Abstract: As animals develop, the initial simple planar epithelia of embryos must be sculpted into complex three-dimensional tissues. However, the architecture and packing of curved epithelia remains largely unknown. Here, by means of computational modelling, we show that cells in bent epithelia are compelled to adopt a novel shape that we name “scutoids”. The detailed image analysis of diverse tissues and organs confirm the generation of apico-basal transitions among cell during morphogenesis. Using biophysics arguments we infer that scutoids allow the minimization of the tissue energy and stabilize the tree-dimensional packing of the tissue. Altogether, we argue that scutoids are nature´s solution to achieve epithelial bending and the missing piece for developing a unifying and realistic model of epithelial architecture. Our results pave the way to understand the biomechanics of morphogenesis in developing organisms and sheds light on the underlying logic of 3D cellular self-organization.

I will present new experiments and the computational tools that we are developing to characterize the molecules that are important for the appearance and maintaining of scutoids.

PUBLICATION: Gómez-Gálvez et al., Nat Commun. 2018. https://www.nature.com/articles/s41467-018-05376-1

LOOSE ENDS: ALMOST ONE IN FIVE HUMAN GENES STILL HAVE UNRESOLVED CODING STATUS

Federico Abascal[1], David Juan[2], Laura Martinez Gomez[3], Jose Manuel Rodriguez[3,4], Jesús Vázquez[5], Irwin Jungreis[6], María Rigau[7] and Michael Tress[3]

[1]*Wellcome Trust Sanger Institute, United Kingdom*
[2]*Institute of Evolutionary Biology, UPF-CSIC, Spain*
[3]*Spanish National Cancer Research Centre, CNIO, Spain*
[4]*Spanish National Bioinformatics Institute, INB, Spain*
[5]*Spanish National Center for Cardiovascular Disease, CNIC, Spain*
[6]*Massachusetts Institute of Technology, MIT, United States*
[7]*Barcelona Supercomputing Center, BSC, Spain*

Abstract: There are three well-maintained manual reference databases for the human genome, RefSeq (1), UniProtKB (2) and Ensembl/GENCODE (3,5). Over the years these three databases have converged on similar numbers of protein coding genes and the number of protein coding genes in the human reference gene set has been more or less stable since 2012 (5). Despite this, the human gene sets are in certain state of flux with coding genes being added and reclassified with each new release.

Even though the three reference sets contain the same number of coding genes, the 20,000 plus coding genes are not the same in each. Here we carry out a manual cross-reference between the three reference catalogues to investigate the differences between the Ensembl/GENCODE, RefSeq and UniProtKB proteomes.

We find that the number of annotated coding genes in the union of the three reference sets exceeds 22,000, while at the same time the three reference sets classify one in every eight coding genes differently. How many of these 2,754 differently annotated genes are protein coding? In fact we find that genes that are annotated as coding by just one or two sets of manual annotators are starkly different from those annotated as coding by all three reference sets. Coding genes that are differently classified across the three sets are rich in non-coding features and large-scale genetic variation data suggests that many of these potential coding genes are unlikely to be functionally important.

This analysis clearly shows the importance of a curated human reference set. Over the years since the human genome sequence was released rigorous manual annotation has brought us considerably closer to a final catalogue of human coding genes. Annotators agree on almost 90% of coding genes, but the final 10% of genes, those with the most conflicting evidence, are proving more difficult to classify.

1. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., 44, D733-745.
2. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase Nucleic Acids Res., 45, D158-D169.
3. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., et al. (2017) Ensembl 2017. Nucleic Acids Res., 2017;45:D635-42.
4. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S., et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res., 22, 1760-1774.
5. Southan,C. (2017) Last rolls of the yoyo: Assessing the human canonical protein count. F1000Research, 6, 448.

PUBLICATION: Abascal et al., Nucleic Acids Res. 2018. https://www.ncbi.nlm.nih.gov/pubmed/29982784

SQANTI: EXTENSIVE CHARACTERIZATION OF LONG-READ TRANSCRIPT SEQUENCES FOR QUALITY CONTROL IN FULL-LENGTH TRANSCRIPTOME IDENTIFICATION AND QUANTIFICATION

Lorena de La Fuente Lorente[1], Ana Conesa[1,2], Manuel Tardaguila[3], Cristina Marti[1], Héctor Del Risco[2], Victoria Moreno[1], Cécile Pereira[2], Francisco Pardo-Palacios[1], Ali Mortazavi[4], Iakes Ezkurdia[5], Marc Ferrel[2], Marissa Macchietto[4], Lennart Martens[6], Maravillas Mellado[1], Susana Rodríguez[1], Michael Tress[7] and Jesús Vázquez[5]

[1]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[2]*University of Florida, United States*
[3]*Sanger Institute, United Kingdom*
[4]*University of California, United States*
[5]*Centro Nacional de Investigaciones Cardiovasculares, CNIC, Spain*
[6]*UGent / VIB, Belgium*
[7]*Spanish National Cancer Research Centre, CNIO, Spain*

Abstract: With the increasing utilization of long read technologies, the necessity for a tool that provides a comprehensive classification of novel transcripts as well as their deep characterization and quality control is ever more pressing. Exhaustive transcriptome curation methods to remove potential artifactual isoforms are essential at a time when long read sequencing and splicing-aware transcriptomics studies are becoming more popular to reveal system regulation triggered by mechanisms as alternative splicing. Here we present SQANTI (Structural and Quality Annotation of Novel Transcript Isoforms), a tool for the analysis long-read transcriptomics data that provides the methods to deliver quality-evaluated (quality control metrics) and curated full-length transcriptomes (machine learning approach). Using SQANTI we revealed that, even well annotated genomes as mouse genome, has still an important fraction of missing transcripts. More importantly, we show that incomplete annotation has a strong negative impact in the accuracy of current isoform expression quantification algorithms. Moreover, since it was released, SQANTI has been applied to multiple organisms, different long-read sequencing platforms (PacBio, Nanopore, etc) and different transcriptome reconstruction pipelines as well as becoming a standard long-read QC procedure for PacBio Iso-seq bioinformatics developers. Results highlight the relevance of SQANTI to fully understand long-read defined transcriptomes.

PUBLICATION: Tardaquila et al., Geome Res. 2018. https://www.ncbi.nlm.nih.gov/pubmed/29440222

A NOVEL PROKARYOTIC MMR PATHWAY

Ana María Rojas[1] and Jesus Blazquez[2]

[1]*CABD-CSIC, Spain*
[2]*CNB-CSIC, Spain*

Abstract: NucS is an endonuclease firstly identified in Archaea involved in Mismatch repair (MMR) with no structural homology to known MMR factors. By genetic screenings we report [1] that this protein is required for mutation avoidance and anti-recombination, hallmarks of the canonical MMR in the surrogate model Mycobacterium smegmatis, lacking classical MutS-MutL factors. Furthermore, phenotypic analysis of naturally occurring polymorphic NucS in a M. smegmatis surrogate model, suggests the existence of M. tuberculosis mutator strains.

Computational and evolutionary studies of NucS indicate a complex evolutionary emergence based in distinct protein domain fusions followed by at least two horizontal gene transfers leading to a disperse distribution pattern in prokaryotes. Together, these findings indicate that distinct pathways for MMR have evolved at least twice in nature. The analyses of these findings in the evolutionary context of the classical MMR proteins open novel and intriguing questions in the emergence of the MMR systems.

[1]. Castaneda-Garcia, A; Prieto, A.I., Rodriguez-Beltran, J., Alonso, N., Cantillon, D., Costas. C., Perez-Lago, L., Zegeye, E.D.., Herranz, M., Plocinski, P., Tonjum, T., Garcia de Viedma, D., Paget, M, Waddell, S.J., Rojas, A.M.*, Doherty, A.J.*, & Blazquez, J*. (2017) A non-canonical mismatch repair pathway in prokaryotes. Nature Communications (27th January) DOI:10.1038/ncomms14246

PUBLICATION: Castaneda-Garcia et al. Nature Comm. 2018. https://www.nature.com/articles/ncomms14246

THE MODULE OF PERSONALIZED MEDICINE: THE PIONEER EXPERIENCE OF INCLUDING GENOMIC DATA INTO THE ANDALUSIAN HEALTH SYSTEM (SAS)

José L. Fernández-Rueda[1], Antonio Rueda[2], Javier López[2], Ignacio Medina[3], Javier Pérez Florido[4] and Joaquín Dopazo[5]

[1]*Clinical Bioinformatics Area. Fundación Progreso y Salud (FPS). CDCA, Hospital Virgen del Rocio. Sevilla. Spain*
[2]*Genomics England, London, United Kingdom*
[3]*HPC Service, UIS, University of Cambridge, United Kingdom*
[4]*Bioinformatics in Rare Diseases (BiER). Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER). FPS. Hospital Virgen del Rocio. Sevilla. Spain*
[5]*INB/ELIXIR-ES, FPS, Hospital Virgen del Rocío, Sevilla, Spain*

Abstract: The Andalusian community launched the first local genomic project in Spain in 2011, the Medical genome Project, and since then has cumulated an enormous experience in the management of genomic data. This experience, in combination with the largest population in Europe (8.5 million) covered by a universal electronic health record (eHR) fostered the innovative Andalusian Personalized Medicine initiative. The novelty in this initiative is the implementation of bioinformatics tools into the corporative systems of the SAS that link the genomic data of patients to their clinical records. In this way, in addition to use massive sequencing for diagnostic (the rare diseases initiative) or treatment recommendation (the ongoing cancer genomics initiative) purposes, the genomic data remain into the system, linked to patient's eHRs. This transforms the database of the SAS in an immense potential prospective clinical study in which, as more genomic data and clinical data are cumulated along the time, the genomic variants of the patients can be associated to different clinical endpoints of interest. This allows an unprecedented capability for biomarker discovery based in growing information storage. The population health database provides a framework for this discovery potential.

This is possible because of the introduction of the MMP into the corporative informatics systems of the SAS, coupled to the sequencing unit, equipped with Illumina NextSeq 550 and HiSeq 2500 sequencers, located in the Hospital Virgen del Rocío (Sevilla). The MMP is a sophisticated web interface, based in previous developments [1, 2], that allows semi-automated diagnosis of genetic diseases. In the pilot phase we experienced a reduction of time to diagnosis since blood extraction from over three weeks to 8 hours. MMP interfaces the advanced genomic data management system OpenCGA, used in the Genomics England 100,000 genomes project, and the knowledge database CellBase [3]. The MMP provides, for the first time, the possibility of carrying out massive sequencing-based diagnosis (and soon cancer treatment recommendation) within a public health system. MMP is based in the IVA architecture, co-developed between the Clinical Bioinformatics Area and the University of Cambridge.

References

1. Aleman A, Garcia-Garcia F, Medina I, Dopazo J: A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications. Nucleic Acids Res 2014, 42(Web Server issue):W83-87.

2.Aleman A, Garcia-Garcia F, Salavert F, Medina I, Dopazo J: A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. Nucleic Acids Res 2014, 42(Web Server issue):W88-93.

3.Bleda M, Tarraga J, de María A, Salavert F, Garcia-Alonso L, Celma M, Martin A, Dopazo J, Medina I: CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. Nucleic Acids Res 2012, 40(Web Server issue):W609-614.

NGS APPLICATIONS

PRECISE SAMPLE CLASSIFICATION OF THE METASUB ENVIRONMENTAL MICROBIOTA USING FUNCTIONAL BIOMARKERS

Carlos S. Casimiro-Soriguer[1], Javier Pérez Florido[1], Daniel López López[1], Carlos Loucera[1] and Joaquín Dopazo[1]

[1]*Fundación Pogreso y Salud, Spain*

Abstract: The availability of hundreds of city microbiome profiles allows the development of predictors of origin based on microbiota composition. Here we use a transformation of the conventional bacterial strain or gene abundance profiles to functional profiles that account for bacterial metabolism and other cell functionalities. We explore the use of functional profiles not only to predict the most likely origin of a sample but also to provide a functional point of view in the study the biogeography of the microbiota.

IDENTIFICATION OF RNA-PROCESSING PROGNOSTIC AND THERAPEUTIC MARKERS IN MLL-REARRANGED INFANT ACUTE LEUKEMIA

Adria Closa[1], Antonio Agraz-Doblas[2,3], Ignacio Varela[3], Pablo Menéndez[2,4,5] and Eduardo Eyras[1,5]

[1]*Universitat Pompeu Fabra, UPF, Spain*
[2]*Josep Carreras Leukemia Research Institute, Spain*
[3]*Instituto de Biomedicina y Biotecnología de Cantabria, Spain*
[4]*Centro de Investigación Biomédica en Red de Cáncer, CIBERONC, Spain*
[5]*Catalan Institution for Research and Advanced Studies, ICREA, Spain*

Abstract: Infant acute lymphoblastic leukemia (ALL) has a poor prognosis, especially with MLL gene rearrangements (MLL-r), which occur in ~80% of the patients during embryonic/fetal hematopoiesis. Genome-sequencing studies of MLL-r ALL patients have shown a very low frequency of somatic mutations, indicating that MLL may not require additional alterations to induce full transformation. However, ALL cannot be recapitulated in a mouse model by only integrating the fusion, suggesting that additional alterations are necessary for leukemogenesis.

MLL fusions have the potential to impact the RNA processing of genes at genome scale through changes in transcriptional elongation, thereby providing a new layer of molecular variation that has remained undetected so far, and which could lead to new prognostic markers and therapeutic strategies. We present here an exhaustive analysis of the RNA-processing alterations in infant ALL samples in relation to the MLL fusions. We have analyzed RNA-sequencing on a cohort of 32 MLL-r and 10 non-MLL infant ALL cases, plus 5 normal B-cell progenitor samples.

In a preliminary analysis we detect a clear differential expression pattern between the different types of MLL-r ALL (MLL-AF4 and MLL-AF9) and non-MLL samples indicating an impact of MLL fusions on gene expression. However, we found a low level of overlap in the differential expressed genes between the fusion groups and non-MLL samples, pointing towards fusion-specific molecular alterations. In particular, we found a pattern of expression alteration in several splicing factors. In agreement with this finding, we found common and specific events that are differentially spliced between MLL-r samples and controls, with a high proportion of skipping exons and alternative first exons specific for both types of MLL fusion. We further present an analysis of the potential functional impacts and the

phenotypic convergence of these alterations across patients. This is the first study of RNA-processing alterations in association to MLL-AF4 and MLL-AF9 fusions in ALL and of their role in leukemogenesis.

IMPLEMENTATION OF A COPY NUMBER VARIANT DETECTION WORKFLOW WITHIN THE URD-CAT PILOT PROJECT ON PERSONALISED MEDICINE

Gemma Bullich[1], Steven Laurie[1], Jordi Morata[1], David Ovelleiro[1], Sandra Redó[1], Ricky Joshi[1], Cristina Luengo[1], Leslie Matalonga[1], Genís Parra[1], Raul Tonda[1] and Sergi Beltran[1,2]

[1]*CNAG-CRG, Spain*
[2]*Universitat Pompeu Fabra (UPF), Spain*

Abstract: The Undiagnosed Rare Disease Program of Catalonia (URDCat) is a pilot project aiming to provide the Catalan Health System with personalised genomic medicine as a fully integrated service for patients with rare diseases. Genotypic and phenotypic data is collated, integrated and analysed through the RD-Cat platform (rdcat.cnag.crg.eu), based on the RD-Connect platform (platform.rd-connect.eu). The RD-Cat platform already integrates pseudo-anonymised clinical and genomic data of more than 1300 individuals, including 800 RD patients.

RD-Cat has collated 413 pre-existing exomes and genomes, and has sequenced 406 new whole exomes, 94 new genomes, and 23 RNA-Seq experiments. Exome sequencing is widely accepted as a robust and cost-effective approach for single-nucleotide variant identification. However, detection of copy number variants (CNV) is still challenging with low sensitivity and high false positive rates due to the targeted nature of exome-capture protocols. With the aim of implementing a standard workflow for the detection of CNVs within the URDCat project, we have compared the results obtained from the analysis of all available exomes with 3 tools, ExomeDepth, XHMM and Conifer, selected after a preliminary evaluation of 6 tools.

Methods

WES data from 809 samples were split by capture kit and independently analysed with ExomeDepth, XHMM and Conifer using default settings. CNVs overlapping at least 50% with CNVs reported by Conrad et al (2010) cohort were discarded. To compare the results obtained by the 3 programs, only CNVs with an observed frequency <1% within our cohort were considered.

Results

Analysis of the output of the 3 tools showed at least one CNV was identified in approximately 90% of the individuals according to ExomeDepth (717 out of 809) and XHMM (710 out of 809), whereas in only 12% of the cohort according to Conifer (100 out of 809). The mean number of CNV calls per sample was 33 for ExomeDepth, 48 for XHMM and 22 for Conifer, considering only those samples with at least 1 CNV identified. The median size of the predicted CNVs was 4.4 Kb for ExomeDepth (ranging from 1 bp to 15863 Kb), 12 Kb for XHMM (ranging from 57 bp to 17966 Kb) and 57 Kb for Conifer (ranging from 392 bp to 23603 Kb). Altogether, only 4% of all CNV's were called by all three tools. The highest overlap (18%) was between ExomeDepth and XHMM, while ExomeDepth and Conifer shared the lowest number of CNVs (6%).

Conclusions

Our preliminary results showed that CNV detection using 3 different tools in parallel might increase sensitivity as it detects a wide range of CNV sizes. However, low concordance exists among the 3 different programs. A confidence score based on the degree of overlap might be useful to stratify the CNVs according to their probability to be real.

COHESIN VARIANTS SA1 AND SA2 PLAY DISTINCT ROLES IN CHROMATIN ORGANIZATION AND GENE EXPRESSION REQUIRED FOR EMBRYONIC STEM CELL IDENTITY

Daniel Gimenez[1], Aleksandar Kojic[1], Marc A. Marti-Renom[2], François Le Dily[3], Ana Cuadrado[1] and Ana Losada[1]

[1]*Spanish National Cancer Research Centre (CNIO), Madrid, Spain*
[2]*CNAG–CRG, Centre for Genomic Regulation, Barcelona, Spain*
[3]*Centre de Regulació Genòmica (CRG), Barcelona, Spain*

Abstract: Cohesin is a DNA entrapping complex essential for genome architecture and inheritance. In somatic cells it consists of SMC1A, SMC3, RAD21 and either SA1 (STAG1) or SA2 (STAG2) resulting in two different cohesin complexes that are not functionally redundant although either one is sufficient for cell proliferation. We are aimed to analyse the specific functions of both cohesin variants in mouse embryonic stem cells (mES) by combining the study of genome-wide cohesin variants distribution by ChIP-seq, how this distribution affects specific mES genomic architecture by HiC, and finally how changes in genomic architecture due to loss of each variant might affect gene expression by RNA-seq. In these cells, genes required for pluripotency must be actively transcribed in order to preserve the pluripotent state. Such high levels of expression are reached thanks to the specific conformation of chromatin in superenhancers that ensures efficient binding of active enhancers to the promoters of pluripotency genes. In addition, mES identity requires the repression of lineage specification genes that must be kept in a "poised" state, silent but ready to be rapidly activated. Such poised state depends on the presence of polycomb complexes (PRC1 and PRC2) that define genomic regions with specific and unique structural features. Our data show that while cohesin SA1 always co-localizes with the insulating protein CTCF, there is an important fraction of cohesin SA2 that specifically co-localizes at superenhancers and polycomb repressed promoters in the absence of CTCF. By means of HiC analysis, we have explored the specific roles of each cohesin variant in different levels of chromatin organization, including A/B compartimentalization, TAD integrity, superenhancer looping and polycomb domains compaction. We have seen that cohesins SA1 and SA2 have unique and specialized functions important for mES cells identity.

IMPROVING RNA-SEQ GERMLINE VARIANT CALLING WITHIN RD-CONNECT CONSORTIUM

Mattia Bosio[1,2], Alfonso Valencia[1,2,3] and Salvador Capella-Gutiérrez[1,2]

[1]*Barcelona Supercomputing Center (BSC), Spain*
[2]*Spanish National Bioinformatics Institute (INB), ELIXIR-ES, Spain*
[3]*Catalan Institution for Research and Advanced Studies, ICREA, Spain*

Abstract: Within the RD-Connect consortium we developed an RNASeq variant calling pipeline, automating sample analysis for the analysis platform. We integrated best practices from ENCODE Consortium, GATK toolkit, and custom filtering to produce reliable variant calls.

Analysing matching WGS and RNA-seq data, we investigated similarities i.e. genotype concordance and overlap, which are in line with the literature, and differences between variant datasets (high number of RNA false positives). We then developed a post-processing strategy for RNA-seq variants to produce more reliable calls and characterize the substantial fraction of novel variants compared to WGS (i.e. which variants are of biological origin and which of technical errors). For this task, we integrated relevant features from best practices and RNA-seq calling protocols e.g. [1-3], and evaluated their impact on concordance, sensitivity, and precision against WGS callset.

Aiming to a better characterization of variants, not limited to known sites, we jointly processed a set of 20 samples with matching DNA and RNA-seq data, modeling how different features can help discriminating variants falling in two sets i.e. RNA private variants like RNA-editing and/or false positives, and shared with WGS; via a random forest estimation. With this, we can estimate each variant likelihood of being a false positive, enabling a deeper characterization of RNA-seq variants, and easing the downstream integration efforts and/or the interpretation process.

The developed classification framework showed to improve the precision-recall performances of the standard variant calling pipeline with respect to finding germline variants. Comparing our framework with similar approaches like [1], on independent samples from RD-Connect and from public gold-standard resources (Gene in a bottle consortium, NA12878 sample), we showed that the random forest classification achieves superior results than hard-filtering strategies.

[1] Piskol et al. Am. J. Hum. Genet. 2013
[2] Oikkonen el al. Wellcome Open Res. 2017
[3] Xu. Computational and Structural Biotechnology Journal, 2018

THE REGULATORY GENOME OF DROSOPHILA REGENERATION

Cecilia Coimbra Klein[1,2,3], Elena Vizcaya-Molina[3], Florenci Serras[3], Roderic Guigó[1,2] and Montserrat Corominas[3]

[1]*Centre for Genomic Regulation, CRG, Spain*
[2]*Universitat Pompeu Fabra, UPF, Spain*
[3]*Departament de Genètica, Microbiologia i Estadística, IBUB, Universitat de Barcelona, Spain*

Abstract: One of the key questions in regenerative biology is to unveil the regulatory regions capable to trigger tissue recovery. Regeneration is the ability to reconstruct missing parts. The capacity to regenerate varies greatly, not only between species, but also between tissues and organs, as well as from one developmental stage to another in the same species. Drosophila imaginal discs show a high regenerative capacity after genetically induced cell death. We performed genome-wide chromatin landscape analyses (ATAC-Seq and RNA-Seq) at different time points (early, mid and late) of Drosophila imaginal disc regeneration to study the transcriptional programs as well as the regulatory elements responsible for tissue regeneration.

We identified sets of upregulated genes located close to one another in the linear genome (herein called clusters), mostly at early and mid regeneration, indicating that large regions, rather than individual genes, may be controlled by the same regulatory elements. Open chromatin regions that presented higher accessibility in regeneration compared to controls (namely damage-responsive regulatory elements: DRREs), were classified according to their position relative to the closest transcription start site (TSS) of a gene: core promoter (±100 bp of the TSS), in the first intron (FI), proximal (±2 kb from the TSS), and distal (more than ± 2 kb away from the TSS). We distinguished two types of DRREs: emerging, open regions detected only after damage (eDRREs); and increasing, regions already open in control, but displaying increased accessibility after damage (iDRREs). We have also validated several DRREs using enhancer reporter fly lines after inducing apoptosis as well as after physical injury. Since spatial chromatin organization connects active enhancers to target promoters to regulate gene expression, we confirmed individual interactions between eDRREs and clusters of co-regulated genes by Chromatin Conformation Capture analysis. Moreover, DRREs contained conserved binding motifs for transcription factors that are upregulated and required for regenerating organs in fly, zebrafish and mouse.

Our findings indicate there is global co-regulation of gene expression where genes localized in genomic clusters may be regulated by the same elements. Furthermore, we found a regeneration program driven by the cooperation among regulatory elements acting exclusively within damaged tissue, with enhancers co-opted from other tissues and other developmental stages, as well as with endogenous enhancers that show increased activity after injury. Such elements

host binding sites for regulatory proteins that include a core set of conserved transcription factors that may control regeneration across metazoans.

Integrative Bioinformatics

PROTEOME-WIDE DISCOVERY OF DEGRONS AND ANALYSIS OF THEIR ROLE IN TUMORIGENESIS ACROSS CANCER TYPES

Francisco Martínez-Jiménez[1], Abel Gonzalez-Perez[1] and Núria López-Bigas[1]

[1]*Institute for Research in Biomedicine, IRB Barcelona, Spain*

Abstract: The ubiquitin mediated proteolysis system (UPS) is involved in both quality control and regulation of protein levels that control crucial cellular processes. Degrons are short linear motifs embedded within the sequences of substrates that are recognized and bound by E3 ligases, forming the basis of the specificity of UPS. Three decades of research have yielded the identification of degrons for only ~6% of the 600 E3-ligases encoded in the human genome. Identifying new degrons is key to understand the operation of UPS and its role in diseases.

We used a combination of machine learning, protein-protein interactions and matched exome, transcriptome and proteome data from 8,167 human tumors to perform a proteome-wide discovery of degrons. The integrative approach enabled the identification of 449 high-confidence instances of annotated degrons. We validated the approach showing that missense mutations and inframe indels affecting new degron instances are significantly more stabilizing than analogous mutations located outside of degradation motifs. A search for degrons of yet unknown E3-ligases pinpointed 41 regions harbouring highly stabilizing mutations and resembling the biochemical properties of annotated degrons. Such regions involve essential proteins such as LCK, ERBB3 or CHEK1. We also showed that degron destroying mutations are positively selected in human tumors. Finally, cohort specific analysis of UPS perturbations revealed a mutually exclusive pattern of mutations between degron mutations and driver alterations in their E3 ligases posing UPS perturbations as a widespread mechanism of tumorigenesis.

ANALYSIS OF TRANSCRIPTION FACTOR ACTIVITY PATTERNS IN SYSTEMIC LUPUS ERYTHEMATOSUS

Raul Lopez-Dominguez[1], Daniel Toro-Domínguez[1], Jordi Martorell-Marugan[1], Christian Holland[2], Guillermo Barturen[1], Julio Saez-Rodriguez[3], Marta Alarcon-Riquelme[1] and Pedro Carmona-Saez[1]

[1]*Centre for Genomics and Oncological Research (GENYO), Spain*
[2]*Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH Aachen University, Germany*
[3]*Institute of Computational Biomedicine, Heidelberg University, Germany*

Abstract: Systemic Lupus Erythematosus (SLE) is a complex and heterogeneous autoimmune disease where the interaction between genetics and environment factors plays an important role in its development (Delgado-Vega et al. 2010). Patients with lupus can suffer from periods of disease activity that may remit spontaneously, or in most cases, require treatment to be controlled.

During disease activity tissue damage occurs, and aggressiveness and duration of these activity periods, as well as response to treatment, are highly heterogeneous across different patients.

During the last two decades, many researchers have focused their efforts towards the characterization of the genetic determinants that are associated with SLE pathogenesis. These efforts have established more than fifty SLE-associated loci and the finding that type I interferon-inducible gene expression signature is commonly deregulated in SLE patients.

Nevertheless, although it is has been shown that most SLE loci occur in regulatory regions of susceptibility genes (Maurano et al. 2012), there is a lack of systematic analyses that explore the gene regulatory network associated with Lupus pathogenesis. Some previously published works (Harley et al. 2018; Dozmorov, Wren, and Alarcón-Riquelme 2014) have analyzed the enrichment of Transcription Factor (TF) binding sites in regulatory regions of SLE-associated loci. Nevertheless, these studies provide evidences about physical interactions among TFs and genomic regions of SLE-loci, but they do not provide information about the TF activity or the deregulation of gene expression programs.

In this work we have explored activity patterns of all human TFs from TFclass database in SLE samples by analyzing expression levels of their direct target genes (the so-called TF regulon)(Garcia-Alonso et al. 2018). TF activity scores can be estimated based on the levels of its associated target genes. The analysis of the TFs by sample activity matrix has allowed us to establish sets of TFs with differential behaviour among SLE and healthy controls, as well as correlations with disease activity patterns. We have analyzed two large independent cohorts finding a set of TFs that show consistent patterns in both sets. Some of these TFs have been previously described in the context of SLE but there are others that provide new insights into molecular mechanisms that might be important contributors for the understanding of SLE pathogenesis.

Phylogeny & Evolution

INTRONIC CNVS CAUSE GENE EXPRESSION VARIATION IN HUMAN POPULATIONS

María Rigau[1], David Juan[2], Alfonso Valencia[1] and Daniel Rico[3]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Institut de Biologia Evolutiva, CSIC-UPF, Spain*
[3]*Institute of Cellular Medicine, United Kingdom*

Abstract: Introns comprise about half of the human non-coding genome and they can have important regulatory roles. However, mutations in introns are usually ignored when looking for normal or pathogenic genomic variation and little is known about their population patterns of structural variation and their functional implication. By combining the most extensive maps of CNVs in human populations, we have found that intronic losses are the most frequent copy number variants (CNVs) in protein-coding genes in human, affecting 4,147 genes (including 1,154 essential genes and 1,638 disease-related genes). This intronic length variation results in dozens of genes showing extreme population variability in size, with 40 genes with 10 or more different sizes and up to 150 allelic sizes. Intronic losses are frequent in evolutionarily ancient genes that are highly conserved at the protein sequence level. This result contrasts with losses overlapping exons, which are observed less often than expected by chance and almost exclusively affect primate-specific genes. By integrating CNV and RNA-seq data, we have showed that intronic loss can be associated with significant differences in gene expression levels in the population (CNV-eQTLs). These intronic CNV-eQTLs regions are enriched for intronic enhancers and can be associated with expression differences of other genes showing long distance intron-promoter 3D interactions. Our data suggests that the frequent gene length variation in protein-coding genes resulting from intronic CNVs might influence their regulation in different individuals.

RECENT EVOLUTION OF THE EPIGENETIC REGULATORY LANDSCAPE IN HUMAN AND OTHER PRIMATES

Raquel Garcia-Perez[1], Gloria Mas-Martin[2], Martin Kuhlwilm[1], Meritxell Riera[1], Antoine Blancher[3], Marc Marti-Renom[2], Luciano Di Croce[2], Jose Luis Gómez-Skarmeta[4], Tomas Marques-Bonet[1] and David Juan[1]

[1]*Institute of Evolutionary Biology. UPF-CSIC, Spain*
[2]*Centro Nacional de Análisis Genómico - Centro de Regulación Genómica, CNAG-CRG, Spain*
[3]*Laboratoire d'Immunogénétique moléculaire, Faculté de Médecine Purpan, Université Toulouse 3, France*
[4]*Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide, Spain*

Abstract: Although comparative epigenomics has intensively studied switch on/off changes in regulatory regions associated to evolutionary differences, the detailed evolutionary changes in activity of enhancers and promoters during primate evolution remain mostly unexplored. Here we characterize the evolutionary dynamics of the epigenomic activity of these regulatory elements in the primate lineage. To that end we have comprehensively profiled lymphoblastoid cell lines (LCLs) from human, chimpanzee, gorilla, orangutan and macaque by performing ChIP-seq for five histone marks, ATAC-seq, WGBS and RNA-seq experiments.

Integration of genome-wide epigenomic and gene expression data has allowed us to identify very characteristic genomic and epigenomic conservation patterns associated to strong, weak and poised activities in promoters and enhancers. In particular, strong promoters are highly epigenomically conserved while poised and weak promoters are mostly species specific. Enhancers show an activity-related bimodal behavior with most of them been either highly

conserved or species specific. We observed a clear activity-dependent association between genomic and epigenomic conservation.

Moreover, building of gene-specific epigenomic architectures based on integration of genomic annotations and available 3D contact maps showed that gene regulation concentrates a high proportion of epigenomically conserved strong promoters in primates and most of the poised promoters while only a very small proportion of strong, poised or weak enhancers could be associated to gene regulation with a weaker connection with epigenomic conservation. In order to study the dynamics of changes of epigenomic gene regulation in the primate lineage we analyzed the differences in calibrated RNA-seq and ChIP-seq and ATAC-seq signals. By exploring the interplay between these evolutionary patterns, we were able to disentangle the connection of gene expression and epigenomic changes at different levels of genomic resolution. Multivariate regulatory models based on the global epigenomic signals in regulatory architectures explain over 60% of intra and inter-species expression variability in differentially expressed genes in primates, showing that H3K27ac and H3K4me3 in promoters and H3K36me3 in genic enhancers are the most informative histone marks.

Interestingly, we could distinguish that different patterns of expression changes implicate different changes in activity in promoters, enhancers or the whole architecture associated to changes in the binding of different histone marks. For instance, human-specific up-regulated genes are mostly associated to changes in H3K36me3 in genic enhancers, while down-regulated ones involves a stronger regulatory changes associated to removal of H3K27ac, H3K4me1 and/or H3K36me3 or of increase of H3K27me3 in the whole architecture. We have established an experimental and computational framework revealing activity-dependent evolutionary constraints and what changes in the epigenomic activity of the gene regulatory architectures are at the source of gene expression changes in the primate lineage.

Translational Bioinformatics

VULCANSPOT: A TOOL TO PRIORITIZE THERAPEUTIC VULNERABILITIES IN CANCER

Javier Perales-Patón[1], Tomás Di Domenico[1], Coral Fustero-Torre[1], Elena Piñeiro-Yáñez[1], Carlos Carretero-Puche[1], Héctor Tejero[1], Alfonso Valencia[2], Gonzalo Gómez-López[1] and Fátima Al-Shahrour[1]

[1]*Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain*
[2]*Barcelona Supercomputing Center, BSC, Spain*

Abstract: Genetic alterations lead to tumour progression and cell survival in cancer, but also uncover genetic dependencies such as oncogenic dependencies and synthetic lethals, which could be exploited to extend the current catalog of molecularly matched treatments for precision medicine [1]. VulcanSpot is a novel computational approach that exploits the notion of collateral therapeutic vulnerabilities caused by the acquisition of cancer mutations. To this aim, our method mines genomic profiles from ~1,500 cancer cell lines and genome-wide gene fitness screening by RNAi and CRISPR technologies to identify potential vulnerabilities in cancer. Then, vulcanSpot prioritizes drugs to target genotype-selective gene dependencies using a weighted scoring system that integrates two complementary strategies for computational drug prescription: i) Pandrugs - a comprehensive database of known gene-drug relationships [2], and ii) a novel drug repositioning method that matches drugs whose transcriptional signature mimics the functional depletion of the target gene while arranges those drugs interacting closer to the vulnerable spot in a Drug-Protein-Protein Interaction Network.

VulcanSpot outcome is in agreement with nominal targeted therapies on oncogenic dependencies (e.g. dabrafenib/vemurafenib in mutant BRAF melanoma tumours, ERBB2 inhibitors such as CP724714 in mutant ERBB2 breast cancer and pictilisib in mutant PIK3CA breast cancer). Moreover, VulcanSpot extends with at least one therapeutic alternative the ~20% of the current undruggable cancer driver genes. Notably, synthetic lethals relationships from the literature were also identified such as CDKN2A-CDK4/6 & SMARCA4-SMARCA2, for which known and novel repurposed targeting compounds are proposed by our tool [3]. VulcanSpot is open-source and has been implemented as publicly accessible web tool at www.vulcanspot.org.

References:

[1] Brunen D and Bernard R (2017). Exploiting synthetic lethality to improve cancer therapy. Nat Rev Clin Oncol. 14(6):331-332.
[2] Piñeiro-Yáñez E et al. (2018). PanDrugs: a novel method to prioritize anticancer drug treatments according to individual genomic data. Genome Med. 10(1):41.
[3] Beijersbergen RL, Wessels LFA and Bernard R. Synthetic Lethality in Cancer Therapeutics. Annual Review of Cancer Biology 2017 1:1, 141-161

MUSCLE INVASIVE BLADDER CANCER STRATIFICATION BY GENOMIC ARCHITECTURE

Sonia González-Alvaredo[1], David Juan[2], Miguel Vázquez[3], Alfonso Valencia[3], Francisco X. Real[1] andEnrique Carrillo-De Santa Pau[4]

[1]*Epithelial Carcinogenesis Group. Spanish National Cancer Research Centre, CNIO, Spain*
[2]*Institute of Evolutionary Biology, UPF-CSIC, Spain*
[3]*Barcelona Supercomputing Centre, BSC, Spain*

[4]*IMDEA Food Institute, Spain*

Abstract: Bladder cancer (BC) is the 4th most common cancer in men and the 11th most common in women. Muscle Invasive bladder cancer (MIBC) is the most aggressive tumor subtype (Shah, McConkey, and Dinney 2011), characterized by genomic instability, copy number alterations and loss-of-function mutations involving TP53 and RB1, among others (Cordon-Cardo 2004). Molecular classifications for MIBC have been proposed including 2-6 different molecular subgroups (Sjödahl et al. 2012; Lindgren et al. 2012; Damrauer et al. 2014; Volkmer et al. 2012; Choi et al. 2014; Rebouissou et al. 2014; Cancer Genome Atlas Research Network 2014; Robertson et al. 2017a; Sjödahl et al. 2017) but there is no consensus regarding the optimal classifier (Lerner et al. 2016).

Genome architecture determines gene expression and it can impact disease through different mechanisms such as copy number variation (Spielmann, Lupiáñez, and Mundlos 2018), mutations (Gerstung et al. 2015; Fredriksson et al. 2014) or epigenetics (Wang et al. 2014; Wagner et al. 2014). Tumor molecular classifications in most cases are exclusively based on expression profiles without consideration of the genome architecture and mechanisms that originate the expression profiles. Therefore, the huge amount of data generated in the last years gives us the opportunity to deepen in molecular stratification from a different perspective.

We here aimed to perform a comprehensive integrative analysis of genome architecture landscape in MIBCs. We took advantage of the rich genomic dataset generated by TCGA for MIBC (Robertson et al. 2017b). We built an innovative data analysis framework to integrate copy number, mutation and methylation data to summarize all the information for each gene and sample in a unique categorical variable and apply clustering analysis to explore the molecular taxonomy of MIBC. We decided to apply a vectorial method to represent the individual tumors in a high dimensional Euclidean space using Multiple Correspondence Analysis (MCA)(Greenacre and Blasius 2006; Hjellbrekke 2005; Torres-Lacomba 2006). MCA cluster is an unsupervised way to stratify samples allowing us to select the key genes and their molecular alterations, associated to the underlying molecular classification of MIBC.

The application of this innovative approach allowed us to identify 3 clusters of MBIC samples with specific clinical and molecular characteristics. Cluster 1 is characterized by genes with non-synonymous mutations, neutral copy number variations, papillary phenotype and lower clinical stages. Cluster 2 by genome deletions, non-papillary phenotypes and high clinical stages. Cluster 3 by genome gains, non-papillary phenotypes and high clinical stages. Our genome architecture perspective for molecular taxonomy provides distinct insights for the understanding of the molecular events leading to urothelial tumorigenesis.

References:
Cancer Genome Atlas Research Network. 2014. Nature 507 (7492): 315–22.
Choi, Woonyoung, et al. 2014. Cancer Cell 25 (2): 152–65.
Cordon-Cardo, Carlos. 2004. Journal of Clinical Oncology: 22 (6): 975–77.
Damrauer, J. S., et al. 2014. Proceedings of the National Academy of Sciences 111 (8): 3110–15.
Fredriksson, Nils J., Lars Ny, Jonas A. Nilsson, and Erik Larsson. 2014. Nature Genetics 46 (12): 1258–63.
Gerstung, Moritz, et al. 2015. Nature Communications 6 (January): 5901.
Greenacre, Michael, and Jorg Blasius. 2006. Multiple Correspondence Analysis and Related Methods. CRC Press.
Hjellbrekke, Johs. 2005. European Sociological Review 21 (5): 529–31.
Lerner, Seth P., et al.. 2016. Bladder Cancer 2 (1): 37–47.
Lindgren, David, et al. 2012. PloS One 7 (6): e38863.
Rebouissou, Sandra, Isabelle et al. 2014. Science Translational Medicine 6 (244): 244ra91.
Robertson, A. Gordon, J et al. 2017a. “Cell 171 (3): 540–56.e25. 2017b. Cell 171 (3): 540–56.e25.
Shah, Jay B., David J. McConkey, and Colin P. N. Dinney. 2011. Clinical Cancer Research: 17 (9): 2608–12.
Sjödahl, Gottfrid, Pontus Eriksson, Fredrik Liedberg, and Mattias Höglund. 2017. Journal of Pathology 242 (1): 113–25.
Sjödahl, Gottfridet al. 2012.” Clinical Cancer Research: 18 (12): 3377–86.
Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. Nature Reviews. Genetics 19 (7): 453–67.

Torres-Lacomba, Anna. 2006. In Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences, 421–32.
Volkmer, et al. 2012. PNAS 109 (6): 2078–83.
Wagner, James R., et al. 2014. Genome Biology 15 (2): R37.
Wang, Fang, et al. 2014. Briefings in Bioinformatics 15 (6): 1028–43.

THE IMMUNOGENIC IMPACTS OF SPLICING ALTERATIONS IN SMALL CELL LUNG CANCER

Juan Luis Trincado Alonso[1], Marina Reixachs[1], Eduardo Eyras[1] and Jun Yokota[2]

[1]*Universitat Pompeu Fabra, UPF, Spain*
[2]*IMPPC, Japan*

Abstract: We describe a novel approach for the exhaustive identification of neo-epitopes from tumor-specific splicing alterations, including aberrant spliced junctions, retained introns and exonizations. Using mass spectrometry for MHC-I associated proteins, we show that splicing derived neo-epitopes are processed and presented by MHC-I complexes. We applied this method to a cohort of 123 small cell lung cancer patients (SCLC) and found that tumor-specific splicing alterations more frequently eliminate than create epitopes, hence uncovering a new mechanism of immune escape in SCLC.

NETWORK, TRANSCRIPTOMIC AND GENOMIC FEATURES DIFFERENTIATE GENES RELEVANT FOR DRUG RESPONSE

Janet Piñero[1], Abel Gonzalez-Perez[2], Emre Guney[1], Joaquim Aguirre-Plans[1], Ferran Sanz[1], Baldo Oliva[1] and Laura I. Furlong[1]

[1]*GRIB (IMIM-UPF), Spain*
[2]*Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Barcelona, Spain*

Abstract: Understanding the mechanisms underlying drug therapeutic action and toxicity is crucial for the prevention and management of drug adverse reactions, and paves the way for a more efficient and rational drug design. The characterization of drug targets, drug metabolism proteins, and proteins associated to side effects according to their expression patterns, their tolerance to genomic variation and their role in cellular networks, is a necessary step in this direction. In this contribution, we hypothesize that different classes of proteins involved in the therapeutic effect of drugs and in their adverse effects have distinctive transcriptomics, genomics and network features. We explored the properties of these proteins within global and organ-specific interactomes, using multi-scale network features, evaluated their gene expression profiles in different organs and tissues, and assessed their tolerance to loss-of-function variants leveraging data from 60K subjects. We found that drug targets that mediate side effects are more central in cellular networks, more intolerant to loss-of-function variation, and show a wider breadth of tissue expression than targets not mediating side effects. In contrast, drug metabolizing enzymes and transporters are less central in the interactome, more tolerant to deleterious variants, and are more constrained in their tissue expression pattern. Our findings highlight distinctive features of proteins related to drug action, which could be applied to prioritize drugs with fewer probabilities of causing side effects.

FUNDING: We received support from ISCIII-FEDER (CP10/00524, and CPII16/00026), IMI-JU under grant agreements no. 116030 (TransQST) and no. 777365 (eTRANSAFE) resources of which are composed of financial contribution from the EU-FP7 (FP7/2007-2013) and EFPIA companies in kind contribution, and the EU H2020 Program 2014–2020 under grant

agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and was supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. The DCEXS is a "Unidad de Excelencia María de Maeztu", funded by the MINECO (ref: MDM-2014-0370).

TRAINING IBM WATSON WITH MELANOMAMINE

Davide Cirillo[1], Andres Cañada[2], Javier Omar Corvi[1], José M. Fernández[1], Jose Antonio Lopez-Martin[3], Salvador Capella-Gutiérrez[1], Martin Krallinger[1] and Alfonso Valencia[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Spanish National Cancer Research Center, CNIO, Spain*
[3]*Research Institute Hospital 12 de Octubre, i+12, Spain*

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

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

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

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

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

Special Session: ELIXIR-ES

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

Audald Lloret-Villas[1] and Jordi Rambla[1]

[1]*Centre for Genomic Regulation, CRG, Spain*

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

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

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

OPENEBENCH. THE ELIXIR PLATFORM FOR BENCHMARKING

Salvador Capella-Gutiérrez[1], Juergen Haas[2], Vicky Sundesha[1], Dmitry Repchevski[1], Javier Garrayo[1], Víctor Fernández-Rodríguez[1], Miguel Madrid[3], Laia Codo[1], José M. Fernández[1], Anália Lourenço[4], J.L. Gelpí[1] and Alfonso Valencia[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Swiss Institute of Bioinformatics (SIB), Switzerland*
[3]*Centre de Recherches en Cancérologie de Touluse (CRCT), France*
[4]*Universidad de Vigo (ESEI), Spain*

Abstract: Benchmarking is intrinsically referred to in many aspects of everyday life from assessing the quality of stock market predictions to weather forecasting to predictions in the life sciences, such as 3D protein structure predictions. On an abstract level, benchmarking is comparing the performance of software under controlled conditions. Benchmarking encompasses the technical performance of individual tools, servers and workflows, including software quality metrics, as well as their scientific performance in predefined challenges. Scientific communities are responsible for defining reference datasets and metrics, reflecting those scientific challenges. In the context of the H2020 ELIXIR-EXCELERATE project, we have developed the OpenEBench platform aiming at transparent performance comparisons across the life sciences. OpenEBench supports the scientific communities by assisting in setting up emerging benchmarking efforts, foster exchange between communities and ultimately aims at making benchmarking not only more transparent, but also more efficient.

We will present the current OpenEBench and a preview on the upcoming implementation, which will be strongly focused on assisting communities to join the platform. Current implementation covers the widgets gallery, which has

been developed to summarize and export OpenEBench data to other platforms; the experts to non-experts visual transformation of scientific benchmarking results, and the assessment of quality metrics for the technical monitoring of bioinformatics resources. For OpenEBench we are working in three levels of operation: level 1 aims to collect and distribute data from established benchmarking communities, via the OpenEBench API; level 2 is based on computing benchmarking metrics within the platform; while level 3 will extend the existing OpenEBench platform to execute benchmarkable workflows (provided as software containers) using identical conditions to ensure an unbiased technical and scientific assessment. Overall, OpenEBench provides an integrated platform to orchestrate benchmarking activities, from the deposition of reference data to test software tools, to the provision of results employing metrics defined by scientific communities.

SPONSORS

PLATINUM

PerkinElmer
For the Better

bioexcel

Fundación InstitutoRoche

FUJITSU

GOLD

Atos

SILVER

SIE

INSTITUTIONAL

INB Spanish National Bioinformatics Institute

elixir SPAIN

ibs.GRANADA
INSTITUTO DE INVESTIGACIÓN BIOSANITARIA

FEDER
Fondo Europeo de Desarrollo Regional
UNIÓN EUROPEA
"Una manera de hacer Europa"

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Instituto de Salud Carlos III

POSTERS

1

José Mg Izarzugaza[1], Søren Brunak[2] and Danish Pangenome Consortium[2]

[1]*DTU Bioinformatics, Denmark*
[2]*NNF Center for Protein Research, Denmark*

Genome Denmark: Sequencing and de novo assembly of the Danish reference genome

Abstract: Hundreds of thousands of human genomes are now being sequenced to characterize genetic variation and use this information to augment association mapping studies of complex disorders and other phenotypic traits Genetic variation is identified mainly by mapping short reads to the reference genome or by performing local assembly. However, these approaches are biased against discovery of structural variants and variation in the more complex parts of the genome. Hence, large-scale de novo assembly is needed. Here we show that it is possible to construct excellent de novo assemblies from high-coverage sequencing with mate-pair libraries extending up to 20 kilobases. We report de novo assemblies of 150 individuals (50 trios) from the GenomeDenmark project. The quality of these assemblies is similar to those obtained using the more expensive long-read technology. We use the assemblies to identify a rich set of structural variants including many novel insertions and demonstrate how this variant catalogue enables further deciphering of known association mapping signals. We leverage the assemblies to provide 100 completely resolved major histocompatibility complex haplotypes and to resolve major parts of the Y chromosome. Our study provides a regional reference genome that we expect will improve the power of future association mapping studies and hence pave the way for precision medicine initiatives, which now are being launched in many countries including Denmark.

(Recently published as Maretty et al., Nature 2017)

2

Ana María Rojas[1] and Jesus Blazquez[2]

[1]*CABD-CSIC, Spain*
[2]*CNB-CSIC, Spain*

A novel porkaryotic MMR pathway

Abstract: NucS is an endonuclease firstly identified in Archaea involved in Mismatch repair (MMR) with no structural homology to known MMR factors. By genetic screenings we report [1] that this protein is required for mutation avoidance and anti-recombination, hallmarks of the canonical MMR in the surrogate model Mycobacterium smegmatis, lacking classical MutS-MutL factors. Furthermore, phenotypic analysis of naturally occurring polymorphic NucS in a M. smegmatis surrogate model, suggests the existence of M. tuberculosis mutator strains.

Computational and evolutionary studies of NucS indicate a complex evolutionary emergence based in distinct protein domain fusions followed by at least two horizontal gene transfers leading to a disperse distribution pattern in prokaryotes. Together, these findings indicate that distinct pathways for MMR have evolved at least twice in nature. The analyses of these findings in the evolutionary context of the classical MMR proteins open novel and intriguing questions in the emergence of the MMR systems.

[1]. Castaneda-Garcia, A; Prieto, A.I., Rodriguez-Beltran, J., Alonso, N., Cantillon, D., Costas. C., Perez-Lago, L., Zegeye, E.D.., Herranz, M., Plocinski, P., Tonjum, T., Garcia de Viedma, D., Paget, M, Waddell, S.J., Rojas, A.M.*, Doherty, A.J.*, & Blazquez, J*. (2017) A non-canonical mismatch repair pathway in prokaryotes. Nature Communications (27th January) DOI:10.1038/ncomms14246

3

Fernando Pozo[1] and Michael Tress[1]

[1]*Spanish National Cancer Research Centre (CNIO), Spain*

Machine Learning Predicts the Functional Importance of Potential Alternative Splicing Isoforms

Abstract: The alternative splicing of messenger RNA produces a vast array of mature RNA transcripts, and many of these alternative transcripts are annotated in reference databases. Despite the overwhelming evidence of alternative splicing at the transcript level, there is little reliable evidence of alternative spliced proteins at the protein level, so the extent to which alternative transcripts will produce functionally relevant protein isoforms is still not clear (1-3).

Here we present a new and automatic computational approach to classify human alternative splice isoforms. This machine learning approach uses more than 60 different features from 4 distinct biological categories to make its predictions. The algorithm classified functional isoforms with high confidence and in contrast to previous attempts to predict functional alternative isoforms, predicted that at least 90% of distinct protein isoforms do have not a functional role in the cell.

The algorithm not only provides valuable insights into the functional importance of alternative splicing, but will also provide a reliable list of the most significant biological splice isoforms. The list of functional splice isoforms will be available from the APPRIS database (4) and will be updated with each new GENCODE release (5) and can be applied to any eukaryotic species.

[1]. Tress, M. L., Abascal, F., and Valencia, A. (2017). Most alternative isoforms are not functionally important. Trends in Biochemical Sciences, 42(6), 408 – 410.

[2]. Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J. M., del Pozo, A., Vázquez, J., Valencia, A., and Tress, M. L. (2015). Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level. PLoS Computational Biology, 11(6), 1–29.

[3]. Tress, M. L., Abascal, F., and Valencia, A. (2017). Alternative Splicing May Not Be the Key to Proteome Complexity. Trends in Biochemical Sciences, 42(2), 98–110.

[4]. Rodriguez,J.M., Rodriguez-Rivas,J., Di Domenico,T., Vázquez,J., Valencia,A., and Tress,M.L. (2018) APPRIS 2017: principal isoforms for multiple gene sets. Nucleic Acids Res., 46, D213-217.

[5]. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S., et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res., 22, 1760-1774.

5

Ángeles Arzalluz-Luque[1] and Ana Conesa[1,2]

[1]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[2]*University of Florida, Spain*

Single-cell RNAseq for the study of isoforms -how is that possible?

Abstract: Single-cell RNAseq and alternative splicing studies have recently become two of the most prominent applications of RNAseq. However, the combination of both is still challenging, and few research efforts have been dedicated to the intersection between them. Cell-level insight on isoform expression is required to fully understand the biology of alternative splicing, but it is still an open question to what extent isoform expression analysis at the single-cell level is actually feasible. In our recent Genome Biology publication, we establish a set of four conditions that are required for a successful single-cell-level isoform study, and evaluate how these conditions are met by both current experimental and computational single-cell methods. We use real data to assess the theoretical limits of each method and provide considerations for experimental design. Finally, we review published findings in single-cell isoform biology, and provide insight for future development of the field.

6

Ouissam El Andaloussi[1], Mhamed Ait Kbir[2] and Badr Din Rossi Hassani[1]

[1]*LABIPHABE, Morocco*
[2]*LIST, Morocco*

Biological data annotation using complementary and alternative Medicine with a collection tracking system (Hirbalink)

Abstract: The web contains huge volume of information related to complementary and alternative medicine. However, healthcare recommendation with medicinal plants has become complicated because tremendous and precious information about medicinal resources are available now. Moreover, the existing scientific search engines are not quite efficient and require excessive manual processing. As a result, the search for accurate and reliable data about herbal plants has become a highly difficult and time-consuming task for scientists. Till date, a wide mapping of already available data concerning herbal plants hasn't been carried out. In this regard, the complementary and alternative medicine collection tracking system (Hirbalink) introduced in this work was created for the purpose of organizing and storing related data.

7

Manuel Ugidos[1], Sonia Tarazona[1], José Manuel Prats[2], Alberto Ferrer[2] and Ana Conesa[1,3]

[1]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[2]*Universitat Politècnica de València, UPV, Spain*
[3]*University of Florida, United States*

MultiBaC: A strategy to remove batch effect between different omic data types.

Abstract: Diversity of omic technologies has expanded in the last years together with the number of omics integration strategies.However, the costs of the different techniques are still high and many research groups cannot afford research

projects where many different omics techniques are analyzed. Nevertheless, as most research share their data in public repositories, there is a possibility of utilization of datasets from other laboratories to construct a multiomic study. An important issue when we want to integrate data from different studies is the batch effect. There are already several methods described which are able to correct batch effect on common omic data between different studies but they cannot be used to correct no common data (i.e. the omic data modality that has been analyzed at only one lab). We have developed MultiBaC, a strategy to correct batch effect on no common omic information which let us integrate different omic data types from different studies. Our results offers a new possibility for combining multi-layered datasets from different reach laboratories to expand of understanding of complex biological systems.

8

Kevin Troulé[1], Miguel Reboiro-Jato[2], Hector Tejero[1], Hugo Lopez-Fernandez[2], Javier Perales-Patón[1], Daniel Glez-Peña[2], Fátima Al-Shahrour[1] and Gonzalo Gómez-López[1]

[1]*Spanish National Cancer Research Centre, CNIO, Spain*
[2]*Universidade de Vigo, Spain*

The druggable immune system: drug repositioning in immune transcriptome.

Abstract: PURPOSE Immune cells can control the fate of tumor, either promoting its growth or diminishing it (1). The goal of this work is finding drugs that can promote a better immune-environment to improve responses to antitumoral treatments. For instance, avoiding the effect of immunosuppressive M2 macrophages, mieloid-derived supressor cells (MDSCs) or regulatory T cells (Tregs) cells by reverting their signatures towards non inmunosupressive states.

METHOD We have employed an in-house version of Connectivity Map (2) to predict single drug treatments and revert expression signatures integrating data from L1000, CCLE, GDSC2.0 and CTRP projects, and comprising more than 5,000 compounds and ~4 million drug-drug interactions. We have applied this approach to study 156 selected immunologic gene expression signatures associated to T-reg, T-helper, MDSC and macrophages obtained from MSigDB (3) and scientific literature. The analysis has been performed both for human and mouse signatures.

RESULTS: Using our methodology we have obtained at least one significant signature-drug prediction for 44% of the immune signatures. In total, we obtained 5,472 immune gene expression signature-drug interactions corresponding to 1,081 drugs (FDR < 0.05). To validate our approach, we have manually reviewed some of our predictions checking scientific literature. For instance as previously reported, we predict that PI3K inhibitors (i.e.wortmannin), suppress Treg activity.

CONCLUSIONS We have built a catalogue of prioritized drug predictions targeting highly relevant immune cells involved in tumor's fate that can either promote or revert a given immune signature. Our approach can be extended to predict drug treatments in other curated immunological studies. The final goal is the creation of a database containing predictions of immune signature-drugs interaction that potentially may revert immune cellular states.

[1] 10.1016/j.cell.2010.01.025
[2] 10.1126/science.1132939
[3] 10.1016/j.cels.2015.12.004

9

Salvador Casani[1], Cecile Pereira[2] and Ana Conesa[3]

[1]*Biobam Bioinformatics, Spain*
[2]*Eura Nova, France*
[3]*University of Florida, United States*

Combining databases and text-mining for biological pathway reconstruction

Abstract: Pathway databases capture the established knowledge on molecular interactions that take place in organisms. Such pathways are the result of an extensive curation process, and depend on the curator´s decisions to define boundaries of the described biological processes. Pathways databases are frequently used for the interpretative analysis of high throughput omics data, as they provide a standardised way to summarise molecular information and provide a biologically meaningful output. Tools like functional enrichment analysis have been developed to make such queries through a robust statistical formulation.

However, pathway databases are by definition incomplete. New knowledge is constantly generated from research projects and reported in the scientific literature. It will normally take years before this dynamic new knowledge in incorporated into the structured databases. Moreover, specific scientific areas might be poorly represented in regular pathway databases, due to either they target rare processes, or they unravel new connections between processes that were not previously studied jointly (e.g. the novel connections between the gut microbiome and the brain function, or the impact of metabolism in chromatin modifications). However, these connections are potentially retrievable from scientific papers.

In this work we propose a novel strategy to infer novel pathways from the scientific literature by text mining approaches. Given a user-provided research domain, we identify the network of genes, protein and metabolites and their relationships that captures the most current state of the art and make this information available for statistical analysis by functional profiling tools.

10

Álvaro Andrades[1,2], Isabel F. Coira[1,2], María I. Rodríguez[1,2], Pedro Carmona-Sáez[2], Javier De Las Rivas[3], Marta Cuadros[1,2] and Pedro P Medina[2,1]

[1]*University of Granada, Granada, Spain*
[2]*Centre for Genomics and Oncological Research, GENYO, Granada, Spain*
[3]*Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), CSIC and University of Salamanca, Salamanca, Spain*

Novel pipeline for prioritizing long non-coding RNA mutations in lung adenocarcinoma

Abstract: Long non-coding RNAs (lncRNAs) are RNA molecules that do not code for protein and have a length over 200 bp. Over the last decades, increasing evidence suggests that many lncRNAs play critical roles in most, if not all, hallmarks of cancer. However, very little is known about the specific alterations that lncRNAs undergo in cancer or about the specific mechanisms by which such alterations affect cancer development. In particular, although alterations in lncRNA expression levels have been thoroughly studied, few putative driver somatic mutations in lncRNAs have been identified and even fewer have been fully characterized yet. This is in part because of the scarcity of computational tools or pipelines for prioritizing somatic mutations in lncRNAs. In this context, we aimed to develop a novel computational pipeline that prioritizes somatic mutations in lncRNAs in order to obtain a small subset of candidate driver mutations for experimental validation. For this purpose, we developed a somatic mutation calling pipeline that combines three state-of-the-art algorithms (Mutect2, Strelka and VarScan 2) in order to obtain high-confidence somatic mutations. In addition, we are designing a pipeline for prioritizing the high-confidence mutations that integrates

genomic and transcriptomic information. We are evaluating our pipeline in data from tumor-normal paired samples of lung adenocarcinoma (LUAD) from two sources: (i) targeted sequencing data from an internal cohort of 27 lung adenocarcinoma patients; and (ii) whole-genome sequencing and RNA-sequencing data from 59 patients from The Cancer Genome Atlas (TCGA-LUAD). Our preliminary results show that our combined somatic mutation calling pipeline is able to discern mutations that are likely to be true somatic mutations. In addition, after combining high-confidence mutation data and a RNA-Seq differential expression analysis, we detected between 8 and 27 somatic mutations that were recurrent in both analyzed cohorts and that affected lncRNAs that were differentially expressed between tumor and normal samples in TCGA-LUAD. The affected lncRNAs included HOTAIR, FENDRR, AFAP1-AS1 and LINC01833, among others. In the near future, we aim to refine the prioritization pipeline and extend our analysis to a larger LUAD cohort. Overall, our method may help researchers to discover true driver mutations in lncRNAs in LUAD and in other tumors, expanding our knowledge on the molecular mechanisms of cancer and allowing for the development of novel procedures for the diagnosis, prognosis or treatment.

Acknowledgements: Genomics Unit and Bioinformatics Unit at GENYO and Bioinformatics Unit at CiC-IBMCC. Funding: SAF2015-67919-R project (MINECO).

11

Juan R Gonzalez[1], Alejandro Caceres[1] and Carlos Ruiz[1]

[1]*Barcelona Institute for Global Health, ISGlobal, Spain*

Genetic inversions in complex diseases

Abstract: Inversions are a common class of structural variants that reverse the orientation of a segment of the genome, and have been shown to have phenotypic consequences in different organisms. However, there is a lack of knowledge on the role of genomic inversions and how they interact with the environment in the aetiology of common complex diseases. The main limitation to perform inversion association studies is the lack of a high-throughput method to genotype inversions on a large scale. Our group has developed several Bioinformatic tools for both detecting and call polymorphic genetic inversions using SNP array data. This has facilitated the reanalysis of existing SNP array data from GWAS studies and assess the impact of inversions in complex diseases. In this talk we will introduce scoreInvHap that is a scalable and efficient method to call genotype inversions from SNP array data. The re-analysis of existing GWAS data has shown significant associations between inversions on 8p23 and 17q21 in different brain-related diseases/disorders such as Autism, neuroticism and risk behavior phenotypes. We have also found an association between inversion 16p11 and asthma/obesity in humans. The analysis of GTEx data also provides significant associations between several inversions and different tissues. Our results also provide pieces of evidence that haplotypes defined by genomic inversions have ancestral origins with out-of-Africa expansions in different human population and traces of selection that could be related to environmental factors such as climate, life style or diet.

13

Jordi Martorell-Marugán[1,2], Víctor González-Rumayor[2] and Pedro Carmona-Sáez[1]

[1]*Pfizer-University of Granada-Junta de Andalucía Centre for Genomics and Oncological Research, GENYO, Spain*
[2]*Atrys Health, Spain*

mCSEA: Detecting differentially methylated regions by enrichment analysis

Abstract: The identification of differentially methylated regions (DMRs) among phenotypes is one of the main goals of epigenetic analysis. Although there are several methods developed to detect DMRs, most of them are focused on detecting relatively large differences in methylation levels and fail to detect moderate, but consistent, methylation changes that might be associated to complex disorders. We present mCSEA, an R package that implements a Gene Set Enrichment Analysis method to identify differentially methylated regions from Illumina 450K and EPIC array data. It is especially useful for detecting subtle, but consistent, methylation differences in complex phenotypes. mCSEA also implements functions to integrate gene expression data and to detect genes with significant correlations among methylation and gene expression patterns. Using simulated data and different previously published datasets we show that mCSEA outperforms other tools in detecting DMRs, and is able to identify significant and biologically meaningful differentially methylated regions not detected by other methods. mCSEA is freely available from the Bioconductor repository.

14

Carlos Ruiz Arenas[1], Carles Hernandez-Ferrer[1], Mariona Bustamante[1], Marta Vives-Usano[2], Angel Carracedo[3], Eulàlia Martí[2], Martine Vrijheid[1] and Juan R Gonzalez[1]

[1]*Barcelona Institute for Global Health, ISGlobal, Spain*
[2]*Center for Genomic Regulation, CRG, Spain*
[3]*Grupo de Medicina Xenómica - Universidade de Santiago de Compostela, Spain*

Association between blood DNA methylation and gene expression in children

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

15

Sonia Tarazona[1], David Gómez-Cabrero[2], Andreas Schmidt[3], Axel Imhof[3], Thomas Hankemeier[4], Jesper Tegnér[5], Johan A. Westerhuis[6] and Ana Conesa[1,7]

[1]*Centro de Investigacion Principe Felipe, CIPF, Spain*
[2]*Navarrabiomed, Spain*
[3]*Ludwig Maximilian University of Munich, Germany*
[4]*Leiden/Amsterdam Center for Drug Research, Netherlands*
[5]*King Abdullah University of Science and Technology, Saudi Arabia*
[6]*University of Amsterdam, Netherlands*
[7]*University of Florida, United States*

Harmonization of quality metrics and power calculation in multi-omic studies

Abstract: Multi-omic studies combine measurements at different molecular levels to build comprehensive models of cellular systems. The success of a multi-omic data analysis strategy depends largely on the adoption of an adequate experimental design and on the quality of the measurements provided by the different omic platforms. However, the field lacks a comparative description of performance parameters across omic technologies and a formulation for experimental design in multi-omic data scenarios.

In this study, we propose a set of harmonized Figures of Merit (FoM) as quality descriptors applicable to different omic data-types, including sequence and mass-spectrometry data. We identify FoM that contribute to determining data variability and omic dimensionality as critical components for statistical power calculations.

We introduce the novel MultiPower method to estimate the optimal sample size in a multi-omics experiment and to assess the final statistical power of each omic dataset, which is critical to trust in statistical analysis results. MultiPower supports different experimental settings, including discrete or continuous data, as well as equal or different sample size per omic, and incorporates practical tools to facilitate informed design decisions in multi-omic experiments.

To illustrate MultiPower usage, we apply the MultiPower method to two multi-omic data sets with different characteristics to assess the statistical power provided by the available sample sizes. The first is the STATegra data set, a multi-omic experiment with 3 replicates per condition and 6 omic data types; and the second is a TCGA cohort study, with between 22 and 155 samples per condition and 4 omic data types. MultiPower returned interesting results on how powerful each data type was, which the optimal sample size was and how data variability affects power in each case.

16

Sara Monzón[1], Luis Chapado[1], Jose Luis Garcia-Pacheco[1], Miguel Juliá[1], Pedro J. Sola-Campoy[1], Ángel Zaballos[1] and Isabel Cuesta[1]

[1]*Institute of Health Carlos III (ISCIII)*

iSkyLIMS, a friendly environment to facilitate the incorporation of High Throughput Sequencing into a genomics core facility.

Abstract: Background: The introduction of high throughput sequencing (HTS) in genomics facilities has meant an exponential growth in data generation, requiring a precise tracking system, for every step, from library preparation to fastq file generation, data analysis and delivery to the researcher. Software designed to handle those tasks are called Laboratory Information Management Systems (LIMS), but it has to be adapted to the genomics laboratory particular needs.

iSkyLIMS is born with the aims to assist with wet laboratory tasks and implements a workflow to guide genomics labs in their sequencing routines reducing potential errors associated to high throughput technology and easing the quality control of the sequencing. Also, iSkyLIMS connects the wet lab with dry lab environments, allowing straight forward bioinformatics data analysis.

Materials/methods: iSkyLIMS has been implemented with Django Framework 2.0 running on Python 3.6 and uses a MySQL database to store the processed data generated by the Illumina sequencer.

Results: iSkyLIMS is an open-source software that run on a linux machine with a web based interface for user-friendly interaction. The sequencing runs inside iSkyLIMS are handled in a state transitions machine where each run is passing through, recorded, sample sent, bcl2fastq executed, and completed state. Run data from Illumina is fetched, processed and stored in a database, which will provide input for quality control analysis, statistics information, and reports done afterwards. The information is searchable by run, sample, project or user. This software has been designed specifically to address sample sheet generation required to start Illumina NextSeq sequencer through BaseSpace Illumina web service, but since our genomics facility has a MiSeq, it is being adapted to support others MS illumina platforms.

Conclusions: iSkyLIMS integrates both steps, sequencing and data analysis, in one automatized efficient workflow, with minimal manual interaction, customized to genomics laboratory's needs. Keeping all sequencing information centralized and running on a virtual environment, iSkyLIMS becomes a scalable solution able to fulfill future needs, since the number of runs are exponentially increasing with the incorporation of HTS to the clinical research laboratory routine. iSkyLIMS is free and available for downloading at https://github.com/BU-ISCIII/iSkyLIMS

17

Jordi Martorell-Marugán[1], Raúl López-Domínguez[1], Daniel Toro-Domínguez[1], Adrián García-Moreno[1], Víctor González-Rumayor[2], Joaquín Dopazo[3], Adoración Martín-Gómez[4], Marta Eugenia Alarcón-Riquelme[1] and Pedro Carmona-Sáez[1]

[1]*GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, Spain*
[2]*Atrys Health S.A., Spain*
[3]*Fundacion Progreso y Salud, Sevilla, Spain*
[4]*Unidad de Nefrología, Hospital de Poniente, Spain*

A centralized database for exploring gene expression and methylation data in Systemic Autoimmune Diseases

Abstract: Systemic autoimmune diseases (SADs) are a group of complex and heterogeneous diseases characterized by immune responses to self-antigens leading to tissue damage and dysfunction in several organs. The pathogenesis of SADs is not fully understood, but both environmental and genetic factors have been linked to their development(Salaman, 2003).

In the last few years, the use of –omics technologies in this field has provided new insights into the molecular mechanisms associated to the development of SADs, opening new scenarios for biomarker discovery and treatments development (Kim et al., 2014). A remarkable example is the characterization of the type I interferon gene expression signature as a key factor in the onset of some autoimmune-related diseases, especially in systemic lupus erythematosus (Crow, 2014).

Nevertheless, although several cohorts and studies have been published and –omics datasets are publicly available, there is a lack of a common resource that facilitates the exploration, comparison and integration of this disseminated information. In this context, we have developed a novel tool named ADEx: Autoimmune Disease Data Explorer, in which we have compiled and processed most of the publicly available gene expression and methylation datasets from Gene

Expression Omnibus (GEO) database (Edgar et al., 2002) for systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), Sjögren's syndrome (SjS) and Systemic sclerosis (SSc) diseases.

ADEx has been developed with a set of useful functionalities to explore gene expression and methylation profiles at a gene or study level. Researchers can explore gene expression variation across diseases or methylation values in CpG sites from individual genes. In addition, potential biomarkers can be explored from an integrative analysis of methylation and expression data and gene signatures can be interpreted in the context of KEGG pathways or signaling networks through integration with HiPathia software (Hidalgo et al., 2017).

The application also implements meta-analysis functionalities to integrate and jointly analyze different and heterogeneous datasets. This is a relevant feature that can be applied in order to define common gene signatures and biomarkers across different pathologies (Toro-Domínguez et al., 2018).

We are totally confident that this new resource will provide to the research community with a unique data portal that will support and promote research in the area of SADs. ADEx is freely accessible at http://bioinfo.genyo.es/adex/.

References

Crow,M.K. (2014) Type I Interferon in the Pathogenesis of Lupus. J Immunol, 192, 5459–5468.

Edgar,R. et al. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, 30, 207–210.

Hidalgo,M.R. et al. (2017) High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget, 8, 5160–5178.

Kim,H.-Y. et al. (2014) Advances in Systems Biology Approaches for Autoimmune Diseases. Immune Netw, 14, 73–80.

Salaman,M.R. (2003) A two-step hypothesis for the appearance of autoimmune disease. Autoimmunity, 36, 57–61.

Toro-Domínguez,D. et al. (2018) ImaGEO: Integrative Gene Expression Meta-Analysis from GEO database. Bioinformatics.

#18

Javier Perales-Patón[1], Tomás Di Domenico[1], Coral Fustero-Torre[1], Elena Piñeiro-Yáñez[1], Carlos Carretero-Puche[1], Héctor Tejero[1], Alfonso Valencia[2], Gonzalo Gómez-López[1] and Fátima Al-Shahrour[1]

[1]*Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain*
[2]*Barcelona Supercomputing Center, BSC, Spain*

vulcanSpot: a tool to prioritize therapeutic vulnerabilities in cancer

Abstract: Genetic alterations lead to tumour progression and cell survival in cancer, but also uncover genetic dependencies such as oncogenic dependencies and synthetic lethals, which could be exploited to extend the current catalog of molecularly matched treatments for precision medicine [1]. VulcanSpot is a novel computational approach that exploits the notion of collateral therapeutic vulnerabilities caused by the acquisition of cancer mutations. To this aim, our method mines genomic profiles from ~1,500 cancer cell lines and genome-wide gene fitness screening by RNAi and CRISPR technologies to identify potential vulnerabilities in cancer. Then, vulcanSpot prioritizes drugs to target genotype-selective gene dependencies using a weighted scoring system that integrates two complementary strategies for computational drug prescription: i) Pandrugs - a comprehensive database of known gene-drug relationships [2], and ii) a novel drug repositioning method that matches drugs whose transcriptional signature mimics the functional depletion of the target gene while arranges those drugs interacting closer to the vulnerable spot in a Drug-Protein-Protein Interaction Network.

VulcanSpot outcome is in agreement with nominal targeted therapies on oncogenic dependencies (e.g. dabrafenib/vemurafenib in mutant BRAF melanoma tumours, ERBB2 inhibitors such as CP724714 in mutant ERBB2 breast cancer and pictilisib in mutant PIK3CA breast cancer). Moreover, VulcanSpot extends with at least one therapeutic alternative the ~20% of the current undruggable cancer driver genes. Notably, synthetic lethals relationships from the literature were also identified such as CDKN2A-CDK4/6 & SMARCA4-SMARCA2, for which known and novel repurposed targeting compounds are proposed by our tool [3]. VulcanSpot is open-source and has been implemented as publicly accessible web tool at www.vulcanspot.org.

References:

[1] Brunen D and Bernard R (2017). Exploiting synthetic lethality to improve cancer therapy. Nat Rev Clin Oncol. 14(6):331-332.

[2] Piñeiro-Yáñez E et al. (2018). PanDrugs: a novel method to prioritize anticancer drug treatments according to individual genomic data. Genome Med. 10(1):41.

[3] Beijersbergen RL, Wessels LFA and Bernard R. Synthetic Lethality in Cancer Therapeutics. Annual Review of Cancer Biology 2017 1:1, 141-161

19

Adrián Muñoz-Barrera[1], Luis A. Rubio-Rodríguez[1], José M. Lorenzo-Salazar[1], Carlos Flores[1,2], Marcos Colebrook[3], Carlos J. Pérez-González[4] and José L. Roda-García[3]

[1]Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Tenerife, Canary Islands, Spain
[2]Hospital N.S. De Candelaria-CIBERES, Spain
[3]Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Tenerife, Canary Islands, Spain
[4]Departamento de Matemáticas, Estadística e Investigación Operativa, Universidad de La Laguna, Tenerife, Canary Islands, Spain

WDL-based pipelines for whole genome and exome sequencing analysis

Abstract: Next Generation Sequencing data analysis comprises a series of computational tasks frequently based on the use of command line tools. These analyses are defined in workflows that group all the necessary tasks, improving data processing performance and results interpretation. Some Domain Specific Languages (DSLs), such as WDL and Nextflow, have been recently created to define and program complex pipelines, as well as to improve the parallelization, the scalability and the reusability.

We have developed a complete pipeline programmed in WDL via scripting and Rabix Composer based on the Broad Institute's best practices and the Genome Analysis Toolkit (GATK4) to analyze whole-genome (WGS) and whole-exome (WES) data. This pipeline is able to process data from one or several WGS/WES of multiple individuals. It allows to parallelize the analysis in multiple tasks and nodes, and runs in environments based on Docker Swarm or in high-performance computing systems such as TeideHPC at ITER (http://teidehpc.iter.es/en/home). Interesting features are: possibility to run on a HPC infrastructure connecting the Cromwell engine and the SLURM scheduler; starts from BCL data; demultiplexing of samples pooled across the flowcell; data processing both on a per-lane and a per-sample basis; possibility to handle hg19 and hg38 reference genomes; programmed to restart from every step in case of fail. For benchmarking, we are following the guidelines of the Truth and Consistency precisionFDA challenges using Genome In A Bottle Consortium released genomes data.

A full pipeline is currently running on TeideHPC to analyze WGS and WES germline data produced by an Illumina HiSeq4000 sequencing platform for research purposes. These resources are available at: https://github.com/genomicsITER/wdl

Funded by Ministerio de Ciencia, Innovación y Universidades (RTC-2017-6471-1; MINECO/AEI/FEDER, UE). This work has been supported by the CEDeI program (Centro de Excelencia de Desarrollo e Innovación, Cabildo de Tenerife). The authors also thankfully acknowledge the computer resources and the technical support provided by TARO Research Group of the University of La Laguna.

20

Sonia González-Alvaredo[1], David Juan[2], Miguel Vázquez[3], Alfonso Valencia[3], Francisco X. Real[1] and Enrique Carrillo-De Santa Pau[4]

[1]*Epithelial Carcinogenesis Group. Spanish National Cancer Research Centre, Madrid, Spain*
[2]*Institute of Evolutionary Biology. UPF-CSIC, Spain*
[3]*Barcelona Supercomputing Centre, BSC, Spain*
[4]*IMDEA Food Institute, Spain*

Muscle Invasive Bladder Cancer stratification by genomic architecture

Abstract: Bladder cancer (BC) is the 4th most common cancer in men and the 11th most common in women. Muscle Invasive bladder cancer (MIBC) is the most aggressive tumor subtype (Shah, McConkey, and Dinney 2011), characterized by genomic instability, copy number alterations and loss-of-function mutations involving TP53 and RB1, among others (Cordon-Cardo 2004). Molecular classifications for MIBC have been proposed including 2-6 different molecular subgroups (Sjödahl et al. 2012; Lindgren et al. 2012; Damrauer et al. 2014; Volkmer et al. 2012; Choi et al. 2014; Rebouissou et al. 2014; Cancer Genome Atlas Research Network 2014; Robertson et al. 2017a; Sjödahl et al. 2017) but there is no consensus regarding the optimal classifier (Lerner et al. 2016).

Genome architecture determines gene expression and it can impact disease through different mechanisms such as copy number variation (Spielmann, Lupiáñez, and Mundlos 2018), mutations (Gerstung et al. 2015; Fredriksson et al. 2014) or epigenetics (Wang et al. 2014; Wagner et al. 2014). Tumor molecular classifications in most cases are exclusively based on expression profiles without consideration of the genome architecture and mechanisms that originate the expression profiles. Therefore, the huge amount of data generated in the last years gives us the opportunity to deepen in molecular stratification from a different perspective.

We here aimed to perform a comprehensive integrative analysis of genome architecture landscape in MIBCs. We took advantage of the rich genomic dataset generated by TCGA for MIBC (Robertson et al. 2017b). We built an innovative data analysis framework to integrate copy number, mutation and methylation data to summarize all the information for each gene and sample in a unique categorical variable and apply clustering analysis to explore the molecular taxonomy of MIBC. We decided to apply a vectorial method to represent the individual tumors in a high dimensional Euclidean space using Multiple Correspondence Analysis (MCA)(Greenacre and Blasius 2006; Hjellbrekke 2005; Torres-Lacomba 2006). MCA cluster is an unsupervised way to stratify samples allowing us to select the key genes and their molecular alterations, associated to the underlying molecular classification of MIBC.

The application of this innovative approach allowed us to identify 3 clusters of MBIC samples with specific clinical and molecular characteristics. Cluster 1 is characterized by genes with non-synonymous mutations, neutral copy number variations, papillary phenotype and lower clinical stages. Cluster 2 by genome deletions, non-papillary phenotypes and high clinical stages. Cluster 3 by genome gains, non-papillary phenotypes and high clinical stages. Our genome architecture perspective for molecular taxonomy provides distinct insights for the understanding of the molecular events leading to urothelial tumorigenesis.

References:

Cancer Genome Atlas Research Network. 2014. “Comprehensive Molecular Characterization of Urothelial Bladder Carcinoma.” Nature 507 (7492): 315–22.

Choi, Woonyoung, Sima Porten, Seungchan Kim, Daniel Willis, Elizabeth R. Plimack, Jean Hoffman-Censits, Beat Roth, et al. 2014. “Identification of Distinct Basal and Luminal Subtypes of Muscle-Invasive Bladder Cancer with Different Sensitivities to Frontline Chemotherapy.” Cancer Cell 25 (2): 152–65.

Cordon-Cardo, Carlos. 2004. “p53 and RB: Simple Interesting Correlates or Tumor Markers of Critical Predictive Nature?” Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 22 (6): 975–77.

Damrauer, J. S., K. A. Hoadley, D. D. Chism, C. Fan, C. J. Tiganelli, S. E. Wobker, J. J. Yeh, et al. 2014. “Intrinsic Subtypes of High-Grade Bladder Cancer Reflect the Hallmarks of Breast Cancer Biology.” Proceedings of the National Academy of Sciences 111 (8): 3110–15.

Fredriksson, Nils J., Lars Ny, Jonas A. Nilsson, and Erik Larsson. 2014. “Systematic Analysis of Noncoding Somatic Mutations and Gene Expression Alterations across 14 Tumor Types.” Nature Genetics 46 (12): 1258–63.

Gerstung, Moritz, Andrea Pellagatti, Luca Malcovati, Aristoteles Giagounidis, Matteo G. Della Porta, Martin Jädersten, Hamid Dolatshad, et al. 2015. “Combining Gene Mutation with Gene Expression Data Improves Outcome Prediction in Myelodysplastic Syndromes.” Nature Communications 6 (January): 5901.

Greenacre, Michael, and Jorg Blasius. 2006. Multiple Correspondence Analysis and Related Methods. CRC Press.

Hjellbrekke, Johs. 2005. “Brigitte Le Roux and Henry Rouanet (with a Foreword by Patrick Suppes): Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis Dordrecht: Kluwer, 2004. 475 Pages, 155 Euros.” European Sociological Review 21 (5): 529–31.

Lerner, Seth P., David J. McConkey, Katherine A. Hoadley, Keith S. Chan, William Y. Kim, François Radvanyi, Mattias Höglund, and Francisco X. Real. 2016. “Bladder Cancer Molecular Taxonomy: Summary from a Consensus Meeting.” Bladder Cancer (Amsterdam, Netherlands) 2 (1): 37–47.

Lindgren, David, Gottfrid Sjödahl, Martin Lauss, Johan Staaf, Gunilla Chebil, Kristina Lövgren, Sigurdur Gudjonsson, et al. 2012. “Integrated Genomic and Gene Expression Profiling Identifies Two Major Genomic Circuits in Urothelial Carcinoma.” PloS One 7 (6): e38863.

Rebouissou, Sandra, Isabelle Bernard-Pierrot, Aurélien de Reyniès, May-Linda Lepage, Clémentine Krucker, Elodie Chapeaublanc, Aurélie Hérault, et al. 2014. “EGFR as a Potential Therapeutic Target for a Subset of Muscle-Invasive Bladder Cancers Presenting a Basal-like Phenotype.” Science Translational Medicine 6 (244): 244ra91.

Robertson, A. Gordon, Jaegil Kim, Hikmat Al-Ahmadie, Joaquim Bellmunt, Guangwu Guo, Andrew D. Cherniack, Toshinori Hinoue, et al. 2017a. “Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer.” Cell 171 (3): 540–56.e25.

2017b. “Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer.” Cell 171 (3): 540–56.e25.

Shah, Jay B., David J. McConkey, and Colin P. N. Dinney. 2011. “New Strategies in Muscle-Invasive Bladder Cancer: On the Road to Personalized Medicine.” Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 17 (9): 2608–12.

Sjödahl, Gottfrid, Pontus Eriksson, Fredrik Liedberg, and Mattias Höglund. 2017. “Molecular Classification of Urothelial Carcinoma: Global mRNA Classification versus Tumour-Cell Phenotype Classification.” The Journal of Pathology 242 (1): 113–25.

Sjödahl, Gottfrid, Martin Lauss, Kristina Lövgren, Gunilla Chebil, Sigurdur Gudjonsson, Srinivas Veerla, Oliver Patschan, et al. 2012. "A Molecular Taxonomy for Urothelial Carcinoma." Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 18 (12): 3377–86.

Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. "Structural Variation in the 3D Genome." Nature Reviews. Genetics 19 (7): 453–67.

Torres-Lacomba, Anna. 2006. "Correspondence Analysis and Categorical Conjoint Measurement." In Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences, 421–32.

Volkmer, Jens-Peter, Debashis Sahoo, Robert K. Chin, Philip Levy Ho, Chad Tang, Antonina V. Kurtova, Stephen B. Willingham, et al. 2012. "Three Differentiation States Risk-Stratify Bladder Cancer into Distinct Subtypes." Proceedings of the National Academy of Sciences of the United States of America 109 (6): 2078–83.

Wagner, James R., Stephan Busche, Bing Ge, Tony Kwan, Tomi Pastinen, and Mathieu Blanchette. 2014. "The Relationship between DNA Methylation, Genetic and Expression Inter-Individual Variation in Untransformed Human Fibroblasts." Genome Biology 15 (2): R37.

Wang, Fang, Shaojun Zhang, Yanhua Wen, Yanjun Wei, Haidan Yan, Hongbo Liu, Jianzhong Su, Yan Zhang, and Jianhua Che. 2014. "Revealing the Architecture of Genetic and Epigenetic Regulation: A Maximum Likelihood Model." Briefings in Bioinformatics 15 (6): 1028–43.

21

Claudia Arnedo-Pac[1], Loris Mularoni[1], Jordi Deu-Pons[1], Iker Reyes- Salazar[1], Abel Gonzalez-Perez[1] and Nuria Lopez-Bigas[1]

[1]*Institute for Research in Biomedicine, IRB Barcelona, Spain*

OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers

Abstract: The characterization of the genomic alterations driving tumorigenesis is one of the main goals of cancer research. The identification of cancer drivers is key to a better understanding of the molecular mechanisms underlying tumors and to the implementation of precision cancer medicine. In this direction, driver genes have been detected using computational methods that analyze specific patterns of somatic mutations known as signals of positive selection: recurrence, clustering and functional impact. As whole genome sequencing (WGS) data are becoming widely available, the contribution of non-coding regions, which account for 98% of the genome (Dunham et al., 2012), to tumorigenesis may become under scrutiny. Therefore, new computational methods able to detect signals of positive selection in non-coding regions from WGS data are needed. While algorithms that exploit the recurrence and functional impact of mutations have been successfully developed, we are still missing a method based on the distribution or clustering of somatic mutations that can be effectively applied to the non-coding genome. This is relevant since different signals of positive selection are complementary in the detection of driver genes (Tamborero et al., 2013; Porta-Pardo et al., 2017). To this aim, we have developed OncodriveCLUSTL, a new algorithm for the detection of significantly mutated regions in the coding and non-coding genome based on the analysis of linear sequence clustering of somatic mutations. OncodriveCLUSTL builds a local cohort-specific background model derived from the nucleotide context mutational probabilities from the cohort under study. It performs a kernel density estimate (KDE) based analysis of the distribution of somatic mutations, reporting variable sized clusters in nucleotide sequences together with a ranked list of potential driver elements. We show OncodriveCLUSTL's versatility by analyzing coding sequences (CDS) as well as promoters, 5'UTRs, 3'UTRs and lncRNAs. In coding regions, OncodriveCLUSTL identifies known cancer drivers included in the Cancer Gene Census (CGC), outperforming the existing OncodriveCLUST method that detects coding drivers in protein sequences (Tamborero, Gonzalez-Perez, y Lopez-Bigas, 2013) as well as the 3D-clustering method HotMAPS (Tokheim

et al., 2016). In non-coding regions, OncodriveCLUSTL is able to identify clusters described by the literature such as telomerase reverse transcriptase (TERT) promoter hotspots, as well as potential candidates for further exploration.

22

Luis A. Rubio-Rodríguez[1], Adrián Muñoz-Barrera[1], José L. Roda-García[2], Carlos J. Pérez-González[3], Marcos Colebrook[2], Pedro González-Yanes[4], Carlos Flores[1,5] and José M. Lorenzo-Salazar[1]

[1]*Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Tenerife, Canary Islands, Spain*
[2]*Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Tenerife, Canary Islands, Spain*
[3]*Departamento de Matemáticas, Estadística e Investigación Operativa, Universidad de La Laguna, Tenerife, Canary Islands, Spain*
[4]*Centro de Cálculo, Escuela Superior de Ingeniería y Tecnología (ESIT-ULL), Universidad de La Laguna, Spain*
[5]*Hospital N.S. De Candelaria-CIBERES, Spain*

Development of a dockerized Bioinformatic architecture based on Big Data technologies to process data from Next Generation Sequencing

Abstract: Next Generation Sequencing (NGS) data imposes major challenges for processing due to its volume and complexity. Analysis of NGS data is generally based on many third party software, which are sometimes complex to install, configure and usually have dependencies that may arise in portability and reproducibility issues. Thus, it is necessary to develop infrastructures to store, manage and analyse massive genomic data in an efficient, scalable and reproducible way.

We have developed a Docker container-based infrastructure comprised of Bioinformatics and Big Data tools (Hadoop and Spark) for NGS data analysis. We have created Docker images for the Hadoop and Spark components and bioinformatics tools, including QC applications (FastQC, MultiQC, Qualimap2), aligners (BWA), variant callers (GATK4, Platypus), and other complementary software (JupyterLab). The infrastructure was deployed as a cluster of 20 commodity hardware nodes in a computer classroom at the ESIT-ULL Computing Center and can be easily scaled. We have used Docker Compose for multi-container definition and Docker Swarm for container orchestration and cluster management. To ensure data integrity, genomic data is stored in a distributed and a redundant manner across all nodes by using HDFS technology. This solution allowed university students to keep using their computers during data processing, so we could take advantage of existing equipment with no additional cost. As this is a non-dedicated infrastructure, it has been necessary to keep in mind the continuous addition and removal of nodes from the cluster. Docker images are available at: https://hub.docker.com/r/taroull/. Dockerfiles for deploying the described infrastructure are available at: https://github.com/lubertorubior/docker-esit

Funded by Ministerio de Ciencia, Innovación y Universidades (RTC-2017-6471-1; MINECO/AEI/FEDER, UE). This work has been supported by the CEDeI program (Centro de Excelencia de Desarrollo e Innovación, Cabildo de Tenerife).

23

Elena Rojano[1], Pedro Seoane-Zonjic[2], James Richard Perkins[2] and Juan Antonio Ranea[1]

[1]*University of Malaga, Spain*
[2]*CIBER of Rare Diseases, CIBERER, Spain*

Predictive software for helping to the clinical diagnosis of patients with complex diseases

Abstract: Complex diseases with genetic causes are the consequence of variants throughout the genome [1], and due to their complexity they are difficult to analyse. These variants usually affect different functional elements that can be involved in the same or in different pathways, influencing the development or homeostasis of the individual [2]. Knowing the genetic causes of the pathological phenotypes that characterize a disease remains a complicated task. On the one hand, the whole-genome sequencing of a patient remains expensive, and on the other hand, the computational analysis of the sequenced data requires time to be performed. To deal with this problem, systemic approaches are used to analyse the possible genetic causes that give rise to the pathological phenotypes of a disease [3]. Here we present a systemic approach for helping to the clinical diagnosis of patients with a complex pathological profile. Using patient information from DECIPHER database [4] and through network analysis, we have associated pathological phenotypes to regions of the genome that are potentially their cause [5,6]. We use these phenotype-genotype associations to predict the potential genetic regions affected from the phenotypic profile of a patient to be analysed. We obtain the genes that are located in these predicted regions and perform a functional annotation of the KEGG pathways that may be affected. In this manner, it can be known the complex systems that are the cause of these diseases.

References

1. Mitchell KJ. What is complex about complex disorders? Genome Biol. 2012; 13:237

2. Scheuner MT, Yoon PW, Khoury MJ. Contribution of Mendelian disorders to common chronic disease: Opportunities for recognition, intervention, and prevention. Am. J. Med. Genet. 2004; 125:50–65

3. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. Nat. Rev. Genet. 2014; 15:34–48

4. Corpas M, Bragin E, Clayton S, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Curr. Protoc. Hum. Genet. 2009; 84:1–17

5. Rojano E, Seoane P, Bueno-Amoros A, et al. Revealing the relationship between human genome regions and pathological phenotypes through network analysis. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 2017; 10208 LNCS:197–207

6. Bueno A, Rodríguez-López R, Reyes-Palomares A, et al. Phenotype-loci associations in networks of patients with rare disorders: application to assist in the diagnosis of novel clinical cases. Eur. J. Hum. Genet. 2018

24

Fernando Moreno-Jabato[1], Elena Díaz-Santiago[1], Elena Rojano[1], Pedro Seoane-Zonjic[2], James Richard Perkins[2] and Juan Antonio Ranea[1]

[1]*University of Malaga, Spain*
[2]*CIBER of Rare Diseases, CIBERER, Spain*

Biological system approaches to better understand patients with undiagnosed rare diseases

Abstract: Rare Diseases (RD) are those with a prevalence lower than 5 per 10,000 people. They are an important health issue, and can be difficult to diagnose due to atypical symptoms and doctor unfamiliarity. The majority have a genetic origin; therefore, it is important to create databases that contain phenotypes and genotypes for RD patients that can be

used to investigate disease patterns and develop tools to improve diagnosis. With this idea the "DatabasE of genomic variation and Phenotype in Humans using Ensembl Resources" (DECIPHER) was created, which includes data about phenotypes (annotated using Human Phenotype Ontology – HPO) and genotypes (CNVs) for patients with undiagnosed RDs. Using these data, we have developed two approaches that aim to i) built and study the comorbid phenotype network of patients with RDs and ii) Obtain communities of RD patients based on similar phenotypic profiles to identify possible new syndromes.

The first approach uses a bipartite network from DECIPHER of phenotype-patient relationships. By searching for phenotypes that co-occur frequently in different patients using this network, it establishes pairs of comorbid phenotypes based on an association index (hypergeometric index), which are used to build a phenotype-phenotype network. This procedure has allowed us to explore and analyze the features of this comorbid phenotype network of not-diagnosed patients with rare genomic disorders.

The second approach starts with the idea that similar phenotypes are likely to have a common genetic basis. First, a weighted patient-patient network is created based on Phenotype-Patient-Genotype data from DECIPHER, and measures of semantic similarity (Robinson and Resnick) are used to compare patients' phenotypic profiles. Then, clustering methods are applied to identify highly connected communities with similar phenotypic profiles. This process has been designed to optimize time and resource consumption, manage data scalability and calculate clusters with potentially biological meaning. This is based on: i) Clique membership, by selecting highly connected clusters; ii) Centrality, with patients potentially clustered into multiple communities, and iii) Granularity, allowing us to detect large communities as well as sub-communities within them. Optimization has been critical to manage scalability problems, commonly related to high density networks like Patient-Patient network obtained on this process, and is necessary to generate an effective tool. Once such communities have been discovered, further work will include adding genetic information to investigate the genetic mechanisms underlying these phenotypes profiles.

The methodologies developed and described in this poster will be useful to identify new potential syndromes and assisting in the diagnoses of patients with rare genomic disorders.

25

Laura Martinez Gomez[1] and Michael Tress[1]

[1]Centro Nacional de Investigaciones Oncológicas, CNIO, Spain

Importance of exon duplications in alternative splicing

Abstract: Alternative splicing and gene duplication have been proposed as two of the major mechanisms providing protein functional diversity (1,2). From the point of view of the protein, there are essentially just two types of alternative splicing, indels (which can be further separated into insertions and deletions) and substitutions. Protein sequence substitutions can be distinguished by their position in the protein sequence (N-terminal, C-terminal, internal) or by whether or not they arose from tandem exon duplications.

Substitutions that arose by exon duplication, make up only a small proportion of the annotated substitutions in the human genome (3). However, studies in human, mouse and Drosophila have shown that alternative isoforms generated from homologous duplicated exons are significantly over-represented in mass spectrometry studies (4). Homologous exons splicing events have very subtle effects in terms of protein folding disruption compared to other splicing mechanisms and a number are implicated in organismal development and disease (5). Despite this little is known about the biological relevance of most homologous exons.

Here we manually retrieved more than 250 pairs of homologous exons. We estimated the duplication dates based on sequence similarity searches in lamprey, fugu, zebrafish, spotted gar and coelacanth and manual curation using the

Ensembl, UniProt, RefSeq and APPRIS databases (6-9). We found that almost 80% of the tandem duplications in the set were conserved all the way back to coelacanth (more than 10 times as many as other alternative exons) and detected peptides for more than 55% of the isoforms generated from these homologous exons (compared to fewer than 1% for all other types of splice events).

Our results suggest that the generation of alternative isoforms from exon duplications, while rare, is likely to be an important means of generating functional diversity in eukaryotes.

1. Ohno, S. 1970. Evolution by gene and genome duplication. Springer, Berlin

2. Mironov, A.A., Fickett, J.W., and Gelfand, M.S. 1999. Frequent alternative splicing of human genes. Genome Res. 9: 1288-1293.

3. Kondrashov, F.A.; Koonin, E.V. Origin of alternative splicing by tandem exon duplication. Hum. Mol. Genet. 2001, 10, 2661–2669.

4. Abascal F, Ezkurdia I, Rodriguez-Rivas J, Rodriguez JM, del Pozo A, et al. Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level PLOS Computational Biology. 2015; 11(6): e1004325.

5. Hatje K, Rahman R, Vidal RO, et al. The landscape of human mutually exclusive splicing. Molecular Systems Biology. 2017;13(12):959.

6. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. et al.(2017) Ensembl 2017. Nucleic Acids Res., 45, D635–D642.

7. The UniProt Consortium. UniProt: the universal protein knowledgebase Nucleic Acids Res. 2017 ;45:D158–D169.

8. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D.et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., 44, D733–D745.

9. Rodriguez,J.M., Rodriguez-Rivas,J., Di Domenico,T., Vázquez,J., Valencia,A. and Tress,M.L. (2018) APPRIS 2017: principal isoforms for multiple gene sets. Nucleic Acids Res., 46, D213–D217.

27

Ernesto Luis Aparicio Puerta[1], David Jaspez[1], Juan Antonio Marchal[2], Danijela Koppers-Lalic[3] and Michael Hackenberg[1]

[1]*Department of Genetics,University of Granada, Spain*
[2]*Department of Human Anatomy and Embryology, University of Granada, Spain*
[3]*Department of Neurosurgery, VU University Medical Center, Netherlands*

liqDB: data and tools for reanalysis of sRNAseq-based liquid biopsy studies

Abstract: MiRNAs are important regulators of gene expression and are frequently deregulated under pathologic conditions. They are highly stable in bodily fluids which makes them feasible candidates to become minimally invasive biomarkers. In fact, several studies already proposed circulating miRNA-based biomarkers for different types of neoplastic, cardiovascular and degenerative diseases. However, many of these studies rely on small RNA sequencing experiments that are based on different RNA extraction and processing protocols, rendering results incomparable. We generated liqDB, a database for liquid biopsy small RNA sequencing profiles that provides users with meaningful information to guide their small RNA liquid biopsy research and to overcome technical and conceptual problems. By means of a user-friendly web interface, miRNA expression profiles from 1609 manually annotated samples can be queried and explored at different levels. Result pages include downloadable expression matrices, differential expression analysis, most stably expressed miRNAs, cluster analysis and relevant visualisations by means of boxplots and heatmaps.

We anticipate that liqDB will be a useful tool in liquid biopsy research as it provides a consistently annotated large compilation of experiments together with tools for reproducible analysis, comparison and hypothesis generation. LiqDB is available at http://bioinfo5.ugr.es/liqdb

---

28

José Córdoba[1], Rafael Núñez-Serrano[1], Lorena Aguilera-Cobos[1], Pedro Seoane[1], Manuel Manchado[2] and M. Gonzalo Claros[1]

[1]*Universidad de Málaga, Spain*
[2]*IFAPA El Toruño, Spain*

Genomic markers to improve the production of Senegalese sole in aquaculture

Abstract: The ongoing de novo sequencing of Solea senegalensis genome has enabled new approaches to study some bottlenecks that constrain sole aquaculture such as the loss of sexual courtship behavior in males (García-López et al., 2005; Howell et al., 2009). We have developed a workflow to locate sex-linked regions as a source of genetic markers associated to genetic sex determination. Our workflow, based in AutoFlow (Seoane et al., 2016), was used to align sequencing data from high throughput sequencing data from male and female specimens. First, we pre-processed raw Illumina libraries to avoid any sequencing bias by means of SeqTrimBB, a tool especially suited for Illumina datasets that can be easily adapted to any other sequencing technology. SeqTrimBB is very flexible since its architecture is based on a set of plugins that specifically address issues known to affect subsequent analysis, such as bad quality or sequencing adapter presence and contains some predefined pre-processing templates to deal with reads from different platforms and experimental approaches. The resulting useful reads were mapped with Bowtie2 onto the Solea senegalensis draft genome available in our laboratory (Manchado et al., 2016) to further compare such males and female-specific alignments using PeakRanger (Feng et al., 2011) and extract some candidate sex-specific genomic regions. In order to reduce false positives and select non-mapped regions in only one sex, reads were re-mapped against the extracted regions. A total of 32 regions (23 female-exclusive and 9 male-exclusive) were found. Their genomic context revealed that they were next to genes related to female reproductive cycle, to sex change, or to stress response/regulation. Some of the regions have been experimentally validated on cultured soles. Finally, the tested workflow is still on development to implement new capacities such as the selection of more suitable aligners and designing an algorithm that outperforms the PeakRanger results in order to increase the amount of candidate regions as well as their reliability.

This work was supported by co-funding by the European Union through the European Regional Development Fund (ERDF) 2014-2020 “Programa Operativo de Crecimiento Inteligente” together with Spanish AEI “Agencia Estatal de Investigación” to RTA2013-00068-C03, AGL2017-83370-C3-3-R and RTA2017-00054-C03-03.

29

Marina Reixachs[1], Jorge Ruiz-Orera[2], Mar Alba[2,3] and Eduardo Eyras[1,3]

[1]*Universitat Pompeu Fabra, Spain*
[2]*GRIB, Hospital del Mar Research Institute, Spain3*
[3]*Catalan Institution for Research and Advanced Studies, ICREA, Spain*

Measuring ribosome profiling at isoform level: towards unveiling the functional impact of alternative splicing

Abstract: The differential production of transcript isoforms through the mechanism of alternative splicing is crucial in multiple biological processes as well as pathologies, including cancer. This has been exhaustively shown at RNA level but

it remains elusive at protein level. Sequencing of ribosome-protected mRNA fragments (ribosome profiling) provides information on the transcripts being translated. We describe a new pipeline for the quantification of individual transcript coding sequences from ribosome profiling using both RNA-seq and Ribo-seq. Using multiple datasets, we find evidence of translation for 50-70% of the isoforms quantified with RNA-seq.

Additionally, we performed differential splicing analysis between glia and glioma samples from human and mouse and found consistent changes occurring in both RNA-seq and Ribo-seq for the majority of cases, indicating that changes in the relative abundance of transcript isoforms lead to changes in the production of protein isoforms in the same direction. Among the cassette exon events changing splicing, we identified an enrichment of orthologous exons with the majority of them preserving the directionality of the change. Interestingly, there was a significant enrichment of microexons that decrease inclusion in glioma compared to glia in both, human and mouse, suggesting a concerted mechanism of dedifferentiation in glioma.

30

Oriol Pich[1], Ferran Muinos[1], Radhakrishnan Sabarinatahan[2], Iker Reyes-Salazar[1], Abel Gonzalez-Perez[1] and Nuria Lopez-Bigas[1]

[1]*Institute for Research in Biomedicine, IRB Barcelona, Spain*
[2]*National Centre for Biological Sciences, Tata Institute of Fundamental Research, India*

Somatic and germline mutation periodicity follow the orientation of the DNA minor groove around nucleosomes

Abstract: Nucleosomes are the most pervasive structural feature of eukaryotic genomes, covering between 75% and 90% (Segal et al., 2006). Their presence creates a periodic pattern formed by alternating histone-covered and linker DNA sequences. Moreover, the DNA wrapped around them also exhibits an alternating pattern of structurally distinct stretches: ~10-bp interspersed segments of DNA with the minor groove facing the histones and away from them. To study the influence of these to alternating structural patterns on the generation of the mutations before selection across the genome, we analyzed the somatic mutations observed across tumors, which are exposed to little selection once the few mutations driving tumorigenesis are filtered out (Frigola et al., 2017; Martincorena et al., 2017).

First, we observed strong periodic patterns in the mutation rate of several cancer types, which track the alternation of nucleosomes and linkers and the rotational orientation of the minor groove of the DNA with respect to histones. In the latter, we observed a relative increase of mutation rate above the expected across several cohorts of tumors showed a significant periodicity of 10 bp, or one helix turn. In some tumors (e.g., esophageal adenocarcinomas) the relative increase of mutation rate peaked at stretches of DNA with the minor groove facing the histones. In others (e.g., melanomas and lung adenocarcinomas), its maxima were at stretches of DNA with the minor groove facing away from the histones.

We then demonstrated that the periodic structure and orientation of the relative increase of the mutation rate within nucleosome-covered DNA is determined by the mutational processes active in each tumor. Ultimately, thus, it is the combined effects of DNA damage and repair efficiency at DNA stretches with the minor groove facing toward and away from histones that shape this periodicity. The germline variation and interspecies C>T divergence also show a significant periodicity, reminiscent of that observed in somatic mutations of certain tumors. Finally, we demonstrate that the widespread WW periodicity across eukaryotic genomes –10-bp A/T dinucleotides periodicity–, in the absence of other evolutionary forces could have arisen as a result of the periodic de novo mutation mutation rate.

31

Abel Gonzalez-Perez[1]

[1]*Institute for Research in Biomedicine, IRB Barcelona, Spain*

Computational genomics at the heart of cancer biology

Abstract: In our lab, we use data on genomic mutations in tumors in three main lines of research. First, we study the distribution of mutations across different regions in the genome to understand basic questions about molecular biology, such as the interplay between DNA damage and repair, and other cellular processes. (I will briefly present one example of this basic research.) This helps us build refined models of the expected background mutation rate of different genomic elements across cell types. Comparing these models with the observed mutational patterns of genomic elements across cohorts of tumors, we are then able to detect which of them are under positive selection in the process of tumorigenesis. We go one further step to identify which amongst all the mutations (point mutations and structural variants) of these elements detected in a particular tumor are actually tumorigenic. (I will present our work on unraveling the panorama of driver mutations of more than 2500 tumor whole genomes.) Translating this basic knowledge in ways that may bridge the gap to personalized cancer medicine is the third line of research in the lab. We search for the potential clinical significance of individual driver mutations, and we build tools that assist clinical oncologists in therapeutic decision-making. (I will exemplify our translational research with our work on the immune-phenotypes of solid tumors.)

32

Marta R. Hidalgo[1], Francisco Salavert Torres[2], Alicia Amadoz[3], Cankut Cubuk[4], José Carbonell-Caballero[5] and Joaquin Dopazo[4]

[1]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[2]*BioBam, Spain*
[3]*Igenomix S.L., Spain*
[4]*Fundación Progreso y Salud, FPS, Spain*
[5]*Centre de Regulació Genòmica, CRG, Spain*

Analyzing signaling pathway and function activity in a new Bioconductor package: HiPathia

Abstract: Hipathia package is a new Bioconductor R package implementing the Canonical Circuit Activity Analysis method for the quantification of the signaling pathways activity presented in [1]. It is a method for the computation of the signal transduction along signaling pathways from transcriptomics data. It is based on an iterative algorithm which is able to compute the signal intensity passing through the nodes of a network by taking into account the level of expression of each gene and the intensity of the signal arriving to it. It also provides a new approach to functional analysis allowing to compute the signal arriving to the functions annotated to each pathway.

The Hipathia methodology has been successfully applied to the study of differences in Breast Cancer in the context of signaling pathways, [1], to the identification of transcriptional changes in blood after death, [2], and to the in-silico analysis of knockouts, over-expressions or drug administration[3].

This methodology has been implemented in the webtool http://hipathia.babelomics.org, allowingthe user to compare signal propagation in an experiment, and train and use a predictor based on the activation of the canonical circuits or subpathways. Here we present the R package hipathia, which has been conceived as a functional tool for R users and allows more control on the analysis pipeline than the web implementation does. The package will set the Hipathia methodology closer to the potential end users, that is, the bioinformatics experts, by allowing to perform a full pathways and function Hipathia analysis in a R environment.

References

[1] Hidalgo,M.R., et al. (2017) High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget, 8(3), 5160–5178

[2] Ferreira,P.G., et al. (2018) The effects of death and Post-mortem Cold Ischemia on Human tissue transcriptomes. Nature Communications, 9, 490.

[3] Salavert,F., et al. (2016) Actionable Pathways: Interactive Discovery of Therapeutic Targets using Signaling Pathway Models. Nucleic Acids Research, 44, Web Server Issue: W212-W216

33

Francisco Martínez-Jiménez[1], Abel Gonzalez-Perez[1] and Núria López-Bigas[1]

[1]*Institute for Research in Biomedicine, IRB Barcelona, Spain*

Proteome-wide discovery of degrons and analysis of their role in tumorigenesis across cancer types

Abstract: The ubiquitin mediated proteolysis system (UPS) is involved in both quality control and regulation of protein levels that control crucial cellular processes. Degrons are short linear motifs embedded within the sequences of substrates that are recognized and bound by E3 ligases, forming the basis of the specificity of UPS. Three decades of research have yielded the identification of degrons for only ~6% of the 600 E3-ligases encoded in the human genome. Identifying new degrons is key to understand the operation of UPS and its role in diseases.

We used a combination of machine learning, protein-protein interactions and matched exome, transcriptome and proteome data from 8,167 human tumors to perform a proteome-wide discovery of degrons. The integrative approach enabled the identification of 449 high-confidence instances of annotated degrons. We validated the approach showing that missense mutations and inframe indels affecting new degron instances are significantly more stabilizing than analogous mutations located outside of degradation motifs. A search for degrons of yet unknown E3-ligases pinpointed 41 regions harbouring highly stabilizing mutations and resembling the biochemical properties of annotated degrons. Such regions involve essential proteins such as LCK, ERBB3 or CHEK1. We also showed that degron destroying mutations are positively selected in human tumors. Finally, cohort specific analysis of UPS perturbations revealed a mutually exclusive pattern of mutations between degron mutations and driver alterations in their E3 ligases posing UPS perturbations as a widespread mechanism of tumorigenesis.

34

Tamara Hernández-Beeftink[1,2], Héctor Rodríguez-Pérez[1], Carolina Gonzalez Navasa[3], José L. Roda-García[4], Carlos J. Pérez-González[5], José M. Lorenzo-Salazar[3], Rafaela González-Montelongo[3], Marcos Colebrook[4] and Carlos Flores[3,6]

[1]*Research Unit, Hospital Universitario N.S. de Candelaria, Universidad de La Laguna, Santa Cruz de Tenerife, Spain*
[2]*Hospital Universitario de Gran Canaria Doctor Negrín, Spain*
[3]*Genomics Division, Instituto Tecnológico y de Energías Renovables (ITER), Santa Cruz de Tenerife, Spain*
[4]*Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Santa Cruz de Tenerife, Spain*
[5]*Departamento de Matemáticas, Estadística e Investigación Operativa, Universidad de La Laguna, Spain*
[6]*Hospital N.S. De Candelaria-CIBERES, Spain*

An interactive species classification instrument with educational purposes using NanoDJ and PCR-free mtDNA enrichment

Abstract: We have developed NanoDJ, an interactive collection of Jupyter notebooks distributed as a Docker software container, to facilitate Oxford Nanopore Technologies (ONT) sequence analyses by integrating capabilities for data manipulation, sequence comparison and assembly for research and educational purposes. It includes base calling, read filtering, simulation and plotting routines with a variety of widely used aligners and assemblers, including procedures for hybrid assembly. In this study, we show the utility of NanoDJ for the development of a hands-on genomics educational tool applying mitochondrial DNA (mtDNA) PCR-free enrichment for species apportionment of food components. After assessing two distinct mtDNA enrichment protocols to pick the best-performing solution, we used a simple 10-minute library preparation (SQK-RAD004) of a mock mixture and a 48-h MinION (ONT) run obtaining 35,424 reads of 3.5 Kbase mean size. NanoDJ allowed a fast BLAST-based identification of all species in the sample, thus validating this hand-on experience as a learning-by-doing instrument for transdisciplinary education adapted to basic laboratory resource settings. In conclusion, this instrument can be used as a routine tool for training and education, as well as for research, contributing to democratize the access to the latest advances in genomics. Software availability: https://github.com/genomicsITER/NanoDJ

Funded by Instituto de Salud Carlos III (PI14/00844; FI17/00177) and Ministerio de Ciencia, Innovación y Universidades (RTC-2017-6471-1; MINECO/AEI/FEDER, UE) co-financed by the European Regional Development Funds, “A way of making Europe” from the European Union; by the agreement OA17/008 with Instituto Tecnológico y de Energías Renovables (ITER) to strengthen scientific and technological education, training, research, development and innovation in Genomics, Personalized Medicine and Biotechnology; and by the CEDeI program (Centro de Excelencia de Desarrollo e Innovación, Cabildo de Tenerife).

35

Federico Abascal[1], David Juan[2], Laura Martinez Gomez[3], Jose Manuel Rodriguez[3,4], Jesús Vázquez[5], Irwin Jungreis[6], Maria Rigau[7] and Michael Tress[3]

[1]*Wellcome Trust Sanger Institute, United Kingdom*
[2]*Institute of Evolutionary Biology, UPF-CSIC, Spain*
[3]*Centro Nacional de Investigaciones Oncológicas, CNIO, Spain*
[4]*Spanish National Bioinformatics Institute, INB, Spain*
[5]*Spanish National Center for Cardiovascular Research, CNIC, Spain*
[6]*Massachusetts Institute of Technology, MIT, United States*
[7]*Barcelona Supercomputing Center, BSC, Spain*

Loose ends: almost one in five human genes still have unresolved coding status

Abstract: There are three well-maintained manual reference databases for the human genome, RefSeq (1), UniProtKB (2) and Ensembl/GENCODE (3,5). Over the years these three databases have converged on similar numbers of protein coding genes and the number of protein coding genes in the human reference gene set has been more or less stable since 2012 (5). Despite this, the human gene sets are in certain state of flux with coding genes being added and reclassified with each new release.

Even though the three reference sets contain the same number of coding genes, the 20,000 plus coding genes are not the same in each. Here we carry out a manual cross-reference between the three reference catalogues to investigate the differences between the Ensembl/GENCODE, RefSeq and UniProtKB proteomes.

We find that the number of annotated coding genes in the union of the three reference sets exceeds 22,000, while at the same time the three reference sets classify one in every eight coding genes differently. How many of these 2,754 differently annotated genes are protein coding? In fact we find that genes that are annotated as coding by just one or two sets of manual annotators are starkly different from those annotated as coding by all three reference sets. Coding genes that are differently classified across the three sets are rich in non-coding features and large-scale genetic variation data suggests that many of these potential coding genes are unlikely to be functionally important.

This analysis clearly shows the importance of a curated human reference set. Over the years since the human genome sequence was released rigorous manual annotation has brought us considerably closer to a final catalogue of human coding genes. Annotators agree on almost 90% of coding genes, but the final 10% of genes, those with the most conflicting evidence, are proving more difficult to classify.

1. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., 44, D733-745.

2. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase Nucleic Acids Res., 45, D158-D169.

3. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K.,

Carvalho-Silva,D., Cummins,C., Clapham,P., et al. (2017) Ensembl 2017. Nucleic Acids Res., 2017;45:D635-42.

4. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S., et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res., 22, 1760-1774.

5. Southan,C. (2017) Last rolls of the yoyo: Assessing the human canonical protein count. F1000Research, 6, 448.

https://www.ncbi.nlm.nih.gov/pubmed/29982784

36

Lorena Aguilera-Cobos[1], Pedro Seoane[1], Rafael Núñez Serrano[1], José Córdoba Caballero[1], Silvana Tapia-Paniagua[1], María Balebona[1], Miguel A. Moriñigo[1] and M. Gonzalo Claros[1]

[1]*Universidad de Málaga, Spain*

Detection of transposons modifying genome background in probiotics

Abstract: The study of probiotic microorganisms is very interesting in the aquaculture field. Administration of live microorganisms in adequate amounts confers some benefits to the host (Kechagia et al. 2013). Even if Shewanella putrafaciens include pathogens and saprophytic strains related to fish spoilage and fish infection (Esteve, Merchán, and Alcaide 2016). The Pdp11 strain of Shewanella putrefaciens has been proved to provide beneficial effects in Sparus aurata (Chabrillón et al. 2005) and Solea senegalensis (Rodrigáñez et al. 2008). Studies focused on Pdp11 could hed light on the origin of this probiotic character.

We have designed a bioinformatic workflow to detect transposons in the newly sequenced Pdp11 genome (Tapia-Paniagua et al, in press). Their presence interrupting genes account for a contribution to its probiotic character due to the lost of virulence or the gain of probiotic effect. The workflow was developed in Ruby programming language and provides: the genomic localisation of known transposons, host coding regions disrupted by complete transposons or their repeated insertion sequences, and transposons and coding regions disrupted identifiers, to stablish the putative functions of Pdp11 that could be affect by the transposons disruption.

These results would support new possible hypothesis about the Pdp11 probiotic character since 14 coding regions related to S. putrefaciens were disrupted by transposons, 4 of which are directly involved in pathogenic mechanisms.

This work was supported by co-funding by the European Union through the European Regional Development Fund (ERDF) 2014-2020 "Programa Operativo de Crecimiento Inteligente" together with Spanish AEI "Agencia Estatal de Investigación" to grants RTA2013-00068-C03, AGL2017-83370-C3-3-R and RTA2017-00054-C03-03.

38

Luis Ángel Rodriguez-Lumbreras[1], Brian Jiménez-García[1], Josep Lluis Gelpí[1,2] and Juan Fernández-Recio[1,3]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Universitat de Barcelona, UB, Spain*
[3]*IBMB-CSIC, Spain*

Structural modeling of protein-DNA interactions with pyDockDNA

Abstract: Protein-DNA interactions are involved in the most critical processes of a cell, such as gene expression, replication and transcription (1). Details on such interactions are needed to understand these processes at atomic level, but unfortunately, for many protein-DNA complexes, structural data is not available. Protein-DNA interactions can be structurally modelled by available protein-protein docking tools, but the specific characteristics of the nucleic acids, such as the energetics and conformational flexibility, are not adequately captured with standard protein docking methods. A few computational methods for protein-DNA docking have been reported (2), but we need new tools that can deal with the specificity of these challenging interactions.

Here we have adapted pyDock (3) for the structural modeling of protein-DNA complexes. The procedure is divided in two major steps. The initial sampling step consists in the generation of 10,000 protein-DNA docking models with FTDock (4), without electrostatics option. In a second scoring step, energy-based functions are computed for all generated

docking poses. Different parameters and combinations of these energetic functions have been tested on an available protein-DNA docking benchmark (5). The best results for the scoring of rigid-body protein-DNA docking poses were obtained through a combination of electrostatics and van der Waals terms. This protocol has been implemented in the MuG VRE platform (6) as a new application called pyDockDNA. To overcome geometrical uncertainty of the grid-based representation of FTDock, more exhaustive sampling could be performed by starting from different random rotations of the interacting molecules, and then using a clustering method (7) to remove redundant docking solutions. This more exhaustive sampling can improve the success rate for the top 10 docking models up to 10%.

The detailed analysis of the docking results on the DNA-protein benchmark clearly showed that the predictive success are mostly related to the structural deviation between the DNA model based on the canonical B-structure and its final bound conformation. Another interesting finding is that, for many complexes, electrostatic term determines the quality of the docking models, as opposed to protein-protein docking, in which desolvation term was more important for the scoring.

REFERENCES

1. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. Nat. Rev. Genet., 11, 751–760.

2. Banitt,I. and Wolfson,H.J. (2011) ParaDock: A flexible non-specific DNA - Rigid protein docking algorithm. Nucleic Acids Res., 39.

3. Cheng,T.M.-K., Blundell,T.L. and Fernandez-Recio,J. (2007) pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. Proteins Struct. Funct. Bioinforma., 68, 503–515.

4. Gabb,H. a, Jackson,R.M. and Sternberg,M.J. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. J. Mol. Biol., 272, 106–120.

5. van Dijk,M. and Bonvin,A.M.J.J. (2008) A protein-DNA docking benchmark. Nucleic Acids Res., 36, e88–e88.

6. MuG Virtual Research Environment. [online] Available at: https://vre.multiscalegenomics.eu.

7. Gil,V.A. and Guallar,V. (2014) PyProCT: Automated cluster analysis for structural bioinformatics. J. Chem. Theory Comput., 10, 3236–3243.

39

Lorena de la Fuente[1], Manuel Tardáguila[2], Hector del Risco[2], Pedro Salguero[1], Sonia Tarazona[1], Victoria Moreno[1] and Ana Conesa[1,2]

[1]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[2]*University of Florida, United States*

tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing

Abstract: Nowadays, the impact of alternative splicing on creating transcriptomics complexity as well as its association to numerous neurodegenerative diseases has been repeatedly demonstrated. However, in spite of counting with dense catalogs of isoforms, strategies to interrogate alternative isoforms from a functional perspective are still lacking. In consequence, analyzing if differential splicing across conditions is impacting specific functional elements such as post-translational modifications or nuclear localization signals, or coming up with genes involving UTR regulation by differential splicing are currently tasks difficult to address.

Here, we present tappAS, a new user-friendly Java application for the functional analysis of alternative isoform usage. Making use of extensive functional annotation at isoform resolution and RNA-seq count data. TappAS provides a diverse

set of approaches and isoform-resolved visualization tools that allow the user to formulate varied functional hypothesis about the role of alternative splicing in a given system of study.

At JBI18, we will present an overview of tappAS, its usability, friendly interface and specially, we will show its application to neural systems, highlighting the Motif Differential Splicing Analysis, a novel approach able to point out specific annotated elements that appear regulated by alternative splicing or alternative polyadenylation. Moreover, capabilities of our implemented visualization engine for functional annotations at splicing variants resolution will be shown.

40

Carlos Torroja[1], Zlatko Trajanoski[2] and Fatima Sanchez-Cabo[2]

[1]*Centro Nacional de Investigaciones Cardiovasculares Carlos III, Spain*
[2]*Division of Bioinformatics, Biocenter Medical University of Innsbruck, Austria*

digitalDLSorter: A Deep Learning algorithm to quantify immune cell populations based on scRNA-Seq data

Abstract: the development of single cell transcriptome sequencing has allowed researchers the possibility to dig inside the role of the individual cell types within a compartment. It also expands to the whole transcriptome what before was only possible for a few tenths of antibodies in cell population analysis. But more importantly it allows to resolve the permanent question of whether the changes observed in a particular bulk experiment are a consequence of changes in cell type proportions or an aberrant behaviour of a particular cell type. However, single cell experiments are still complex to perform and expensive to sequence making bulk RNA traditional approaches more common.

From a statistical perspective single cell data are particularly interesting due to its high dimensionality, overcoming the limitations of the “skinny matrix” characteristics of traditional bulk RNA-Seq experiments. Therefore scRNA-Seq data are especially suitable for the application of machine learning algorithms and in particular of Deep Learning (DL) methods.

We present here a DL based method to enumerate and quantify the immune infiltration in colorectal and breast cancer bulk RNA-Seq samples starting from scRNA-Seq (Tirosh et al. 2016, Chung et al. 2017, Li et al. 2017). Our method makes use of a Deep Neural Network (DNN) model that allows quantification not only of lymphocytes as a general population but also of specific CD8+, CD4Tmem, CD4Th and CD4Tregs subpopulations, as well as B-cells and Stromal content. Moreover, the signatures were built from scRNA-Seq data from the tumour, preserving the specific characteristics of the tumour microenvironment as opposite to other approaches in which cells were isolated from blood. Our method was applied to synthetic bulk RNA-Seq and to samples from the TCGA project (n=1208) yielding very accurate results in terms of quantification and survival prediction.

This methodology can be extended to other clinical questions for which accurate quantification of immune cell subtypes might be important, i.e. atherosclerosis, auto-immune disorders.

41

Marco Trevisan-Herraz[1], Jose Rodriguez[2], Jesús Vázquez Cobos[2], Navratan Bagwan[2] and Elena Bonzon-Kulichenko[2]

[1]*Institute of Cellular Medicine, Newcastle University, United Kingdom*
[2]*Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Spain*

SanXoT: a workflow for the quantitative analysis of high-throughput proteomics experiments

Abstract: Mass spectrometry-based proteomics has had a formidable development in recent years, increasing the amount of data handled and the complexity of the statistical resources needed. Here we present SanXoT, an open-source software package, for the statistical analysis of high-throughput, quantitative proteomics experiments. SanXoT is based on our previously developed WSPP [1,2] statistical model and has been specifically designed to be modular, scalable, and user-configurable. It allows limitless workflows that adapt to most experimental setups, including quantitative protein analysis in multiple experiments, systems biology, quantification of post-translational modifications and comparison and merging of experimental data from technical or biological replicates [2-6]. Furthermore, it can be used as a standalone program, or, alternatively, as a package of routines ready to be used in combination with third party software. SanXoT has been used successfully to analyse hundreds of experiments at CNIC, becoming a key element for the interpretation of data generated in large-scale clinical projects.

Further information: https://www.cnic.es/wiki/proteomica/index.php/SanXoT_software_package

References:

1.Navarro P, et al. General statistical framework for quantitative proteomics by stable isotope labelling. J Proteome Res. 2014 Mar 7;13(3):1234-47. doi: 10.1021/pr4006958.

2.García-Marqués F, et al. A Novel Systems-Biology Algorithm for the Analysis of Coordinated Protein Responses Using Quantitative Proteomics. Mol Cell Proteomics. 2016 May;15(5):1740-60. doi:10.1074/mcp.M115.055905.

3.Latorre-Pellicer A, et al. Mitochondrial and nuclear DNA matching shapes metabolism and healthy ageing. Nature. 2016 Jul 28;535(7613):561-5.

4.Guarás A, et al. The CoQH2/CoQ Ratio Serves as a Sensor of Respiratory Chain Efficiency. Cell Rep. 2016 Apr 5;15(1):197-209. doi:10.1016/j.celrep.2016.03.009.

5.Martin-Lorenzo M, et al. Immune system deregulation in hypertensive patientschronically RAS suppressed developing albuminuria. Sci Rep. 2017 Aug21;7(1):8894. doi: 10.1038/s41598-017-09042-2.

6.Jorge I, et al. The human HDL proteome displays highinter-individual variability and is altered dynamically in response toangioplasty-induced atheroma plaque rupture. J Proteomics. 2014 Jun 25;106:61-73.doi: 10.1016/j.jprot.2014.04.010.

42

Lorena de La Fuente Lorente[1], Ana Conesa[1,2], Manuel Tardaguila[3], Cristina Marti[1], Héctor Del Risco[2], Victoria Moreno[1], Cécile Pereira[2], Francisco Pardo-Palacios[1], Ali Mortazavi[4], Iakes Ezkurdia[5], Marc Ferrel[2], Marissa Macchietto[4], Lennart Martens[6], Maravillas Mellado[1], Susana Rodríguez[1], Michael Tress[7] and Jesús Vázquez[5]

[1]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[2]*University of Florida, United States*
[3]*Sanger Institute, United Kingdom*
[4]*University of California, United States*
[5]*Centro Nacional de Investigaciones Cardiovasculares, CNIC, Spain*
[6]*UGent / VIB, Belgium*
[7]*Spanish National Cancer Research Centre, CNIO, Spain*

SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification.

Abstract: With the increasing utilization of long read technologies, the necessity for a tool that provides a comprehensive classification of novel transcripts as well as their deep characterization and quality control is ever more pressing. Exhaustive transcriptome curation methods to remove potential artifactual isoforms are essential at a time when long read sequencing and splicing-aware transcriptomics studies are becoming more popular to reveal system regulation triggered by mechanisms as alternative splicing. Here we present SQANTI (Structural and Quality Annotation of Novel Transcript Isoforms), a tool for the analysis long-read transcriptomics data that provides the methods to deliver quality-evaluated (quality control metrics) and curated full-length transcriptomes (machine learning approach). Using SQANTI we revealed that, even well annotated genomes as mouse genome, has still an important fraction of missing transcripts. More importantly, we show that incomplete annotation has a strong negative impact in the accuracy of current isoform expression quantification algorithms. Moreover, since it was released, SQANTI has been applied to multiple organisms, different long-read sequencing platforms (PacBio, Nanopore, etc) and different transcriptome reconstruction pipelines as well as becoming a standard long-read QC procedure for PacBio Iso-seq bioinformatics developers. Results highlight the relevance of SQANTI to fully understand long-read defined transcriptomes.

43

Angela Del Pozo[1], Mario Solís[1], Beatriz Ruz[1], Pablo Lapunzina[1] and Kristina Ibáñez[2]

[1]*INGEMM - Hospital Universitario La Paz Madrid - IdiPaz – CIBERER, Spain*
[2]*Genomics England, Queen Mary University London, United Kingdom*

LACONv: a tool for CNV detection in gene panel for diagnosis

Abstract: Since the implantation of next-generation sequencing technologies (NGS) in clinical setting there has been a great development to scan the genome for detection of SNPs and short INDELs. As consequence, currently it is available a plethora of tools and methodologies that could consider as the gold standard in NGS for diagnosis.

However, structural variation, such as copy number variation (CNV), which includes insertions, deletions and duplications are routinely identified in clinical domain by array aCGH or MLPA as there are no “best practice guidelines” in the NGS community to assess the validity and quality of CNV detection neither reference material that could be used to perform benchmarking for the clinical study of the methods.

In addition, many methods do not perform correctly in data from targeted gene sequencing (i.e. gene panels) as they are developed for whole exome sequencing (WES) and they are not adapted for shorter sizes of genome capture.

In this work, it is presented an in-house tool for CNV detection in patients that had been screened because a particular genetic condition was suspected. The tool (LACONv) has been routinely tested in a cohort of 10,000 NGS samples that have been sequenced with NGS gene panels. When a CNVs detected was considered as candidate of being involved in the phenotype of the patient, it was validated with MLPA resulting in a 70% of precision rate.

LACONv is based in two metrics: normalized weighted ratio of depth-of-coverage with respect to controls and allelic imbalance of heterozygous variants. CNVs are functionally enriched with some public databases in order to facilitate clinical interpretation.

Finally, robust and validated CNV detection in NGS samples provides new diagnostic opportunities in those patients that remains undiagnosed as they carry mutations in criptic genomic positions that are not able to be detected in commercial testing kits.

44

Andrea Rubio-Ponce[1], Iván Ballesteros[1], Andrés Hidalgo[1] and Fátima Sánchez-Cabo[1]

[1]*Centro Nacional de Investigaciones Cardiovasculares, CNIC, Spain*

A statistical framework for accurate detection of circadian genes in mammalian systems

Abstract: Circadian genes are found in most animal cells. They have been shown to be of paramount importance in tissue homeostasis and organ function, thereby making them important targets of genomic analyses. A caveat of most algorithms available for their detection is the need for frequent sampling (every hour) for 24 or even 48 hours to accurately detect rhythmic genes. In research with mammalian systems, however, repeated sampling can be expensive, unethical and often unfeasible given the cost and difficulty of access to certain models. This has in some cases led to work without replicates or to lower the rigor of the tests (i.e. using raw p-values instead of q-values). While this approach certainly increases the amount of detected genes, it also escalates the number of false positives, hindering validation and downstream analysis. In this study we show that the use of nonlinear mixed models is a more adequate framework for accurate detection of circadian genes in real experimental setups with limited sampling. Furthermore, comparison of our approach with other widely used non-parametric methods such as JTK and Biocycle under different experimental scenarios reveals more robust identification of circadian genes. We use these models to provide researchers with guidelines for experimental design and data analysis under different sampling conditions. Finally, we provide a web application and an R package to make this algorithm available to the community.

45

Inés Sentís[1]

[1]*Institute for Research in Biomedicine, IRB Barcelona, Spain*

The evolution of T-Cell Acute Lymphoblastic Leukemia in adult patients under treatment

Abstract:

Acute lymphoblastic leukemia (ALL) is an heterogeneous disease caused by the high proliferation of the lymphoblasts in the bone marrow which are either B limphoid precursors (B-ALL) or T limphoid precursors (T-ALL) 1. The majority of the ALL cases (75-85%) corresponds to B-ALL whereas the rest of them are T-ALL 1,2. Concretely, T-ALL accounts for 25% of ALL cases in adults 3 whereas pediatric T-ALL represent a 15% of ALL disease in children 4. Despite the increasing success in survival rates in ALL 5 there is still a fatal prognosis for those patients that present a recurrence in

the disease (relapse) specially in adults 6–8 . Consequently, understanding the molecular mechanisms leading to treatment resistance in ALL might create new clinical opportunities to reduce the fatal outcome of a recurrence in this cancer. In collaboration with the Josep Carreres Leukemia Research Institute and the Hospital del Mar Medical Research Institute (IMIM) we are currently sequencing the whole genome of 20 patients presenting T-ALL with paired samples at diagnosis (primary tumor) and relapse time points. We are also studying the phenotypes of cell leukemic populations with single-cell sequencing of the transcriptome. Here, we detail the landscape of the driver alterations across patient samples and the heterogeneity among patients. We focus on a comparison of genetic recurrent mechanisms underlying primary and relapse in ALL. Preliminary results will be presented.

Bibliography

1.Ustwani, O. Al et al. Clinical updates in adult acute lymphoblastic leukemia. Crit. Rev. Oncol. Hematol. 99, 189–199 (2016).

2.Chiaretti, S. & Foà, R. T-cell acute lymphoblastic leukemia. Haematologica 94, 160–162 (2009).

3.Neumann, M. et al. Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations. Blood 121, 4749–4752 (2013).

4.Kunz, J. B. et al. Pediatric T-cell lymphoblastic leukemia evolves into relapse by clonal selection, acquisition of mutations and promoter hypomethylation. Haematologica 100, 1442–1450 (2015).

5.Hunger, S. P. & Mullighan, C. G. Acute Lymphoblastic Leukemia in Children. N. Engl. J. Med. 373, 1541–1552 (2015).

6.Einsiedel, H. G. et al. Long-term outcome in children with relapsed ALL by risk-stratified salvage therapy: Results of trial Acute Lymphoblastic Leukemia-Relapse Study of the Berlin-Frankfurt-Münster Group 87. J. Clin. Oncol. 23, 7942–7950 (2005).

7.Oriol, A. et al. Outcome after relapse of acute lymphoblastic leukemia in adult patients included in four consecutive risk-adapted trials by the PETHEMA study group. Haematologica 95, 589–596 (2010).

8.Fielding, A. K. et al. Outcome of 609 adults after relapse of acute lymphoblastic leukemia (ALL); an MRC UKALL12/ECOG 2993 study. Blood 109, 944–950 (2007).

46

Héctor Tejero[1], Diana De La Iglesia[1], Javier Perales-Paton[1], Gonzalo Gómez[1] and Fatima Al-Shahrour Núñez[1]

[1]*Spanish National Cancer Research Centre, CNIO, Spain*

Predicting synergistic combinations using random forest, gradient boosting and deep neural networks

Abstract: Cancer disease is one of the leading causes of death worldwide (Ferlay et al, 2015). Drug combinations have been proposed in order to tackle innate resistance, to delay the development of acquired resistance and to reduce the toxicity and side effects of the treatment as well (Al-Lazikani et al, 2012), three of the most important mechanisms of drug failure. The computational prediction of antitumor drug combinations that could act synergistically has been extensively studied in recent years (Brown et al, 2016) as a tool to increase the repertoire of therapeutic options.

We use data coming from two high-throughput datasets studying the interaction between antitumour drugs published recently (O'Neil 2016; Koplev, 2017) and several machine learning algorithms (random forest, gradient boost machine and deep neural networks) to predict new drug combinations. The cell line gene expression and the chemical structure of the drugs were used as features. To codify the chemical structure of the moleculas we used Mol2Vec (Jaeger et al, 2018), an unsupervised machine learning algorithm used to codify molecular substructures as numerical vectors.

A comparison of the results obtained by the three algorithms is presented, as well as with previously presented results in the literature (Preuer et al, 2018) . Results are also presented for each cell line and drug combination showing a high variability of the accuracy among them.

Al-Lazikani, B., Banerji, U., & Workman, P. (2012). Nature Biotechnology, 30(7), 679–692.

Brown AS and Patel CJ. (2016) Briefings in Bioinformatics, 1–4.

Ferlay, J., Soerjomataram, I., (...) Bray, F. (2015). International Journal of Cancer, 136(5), E359–86.

Jaeger, S., Fulle, S., Turk, S., (2018) J. Chem. Inf. Model , 58 (1)

Koplev, S., Longden, J., (...) Linding, R. (2017) Cell reports, 20(12), 2784-2791

O'Neil, J., Benita, Y., (...) Shumway, S. D. (2016). Molecular Cancer Therapeutics, 15(6), 1155–1162.

Preuer, K., Lewis, R. P.I, (...) Klambauer, K. Bioinformatics, (2018) 34(9), 1538–1546

47

Cankut Cubuk[1], Marta R. Hidalgo[2], Alicia Amadoz[3], José Carbonell-Caballero[4] and Joaquin Dopazo[1]

[1]*Fundacion Progreso y Salud, FPS, Spain*
[2]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[3]*Igenomics SL. Valencia, Spain*
[4]*Center for Genomic Regulation, CRG, Spain*

Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape

Abstract: Metabolic reprogramming plays an important role in cancer development and progression and is a well-established hallmark of cancer. Despite its inherent complexity, cellular metabolism can be decomposed into functional modules that represent fundamental metabolic processes. Here we performed a pan-cancer study involving 9428 samples from 25 cancer types to reveal metabolic modules whose individual or coordinated activity predict cancer type and outcome, in turn highlighting novel therapeutic opportunities. Integration of gene expression levels into metabolic modules suggests that the activity of specific modules differs between cancers and the corresponding tissues of origin. Some modules may cooperate, as indicated by the positive correlation of their activity across a range of tumors. The activity of many metabolic modules was significantly associated with prognosis at a stronger magnitude than any of their constituent genes. Thus, modules may be classified as tumor suppressors and onco-modules according to their potential impact on cancer progression. Using this modeling framework, we also propose novel potential therapeutic targets that constitute alternative ways of treating cancer by inhibiting their reprogrammed metabolism. Collectively, this study provides an extensive resource of predicted cancer metabolic profiles and dependencies.

48

Lorena de La Fuente Lorente[1], Ana Conesa[1,2], Victoria Moreno[1] and Cristina Marti[1]

[1]*Centro de Investigación Príncipe Felipe (CIPF), Spain*
[2]*University of Florida, Spain*

Bioinformatics approach to deceipher the functional consequences of post-transcriptional regulation in neural differentiation systems

Abstract: Transcriptomes of higher eukaryotes are characterized by generating multiple isoforms per gene by using mechanisms as Alternative Splicing (AS) or Alternative Polyadenylation (APA). These mechanisms have been generally cited as responsible for creating the greater amount of protein diversity present in higher eukaryotes in relation to the overall number of genes. However, the study of how and at what extent alternative isoform expression impacts function is still difficult to address due to the lack of approaches and tools managing extensive functional annotation at isoform-resolution and integrating this isoform-resolved functional information with isoform expression dynamics.

Hence, we have developed a new pipeline for the Functional Analysis of Alternative Isoform Usage that involves the use of long-read sequencing data to defined expressed isoforms, pipelines to get extensive isoform-resolved functional annotation as well as novel methods able to profile the magnitude and nature of changes triggered for both AS and APA mechanisms.

We have applied our framework of analysis to a multiple differentiation system in mouse, which encompasses several differentiation time points from Neural Stem Cells to Oligodendrocytes and Motor Neurons. Results show how post-transcriptional mechanisms impact on the function by extensively altering the load of splicing variants through the differential inclusion of a great diversity of motifs and domains, both at the coding parts (CDS) and the un-translated regions of isoform (UTR). Moreover, we will show how our method is able to formulate de novo hypothesis about the functional outcome of differential splicing.

49

Carlos Martí-Gómez[1], Claus Vogl[2], Enrique Lara-Pezzi[1] and Fátima Sánchez-Cabo[1]

[1]Fundación Centro Nacional de Investigaciones Cardiovasculares Carlos III, Spain
[2]University of Veterinary Medicine, Vienna, Austria

A bayesian model of phenotypic evolution unveils random forces as drivers of different evolutionary rates of alternative splicing

Abstract: Alternative splicing (AS) is thought to underlie major phenotypic changes along species evolution and divergence. However, these results are based on descriptive statistics that are compatible also with neutral evolution of AS patterns. To gain further insight into the evolution of AS, one can use models of phenotypic evolution over a phylogenetic tree, such as the Browniam Motion (BM) model. However, application of such models to transcriptomic data remains challenging due to the small number of species for which data are available and the noisy nature of such data.

In this study, we use a generalization of the BM model, an Ornstein-Uhlenbeck (OU) process, to model AS evolution. The OU model combines directional evolution towards a phenotypic optima with random changes accumulating over evolutionary time as in the BM model. We extended the basic OU model to include intraespecific variability and measurement error and to take into account the binomial nature of AS data from RNA-seq experiments. We then performed bayesian inference of the underlying parameters by sampling from the joint posterior distribution using a Markov Chain Monte Carlo (MCMC) algorithm.

We showed that the limitation of using small phylogenies can be circumvented by pooling information from a relatively high number of exons and reach accurate parameter inferences using simulated data. Fitting the model to a real dataset, we find that most putatively alternatively spliced exons are actually nearly constitutively spliced and that the set of high confidence alternatively spliced exons is small. Inclusion levels of alternatively spliced exons evolved relatively slowly mainly due to a reduced stochastic rather than to stronger deterministic effects.

50

Carlos S. Casimiro-Soriguer[1], Javier Perez Florido[1], Daniel Lopez Lopez[1], Carlos Loucera[1] and Joaquin Dopazo[1]

[1]*Fundacion Progreso y Salud, FPS, Spain*

Precise sample classification of the MetaSUB environmental microbiota using functional biomarkers

Abstract: The availability of hundreds of city microbiome profiles allows the development of predictors of origin based on microbiota composition. Here we use a transformation of the conventional bacterial strain or gene abundance profiles to functional profiles that account for bacterial metabolism and other cell functionalities. We explore the use of functional profiles not only to predict the most likely origin of a sample but also to provide a functional point of view in the study the biogeography of the microbiota.

51

Carlota Rubio-Perez[1]

[1] *Vall d'Hebron Institute of Oncology, VHIO, Spain*

Pan-cancer study of tumor omics data to pinpoint new therapeutic targets modulating its interaction with the immune system

Abstract: Throughout their development, tumors are challenged by the immune system and acquirefeatures to escape its surveillance. While some of these features have been identified in recent years, they are for the most part still unknown. On the other hand, when the tumor transcriptome is profiled only tumor cells are sequenced but also stromal cells, including immune infiltrating cells. In line with this, several bioinformatic approaches that deconvolute the tumor stroma by analysing transcriptomic data have been developed. I will explain how the pancancer study of tumor omics data allows to identify potential new therapeutic targets and their implications in the response to immune checkpoint blockers.

52

Francisco Garcia-Garcia[1], Marta R. Hidalgo[1], José F. Català Senent[1], Irene Pérez Díez[1], Pablo M. Malmierca Merlo[1], Franc Casanova Ferrer[1], Raúl Pérez Moraga[1], Rubén Grillo Risco[1], Adolfo López Cerdán[1], Rubén Sánchez García[1], Marina Berenguer[2], María Pascual[3], Consuelo Guerri[3], Miguel Barquín[4], Atocha Romero[4], Rosa Farrás[5], María de la Iglesia-Vayá[6], Laura Galiana[7], José M. Tomás[7], Amparo Oliver[7] and Deborah Burks[8]

[1]*Bioinformatics and Biostatistics Unit, Principe Felipe Research Center, CIPF, Spain*
[2]*Hepatology & Liver Transplantation Unit, Hospital Universitari I Politècnic La Fe, University of Valencia & Ciberehd, Spain*
[3]*Department of Molecular and Cellular Pathology of Alcohol, Principe Felipe Research Center, CIPF, Spain*
[4]*Medical Oncology Department, Hospital Universitario Puerta de Hierro-Majadahonda, Majadahonda, Spain*
[5]*Department of Oncogenic Signalling, Principe Felipe Research Center, CIPF, Spain*
[6]*CIPF-FISABIO Joint Research Unit of Biomedical Imaging, Principe Felipe Research Center, CIPF, Spain*
[7]*Department of Methodology at Behavioral Sciences, University of Valencia, Spain*
[8]*Molecular Neuroendocrinology Unit, Principe Felipe Research Center, CIPF, Spain*

Big data approaches to detect and understand gender differences in health

Abstract: Clinical and epidemiological indicators show a large number of diseases and health scenarios where gender differences are detected but the underlying causes of these changes are still unknown. The detection of these biological, clinical and psychosocial causes would allow us to be more precise in the diagnosis and in the personalized selection of treatments for each group of patients, and to improve the selection of health interventions.

In this study, we present several big data approaches whose goal is to improve the knowledge that explain gender differences in 1) non small cell lung cancer, 2) non alcoholic fatty liver disease, 3) effect of alcohol on development, 4) schizophrenia and 5) loneliness.

For the three first scenarios, the strategy includes three phases: i) the systematic review and selection of omics studies available in public repositories such as Gene Expression Omnibus, Sequence Read Archive, The Cancer Genome Atlas... ii) the analysis of the data of each study in both signalling pathways and molecular functions contexts, highlighting those with clear alterations in their activity. iii) finally, the application of a functional meta-analysis to all results to provide a better interpretation in a Systems Biology approach. For the schizophrenia study, we explain gender differences evaluating biomedical images and for loneliness we do so through systematic review of recent literature.

From these big data approaches we provide bottom-up prototypes to generalize at a population level in a next step.

These results allow us to know the common functions in the set of the omics studies for one or more diseases, offering a greater power in the detection of signaling routes or functions of interest, which will provide us with new and more effective therapeutic targets within the framework of Precision Medicine. This knowledge would provide evidence based useful information in prevention and health stakeholders' decision-making.

53

Alba Álvarez-Franco[1], Raquel Rouco[1], Maria Tiana[1], Guadalupe Guerrero-Serna[2], Kaur Kuljeet[2], José Jalife[2] and Miguel Manzanares[1]

[1]*Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain*
[2]*Michigan University, United States*

Transcriptome and Proteome Dynamics of a Large Animal Model of Induced Atrial Fibrillation

Abstract: Atrial fibrillation, the most common arrhythmia seen by the clinician, is a highly morbid condition and the leading cause of hospitalizations for all arrhythmias. Some patients suffer relatively short (<7 days) self-terminating episodes (i.e., paroxysmal) of AF indefinitely, but a large proportion progress to long-lasting forms of AF. When AF lasts continuously for more than 7 days it is considered persistent AF. Sustained AF leads to electrical remodeling and fibrosis of the atria but the mechanism(s) remain poorly understood.

Remodeling can be due to aging, inflammation, underlying cardiac conditions, or AF itself. Current knowledge about the molecular basis of AF is mainly limited to particular genetic variants identified by genome-wide association studies and known mechanisms affecting myocardial structure, electrophysiology or signaling pathways. However, there is a clear demand for more inclusive and large-scale approaches to understand the molecular mechanisms responsible for the disease, as well as its progression and perpetuation.

Here we sought to characterize the dynamics of the transcriptome and proteome taking advantage of a well-established model of tachypacing-induced long-standing AF in the sheep. We collected samples at clinically relevant time points from anatomically distinct regions of both atria, together with cardiomyocytes isolated from these tissues. We applied multiple co-inertia analysis to integrate multiple layers of information and build a regulatory model that explains the observed changes in AF. Our approach overcomes limitations like the low correlation between transcription and translation, as well as the curse of dimensionality. Our data-driven analysis consolidates most of the previous knowledge about AF pointing at structural remodelling, specific channel imbalance or inflammatory pathways. Moreover our results suggest impaired chromatin remodelling and defects in mitochondrial biogenesis as relevant and consistent changes. Altogether, these findings are a valuable resource for the study of the molecular dynamics underlying the progression from paroxysmal to persistent AF.

54

Sandra Alandes Esteve[1], Francisco Garcia-Garcia[2], Gema Garcia-Garcia[1] and José María Millán[1]

[1]*Bioinformatics and Biostatistics Unit, CIPF; IIS La Fe; CIBERER, Spain*
[2]*Bioinformatics and Biostatistics Unit, Principe Felipe Research Center, CIPF, Spain*

Application of genomics for a precision medicine in hereditary dystrophies of the retina: Usher syndrome

Abstract:

Usher syndrome (USH) is a rare autosomal recessive disease and the most common inherited form of combined visual and hearing impairment. From a clinical point of view, USH is classified into three types and each one of them associates a series of responsible genes: type I (MYO7A, CDH23, PCDH15, USH1C, USH1G, CIB2), type II (USH2A, GPR98, DFNB31) and, type III (CLRN1, HARS). Other genes implicated in this disease have also been detected PDZD7 and CEP250. The Next Generation Sequencing (NGS) has shown that the different clinical types are not genetically sealed and that there is still an important group of patients where we do not know the genes involved in this disease.

This work has a double objective: 1) clinical and molecular characterization of a cohort of patients with USH using gene panels and Whole Exome Sequencing (WES); 2) identification of new genes and molecular mechanisms, responsible for

the USH with Whole Genome Sequencing (WGS) together with transcriptome (RNA-Seq) for patients in whom, after the panel study and WES, mutations in known genes have not been found .

To achieve these objectives, we have developed a panel that includes all the known genes that cause USH by NGS and that allows us to detect at least one mutation in 80% of our patients and the two responsible mutations in 70%. The combination of WES, Copy Number Variation (CNV), WGS and RNA-seq will allow the development of a clinical application in the diagnosis and treatment of this disease, establishing a circuit within the framework of Personalized Medicine.

55

José F Català-Senent[1], Marta R Hidalgo[1], Marina Berenguer[2] and Francisco Garcia-Garcia[1]

[1]*Bioinformatics and Biostatistics Unit, Principe Felipe Research Center, CIPF, Spain*
[2]*Hepatology & Liver Transplantation Unit, Hospital Universitari i Politècnic La Fe, University of Valencia & CIBERehd, Spain*

Gender differences in non-alcoholic fatty liver disease from an omics approach

Abstract: Non-alcoholic fatty liver disease (NAFLD) is one of the most important illnesses that can lead to advanced liver disease, cirrhosis and hepatocellular carcinoma (HCC). The prevalence of NAFLD has increased dramatically in recent years and is estimated at 24% throughout the world[1]. As also occurs in other diseases (such as lung cancer or some autoimmune diseases), NAFLD affects in different ways men and women, being more frequent in the former than in the latter[2|3].

In order to understand the grounds for these differences and open the door to more specific treatments, this work aims to identify the molecular mechanisms affected differentially between men and women in the NAFLD. For this, an in silico approach will be carried out, with the goal of perform a meta-analysis combining the information of a selection of omic studies.

To achieve this objective, we started with a systematic review of public databases to found omic studies that take into account the gender of patients affected by the disease, followed by the data download of all the selected studies and the performance of an individual exploratory analysis to find non-expected patterns in the data. Then, in each study, two groups were distinguished (pairwise analysis if more than two groups are present in the study) to carry out a differential intensity assessment of all comparisons. The results of each were enriched with a functional profiling in order to find out functional blocks enriched in any of the conditions. Simultaneously, also for each study, Hipathia methodology [4] was used to look into the affected mechanisms by calculating the activity values of both routes and functions and obtaining the differences between groups.

Finally, we carried out a meta-analysis with the results of each method to detect functional results of global interest, reducing the effect of the specific experiments. This methodology also checks the consistency of the experiments and generates an estimate of the effect having a greater statistical power than that obtained for each experiment separately. We will also perform several functional meta-analysis for each annotation (Gene Ontology, KEGG pathways and KEGG reactions).

The results of this work allow us to understand how NAFLD affects men and women differentially, and will help design more personalized treatments for each patient.

References

[1] Anstee QM et. al. Progression of NAFLD to diabetes mellitus, cardiovascular disease or cirrhosis. Nat. Rev. Gastroenterol. Hepatol. 10, 330–344 (2013).

[2] Pan JJ and Fallon MB. Gender and racial differences in nonalcoholic fatty liver disease. World J Hepatol. 6(5), 274–283 (2014).

[3] Ballestri S et al. NAFLD as a Sexual Dimorphic Disease: Role of Gender and Reproductive Status in the Development and Progression of Nonalcoholic Fatty Liver Disease and Inherent Cardiovascular Risk. Advances in Therapy 34(6), 1291–1326 (2017).

[4] Hidalgo MR et. al. High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget 3(8), 5160–5178 (2017).

56

Cankut Cubuk[1], Marta R. Hidalgo[2], Carlos Loucera[1], Kinza Rian[1], Maria Peña-Chilet[1], Matias M. Falco[1], Isabel A. Nepomuceno-Chamorro[3], Helena Molina-Abril[4] and Joaquin Dopazo[1]

[1]*Fundacion Progreso y Salud, FPS, Spain*
[2]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[3]*Dpto. Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Spain*
[4]*Departamento de Matemática Aplicada I. Universidad de Sevilla, Spain*

Interpreting genomic profiles with mechanistic models of pathways

Abstract: Despite the increasing availability of genomic and transcriptomic data, there is still a gap between the detection of perturbations in gene expression and the understanding of their contribution to the molecular mechanisms that ultimately account for the phenotype studied. Disease maps (http://disease-maps.org/projects) and other generic maps that recapitulate cell signaling, metabolism and functionality (e.g. KEGG, WikiPathways, etc.) offer a detailed picture on the complex network of interrelationships among genes that result in cell activity decisions.

Alterations in the in the functioning of such networks are behind the initiation and progression of many diseases, including cancer. The wealth of available knowledge on biological networks can therefore be used to derive mechanistic models that link gene expression perturbations to changes in metabolic, signaling, etc. activities that provide relevant clues on molecular mechanisms of disease and drug modes of action (MoA).

Here we describe simple mechanistic models of signaling (hipathia) [1] and metabolic (Metabolizer) [2, 3] activity based on modules defined as functionally substantiated circuits within pathways. Such mechanistic models that consider the interaction network have proven to be more sensitive and specific than other alternatives proposed [4] and have demonstrated to be highly precise in revealing the molecular bases of cancer hallmarks [1, 5], predicting drug effect [6] or predict new therapeutic targets in cancer [2, 3].

The models have been implemented as web-based applications, which offer intuitive, easy-to-use interactive interfaces to analyze differences in pathway activities that can also be used for class prediction and in silico prediction of Knock-Out (KO) effects [7].

A working group composed by the authors is actively working on the application of machine learning methods to provide meaningful interpretations of genomic data helping to understand the subjacent molecular mechanisms of disease and drug modes of action (MoA).

We provide different types of validations of some of the predictions made by the models.

Metabolizer can be found at: http://metabolizer.babelomics.org.

Hipathia can be found at: http://hipathia.babelomics.org. A Bioconductor/R package can be found at: http://bioconductor.org/packages/devel/bioc/html/hipathia.html

Pathact, that predict the effect of KOs on signaling pathways, can be found at: http://pathact.babelomics.org.

References

1.Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J: High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget 2017, 8(3):5160-5178.

2.Cubuk C, Hidalgo MR, Amadoz A, Pujana MA, Mateo F, Herranz C, Carbonell-Caballero J, Dopazo J: Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape. Cancer Res 2018, In press.

3.Cubuk C, Hidalgo MR, Amadoz A, Rian K, Salavert F, Pujana MA, Mateo F, Herranz C, Caballero JC, Dopazo J: Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. bioRxiv 2018:367334.

4.Amadoz A, Hidalgo M, Cubuk C, Carbonell-Caballero J, Dopazo J: A comparison of mechanistic signaling pathway activity analysis methods. Brief Bioinform 2018, In press.

5.Hidalgo MR, Amadoz A, Cubuk C, Carbonell-Caballero J, Dopazo J: Models of cell signaling uncover molecular mechanisms of high-risk neuroblastoma and predict disease outcome Biology direct 2018, In press.

6.Amadoz A, Sebastian-Leon P, Vidal E, Salavert F, Dopazo J: Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity. Scientific reports 2015, 5:18494.

7.Salavert F, Hidalgo MR, Amadoz A, Cubuk C, Medina I, Crespo D, Carbonell-Caballero J, Dopazo J: Actionable pathways: interactive discovery of therapeutic targets using signaling pathway models. Nucleic Acids Res 2016, 44(W1):W212-216.

57

Patricia Sebastian-Leon[1] and Patricia Diaz-Gimeno[1]

[1]*IVI-RMA Global Foundation, Spain*

Gene selection meaning comparison of endometrial transcriptomic signatures in Reproductive Medicine

Abstract: Since the publication of the guidelines for microarray-based predictive models in the MACQ-II project (Shi et al. 2010), signature predictor from microarray data have risen as useful tools to improve diagnosis in medicine. In addition, several studies have demonstrated the ability of using gene-expression signatures derived from DNA microarray data to classify between different diseases and predict toxicity or clinical response (Chibon et al. 2013). However, gene signatures are not homogeneous and usually signatures associated to the same disease have not many genes in common. Many studies (Shen et al. 2008; Shi et al. 2010) compared signatures in a functional context, showing that although they share few genes, there is a high overlap between the functions they shared.

In reproductive medicine, there were exist a controversy about the Recurrent Implantation Failure (RIF) causes: a window of implantation (WOI) displacement (Valdes et al. 2017) or disruption (Macklon 2017). In last years, sixteen heterogeneous signatures of different sizes were proposed for endometrial evaluation in both contexts, but their predictive value and meaning has not been compared. Comparison of them will help clinicians to improve the diagnosis using a new taxonomy of RIF.

Higher number of genes implies higher predictive value (Mitra et al. 2002), so the main challenge of this work was to compare the predictive value between signatures of different sizes and to determine if the predictive value obtained was higher or not than expected by chance, evaluating, therefore, the gene signature selection meaning related to the trait.

Individual signatures predictive value were estimated using caret R-package by selecting the model with the highest accuracy from a cross-validation (10-fold). Accuracy and kappa values were estimated and compared with the obtained

from random signatures of the same size to achieve if each predictive value is higher than expected by chance or not. The distance between the random signatures accuracy confidence interval and the real accuracy estimated indicates how much better this signature is than expected by chance. This measure is directed related with the gene signature selection meaning and can be used to compare signatures from different sizes in the same context.

Once this setback was got over, we were able to understand the gene signature selection meaning related to displacement and disruption demonstrating for the first time that molecular causes of RIF are different and also could coexist in the same patient.

Valdes et al. Fertility and sterility 108.1 (2017): 15-18.

Macklon. Fertility and sterility 108.1 (2017): 9-14.

Shen et al. BMC Medical Genomics 1 (2008) 1.

Chibon. European journal of cancer 49.8 (2013): 2000-2009.

Shi et al. The pharmacogenomics journal 10.4 (2010): 310.

Mitra et al.. IEEE transactions on pattern analysis and machine intelligence 24.3 (2002): 301-312.

Sebastian-Leon, P., et al. Human Reproduction 33.4 (2018): 626-635.

*This work was presented partially in the 33rd Annual Meeting of the European Society of Human Reproduction Biology (ESHRE) and published in Human Reproduction (Sebastian-Leon et al. 2018).

58

Mario Solís[1], Beatriz Ruz[1], Elena Vallespín[1], Pablo Lapunzina[1] and Angela Del Pozo[1]

[1]*Instituto de Genética Médica y Molecular (INGEMM) - Hospital Universitario La Paz Madrid – IdiPaz, Spain*

An overview of a quality control protocol for high-throughput sequencing data in a clinical laboratory

Abstract: In the day of precision medicine, human genetics diagnostic laboratories are using high-throughput sequencing (HTS) of patient's DNA samples in the diagnostic routine to discover probable causal mutations. The clinical bioinformaticians analyze the HTS data generated by these experiments and provide a report to clinical geneticist containing functional enriched variants discovered in the sample aiming to support physicians to filter and select a set of candidate variants for further validation. The diagnostic power of the analysis could, however, be compromised by multiple issues. A routine Next-Generation Sequencing (NGS) wet-lab workflow usually involves the handling of multiple samples from different patients at the same time causing frequent human errors that produce cross-contamination or swapping of samples; furthermore, clinical laboratories are storing, receiving and managing hundreds of samples for different tests on a daily basis, so opportunity for error arises, such as mislabeling samples. Even if there are no human errors in any of the multiple protocols, mechanical failures and defective reagents are, unfortunately, also possible. To ensure the suitability of the analysis of the samples for diagnosis, the clinical bioinformaticians must perform a strict quality assessment.

In this work, we present our quality control protocol, which allows the monitorization of all the steps in the variant discovery analysis pipeline with the goal to establish whether a sample is suited for diagnosis purposes. Our protocol is designed to measure various metrics throughout the bioinformatics analysis to then perform a comparison with our historical data and warn if some of the quality markers deviate from the expected distribution. The reference ranges of a set of 80 quality parameters have been established after the study of over 12000 samples, sequenced in different platforms in our laboratory over the last 3 years.

The quality evaluation of this historical dataset has provided valuable insight into quality markers behavior under different experimental conditions; examples include: the increase of rate of bases mismatching the reference as the number of called variant decreases with expired reagents, or how the median of the ratios of the alternative allele reads of all the heterozygous calls is below 0.4 when the sequenced sample is contaminated.

As a result of this work, we have been able to establish quality thresholds that are supported in this multi-factor model. With this protocol, problematic sequencing runs have been identified in the past; however, the evaluation of the metadata associated with each sample is also required to identify some errors, such as sample mislabeling or sample swapping. To detect those events, we identify sex-specific markers, and we infer relatedness amongst samples.

To conclude, this quality assessment is a powerful tool to ensure: (1) that the quality of the sequencing run and the bioinformatics data generated are suitable for a clinical study, (2) that the diagnostic power of the analysis has not been compromised and (3) that the discovered variants are from the expected individual.

59

Rafael Hernández-de-Diego[1], Tarazona Sonia[1], Carlos Martínez-Mira[1], Leandro Balzano-Nogueira[2], Pedro Furió-Tarí[1], Georgios Pappas[3] and Ana Conesa[1,2]

[1]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[2]*University of Florida, United States*
[3]*University of Brasilia, Brazil*

PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data

Abstract: The increasing availability of multi-omic platforms poses new challenges to data analysis. Joint visualization of multi-omics data is instrumental in better understanding interconnections across molecular layers and in fully utilizing the multi-omic resources available to make biological discoveries. We present here PaintOmics 3, a web-based resource for the integrated visualization of multiple omic data types onto KEGG pathway diagrams. PaintOmics 3 combines server-end capabilities for data analysis with the potential of modern web resources for data visualization, providing researchers with a powerful framework for interactive exploration of their multi-omics information. Unlike other visualization tools, PaintOmics 3 covers a comprehensive pathway analysis workflow, including automatic feature name/identifier conversion, multi-layered feature matching, pathway enrichment, network analysis, interactive heatmaps, trend charts, and more. It accepts a wide variety of omic types, including transcriptomics, proteomics and metabolomics, as well as region-based approaches such as ATAC-seq or ChIP-seq data. The tool is freely available at www.paintomics.org

---

60

Marta Coronado-Zamora[1], Irepan Salvador-Martínez[2], David Castellano[3], Antonio Barbadilla[1] and Isaac Salazar-Ciudad[4]

[1]*Universitat Autònoma de Barcelona, Spain*
[2]*University College London, United Kingdom*
[3]*Aarhus University, Denmark*
[4]*University of Helsinki, Finland*

Mapping natural selection through D. melanogaster development: Towards a population -omics synthesis

Abstract: The current genomic era has led to the paradoxical situation in which much more evidence of selection is available at the genomic level than at the phenotypic one. The first high-resolution map of natural selection showed that natural selection is rampant in the genome of D. melanogaster [1]. In contrast to Kimura's neutral theory expectations, between 30 and 50% of mutations that become incorporated into the genome of this species are adaptive [2]. Which is the adaptive significance of these mutations? Natural selection acts primarily on the phenotypic properties of organisms and only, to the extent that these properties are heritable, does it act secondarily on the genotype. The action of the selection in the whole phenotype of an organism has not yet been approached, neither any study integrating both levels of selection at a genomic scale. Here we carry out an organismal selection-phenotype-genotype integration, specifically we draw an exhaustive map of selection acting on the complete embryonic anatomy of species D. melanogaster. We have developed a new approach that integrates high-throughput data of genomic variation from the DGRP [1], gene expression and development [3] to map adaptation over the entire embryo's anatomy of D. melanogaster. The selection map indicated that selective constraint is pervasive over most of the embryo's anatomy, specifically affecting anatomical organs related to the nervous and digestive systems. Adaptation is found in the structures that also show evidence of adaptation in the adult, the immune and reproductive systems. It is also found a relaxation of selection in the first stage, attributed to the maternal-effect genes. Finally, it was observed that genes that are expressed in one or a few different anatomical structures are younger and have higher rates of evolution, unlike genes that are expressed in all or almost all structures. The integration of population genomics with other phenotypic multi-omics data is necessary to obtain a global picture of how adaptation occurs [4].

[1] Mackay, et al. (2012) Nature. 482:173–178

[2] Eyre-Walker (2006) Trends Ecol Evol. 21:569–575

[3] Tomancak, et al. (2007) Genome Biol. 8(7):R145

[4] Casillas & Barbadilla (2017) Genetics. 205(3): 1003–1035

61

Beatriz Ruz-Caracuel[1], Carlos Rodríguez Antolín[1], Isabel Vallcorba[1], Mario Solís[1], Ángela Del Pozo[1], Adela Escudero[1] and Antonio Pérez Martínez[1]

[1]*INGEMM-Hospital Universitario La Paz-IdiPaZ, Spain*

Detection of somatic gene rearrangements in custom Next Generation Sequencing (NGS) DNA Panels for pediatric leukemia

Abstract: Pediatric Acute Lymphoblastic Leukemias (ALL) can be classified according to the fusion protein expressed. Moreover, fusion protein define a prognosis and guide the treatment. For this reason, it is important to be able to detect them in a rigorous way.

Traditionally, gene fusion detection has been made by fluorescence in situ hybridization (FISH), RT-PCR or RNA-seq. RNA-seq allow to detect directly the fusion transcript, but it is a difficult technique due to the instability of the molecule. Targeted DNA next-generation sequencing (NGS) used for detecting single nucleotide variants or small indels can also be adapted to detect chromosomal rearrangements with the help of different bioinformatics softwares. Finding a way

to detect gene rearrangements with SNVs and indels in DNA using only one technique, can be very useful in cancer in which there is a small sample size.

We have compared the reproducibility of the results obtained with RNA-seq in DNA-seq and the possible characterization of translocations with a custom clinical DNA panel. This panel contains exons from most of the genes and few complete genes relevant for ALL diagnosis. We have explored the use of different algorithms (Manta, Socrates, GeneFuse and GRIDSS) for detecting chromosomal rearrangements in 6 samples with a translocation already validated by FISH technique.

Firstly, we have confirmed the proof of concept with 3 control samples with a gene translocation validated by FISH, confirming that we obtained the same results using both RNA-seq and DNA-seq. Then, from our 6 samples, we were able to detect only 3 translocations using the custom DNA panel. The gene rearrangements not detected could be due to several reasons: (a) the breakpoint of the translocation was not in the gene regions included in the panel or (b) the sequencing coverage obtained was not enough to detect translocations present in a low mosaicism. Characterizing the breakpoint regions of the rearrangements of interest would allow to build custom DNA to detect both somatic gene fusion and point mutations to guide treatment using a single detection method.

62

Maria Peña-Chilet[1], Gema Roldan[1], Rosario Carmona[1], Francisco García-García[2], Jose L. Fernández-Rueda[1], Javier Perez Florido[1] and Joaquin Dopazo[1]

[1]*Clinical Bioinformatics Area, Fundacion Progreso y Salud, FPS, Spain*
[2]*Bioinformatics and Biostatitics Unit. Prince Felipe Research Center, CIPF, Spain*

The database of Spanish population variability: a crowdsourcing resource for collecting local genomic data.

Abstract: The knowledge of the genetic variability of the local population has revealed as a critical factor for the discovery of new disease variants. However, reference repositories of genetic variability of normal population are scarce and typically contain mixtures of different ethnical groups and geographic locations. Here we present the Collaborative Spanish Variability Server (CSVS), a user-friendly web interface to a database containing the frequencies of the naturally occurring variants found in the Spanish population. Currently, more than 1600 unrelated Spanish individuals have been used to derive the genetic variant frequencies of the Spanish population. The genomic data are aggregated in a way that does not compromise the identity of the donors and the access makes virtually impossible individual re-identification. Interestingly, this database has been populated with data produced in diverse research projects. For any submission the system checks that it corresponds to the local (Spanish) population and that the individual is not related to any other in the database (at least not to the degree of cousin or closer). The information provided by the database consists on aggregated counts of allele frequencies in groups representing ICD10 disease categories. In this way, ICD10 groups can be used to obtain pseudo-controls for any disease by simply excluding its ICD10 category. This is the first local repository of variability entirely produced by crowdsourcing and constitutes an example for future initiatives to characterize of local variability worldwide.

The database can be found at: http://csvs.babelomics.org

63

Carlos Carretero-Puche[1,2], Beatriz Soldevilla[1,2], Gonzalo Gomez-Lopez[2], Julia Martínez-Pérez[3], María C. Riesco-Martinez[3], Beatriz Gil-Calderón[1,2], Fátima Al-Shahrour[2] and Rocío García-Carboner[3]

[1]*Instituto de Investigacion i+12, Spain*
[2]*Centro Nacional de Investigación Oncológica, CNIO, Spain*
[3]*Hospital Universitario Doce de Octubre, Spain*

Resistance mechanisms to oxaliplatin in metastatic colorectal cancer and the importance of the molecular scenario.

Abstract: MOTIVATION: Metastatic colorectal cancer (mCRC) is diagnosed in the 25% of the patients and has a critical prognosis with a 5-year survival rate of less than 10%. One of the first-line treatment widely used in mCRC is based on the combination of oxaliplatin and fluoropyrimidines, but about 50% of patients show innate or acquired resistance1. Therefore, it is of capital importance to identify genes and mechanisms involved in this resistance that allow us to discriminate mCRC patients which could obtain a great benefit from these therapies. On the other hand, CRC is not only heterogeneous to the drug response, but also at molecular features. This heterogeneity has been collected inside the Consensus Molecular Subtypes (CMS). Colorectal tumors can be classified into four Consensus Molecular Subtypes (CMSs); Immune (CMS1), Canonical (CMS2), Metabolic (CMS3) and Mesenchymal (CMS4), which present different clinical, molecular, functional and immune patterns2.

AIM: The aim of our study was to assess if the genes and molecular pathways that are differentially expressed between responders and non responders to oxaliplatin in mCRC depends on the CMSs.

METHODS: We analyzed gene-expression profile from 86 mCRC tumors using RNA-sequencing by Nextpresso RNA-seq pipeline3. The samples were classified into CMS groups by CMSclassifier R package. Genset enrichment analysis (GSEA) was performed with 0.05 FDR threshold.

RESULTS: We observed that 55 genes were differentially expressed (FDR < 0.05) between all responders and non responders. Only 3 of these genes appeared in the same analysis considered CMS3 tumors (62 genes), 4 in CMS4 tumors (14 genes) and not one in CMS2 tumors (80 genes). In addition to that, 6 genes were differentially expressed both in CMS2 and CMS3 tumors. We also applied a GSEA preranked analysis in order to obtain functional differences between both phenotypes. We saw that oxidative phosphorylation, ribosome and adipogenesis gensets were enriched in all responder patients, while TNFα signaling via NFκ-B, inflammatory response, interferon gamma response, E2F targets, G2M checkpoint and EMT gensets were enriched in all no responder patients. Surprisingly, EMT genset appeared more enriched in CMS2 responders, while E2F targets and G2M checkpoints appeared more enriched in CMS2 and CMS3 non responders. Furthermore, Ribosome and Interferon gamma response gensets were enriched in CMS3 responders. In CMS4 oxidative phosphorylation genset was more enriched in responder patients and TNFα signaling via NFκ-B, inflammatory response, interferon gamma response and EMT gensets were enriched in non responders.

CONCLUSIONS: We have applied a bioinformatic approach to study differential expression and functional patterns between responders and non responders to oxaliplatin. We found that differences in genes expression and gensets enrichment are not the same in all CMSs. These findings show the relevance of molecular scenario in the resistance mechanisms to oxaliplatin in mCRC patients, that should be considered in future studies of predictive biomarker genes response.

REFERENCES

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2017. CA. Cancer J. Clin. 67, 7–30 (2017).

2. Guinney, J. et al. The Consensus Molecular Subtypes of Colorectal Cancer. Nat. Med. 21, 1350–1356 (2015).

3. Grana, O., Rubio-Camarillo, M., Fdez-Riverola, F., Pisano, D. G. & Glez-Pena, D. Nextpresso: Next Generation Sequencing Expression Analysis Pipeline. Curr. Bioinforma. 12, (2017)

64

Teresa Rubio[1], Carmina Montoliu[2], Vicente Felipo[1], Sonia Tarazona[1] and Ana Conesa[1,2]

[1]*Centro de Investigacion Principe Felipe, CIPF, Spain*
[2]*INCLIVA, Spain*
[3]*University of Florida, United States*

Integrative multi-omics analysis to explain immune alterations in minimal hepatic encephalopathy patients

Abstract: Minimal hepatic encephalopathy (MHE) produces slight cognitive impairment, attention deficits, psychomotor slowing and impaired coordination in cirrhotic patients. Two million people in European Union show MHE which becomes in a serious health, social and economic problem. No specific biomarkers in liquid biopsy exist to detect the appearance of MHE up to now, so most patients remain undiagnosed and untreated worldwide. With the aim of compensate for this lack, our study includes a multi-omics dataset (transcriptomics, metabolomics and panel of interleukins) from human peripheral blood cells. We perform an integrative analysis to elucidate the appearance of neurological alterations from liver damage via peripheral inflammation.

After differential expression and enrichment analysis of transcriptomics data we obtained different GO terms disrupted in MHE (i.e.: olfactory receptor dysfunction, NF-kB or Wnt T cells signalling pathways). With this information, we integrated transcriptomics with metabolomics and interleukins making correlation networks based on partial least square (PLS) regression. Selecting a subset of the most variable genes between groups of patients, we connected them with the most important metabolites/interleukins. This is a new strategy which allows finding covariance between long-non-coding genes, transcription factors, coding genes, metabolites and interleukins.

65

Jesus Murga Moreno[1], Marta Coronado-Zamora[1], Alejandra Bodelón[1], Antonio Barbadilla[1] and Sonia Casillas[1]

[1]*Universitat Autònoma de Barcelona, UAB, Spain*

PopHumanScan: the online catalog of human genome adaptation

Abstract: Since the split with chimpanzees, and especially since the migrations that lead humans to colonize almost every single place on Earth, we have been exposed to environmental and social challenges that have shaped our genomes through the action of natural selection. The availability of a comprehensive worldwide nucleotide variation data set from the 1000 Genomes Project, contained in PopHuman database, provides the human lineage with an invaluable resource, allowing the testing of molecular population genetics hypotheses and eventually understand the evolutionary dynamics of genetic variation in human populations. Here we analyze a battery of population genetics metrics calculated along the human genome from PopHuman to: (i) describe the patterns of diversity and divergence for different chromosomes and genomic regions showing distinct rates of recombination, and (ii) perform a genome-wide scan of selection to identify regions of the genome that have been subjected to either recent sweeps or recurrent selection since the split between our species and chimpanzees.

Our genome-wide scan of selection includes 2,859 genomic regions showing signatures of positive selection, of which, 1,453 (~50.8%) overlap genes. Well-known examples of human genetic adaptation published elsewhere are included in the catalog, as well as hundreds of other interesting candidates that will require more thoroughly analyses. The catalog facilitates comparisons of each signature of selection with empirical distributions of the corresponding DNA diversity metric across the human genome and among populations and structural and functional annotations of the region. Our catalog is freely available at pophumanscan.uab.cat and presented as a collaborative database to compile and annotate adaptation events along the human evolutionary history.

66

Jose L. Fernández-Rueda[1], Antonio Rueda[2], Javier Lopez[2], Ignacio Medina[3], Javier Perez Florido[1] and Joaquin Dopazo[1]

[1]*Clinical Bioinformatics Area, Fundacion Progreso y Salud, FPS, Spain*
[2]*Genomics England, Charterhouse Square, London, United Kingdom*
[3]*HPC Service, UIS, University of Cambridge, Cambridge, United Kingdom*

The Module of Personalized Medicine: the pioneer experience of including genomic data into the Andalusian Health System (SAS)

Abstract: The Module of Personalized Medicine: the pioneer experience of including genomic data into the Andalusian Health System (SAS)

José L Fernández-Rueda1, Antonio Rueda2, Javier Lopez2, Ignacio Medina3, Javier Perez-Florido1*, Joaquin Dopazo1,4,5*

1 Clinical Bioinformatics Area. Fundación Progreso y Salud (FPS). CDCA, Hospital Virgen del Rocio. Sevilla. Spain

2 Genomics England, Charterhouse Square, London, EC1M 6BQ, UK

3 HPC Service, UIS, University of Cambridge, Cambridge, UK

4 Bioinformatics in Rare Diseases (BiER). Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER). FPS. Hospital Virgen del Rocio. Sevilla. Spain;

5 INB-ELIXIR-es, FPS, Hospital Virgen del Rocío, Sevilla, Spain

* Corresponding authors

The Andalusian community launched the first local genomic project in Spain in 2011, the Medical genome Project, and since then has cumulated an enormous experience in the management of genomic data. This experience, in combination with the largest population in Europe (8.5 million) covered by a universal electronic health record (eHR) fostered the innovative Andalusian Personalized Medicine initiative. The novelty in this initiative is the implementation of bioinformatics tools into the corporative systems of the SAS that link the genomic data of patients to their clinical records. In this way, in addition to use massive sequencing for diagnostic (the rare diseases initiative) or treatment recommendation (the ongoing cancer genomics initiative) purposes, the genomic data remain into the system, linked to patient's eHRs. This transforms the database of the SAS in an immense potential prospective clinical study in which, as more genomic data and clinical data are cumulated along the time, the genomic variants of the patients can be associated to different clinical endpoints of interest. This allows an unprecedented capability for biomarker discovery based in growing information storage. The population health database provides a framework for this discovery potential.

This is possible because of the introduction of the MMP into the corporative informatics systems of the SAS, coupled to the sequencing unit, equipped with Illumina NextSeq 550 and HiSeq 2500 sequencers, located in the Hospital Virgen del Rocío (Sevilla). The MMP is a sophisticated web interface, based in previous developments [1, 2], that allows semi-automated diagnosis of genetic diseases. In the pilot phase we experienced a reduction of time to diagnosis since blood extraction from over three weeks to 8 hours. MMP interfaces the advanced genomic data management system OpenCGA, used in the Genomics England 100,000 genomes project , and the knowledge database CellBase [3]. The MMP provides, for the first time, the possibility of carrying out massive sequencing-based diagnosis (and soon cancer treatment recommendation) within a public health system. MMP is based in the IVA architecture, co-developed between the Clinical Bioinformatics Area and the University of Cambridge.

References

1.Aleman A, Garcia-Garcia F, Medina I, Dopazo J: A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications. Nucleic Acids Res 2014, 42(Web Server issue):W83-87.

2.Aleman A, Garcia-Garcia F, Salavert F, Medina I, Dopazo J: A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. Nucleic Acids Res 2014, 42(Web Server issue):W88-93.

3.Bleda M, Tarraga J, de Maria A, Salavert F, Garcia-Alonso L, Celma M, Martin A, Dopazo J, Medina I: CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. Nucleic Acids Res 2012, 40(Web Server issue):W609-614.

67

Victor Jimenez Jimenez[1], Miguel Sanchez Alvarez[1], Fátima Sanchez-Cabo[1] and Miguel Angel del Pozo[1]

[1]*Spanish National Center for Cardiovascular Research, CNIC, Spain*

A boolean network to understand the role of Caveolin-1 (CAV1) as a potential integrator of mechanoadaptation, signaling and metabolism

Abstract: Caveolin-1 (CAV1) participates of multiple processes in the cell, including plasma membrane organization (through structures termed caveolae), lipid metabolism and trafficking, signaling integration and regulation, and mechanotransduction, but we are far from fully understanding its contribution to cell behavior and disease. As a signaling molecule, CAV1 can directly interact with many ligands that regulate fundamental cellular processes such as cell cycle or mechanotransduction, however, one of CAV1 fundamental contributions to cellular physiology is due to its ability to induce specific membrane nanodomains which are responsible for the indirect modulation of the activity of other protein receptors and effectors by controlling their spatiotemporal distribution along the different cellular membranes.

How CAV1 exerts its multiple and, sometimes, contradictory functions is still unknown. In order to unravel this question a CAV1-centered, systems-biology, global picture of cell physiology is needed. To arrive to such a picture, this project adopts a bioinformatics approach to analyze different transcriptomic datasets present in public repositories. A CAV1-dependent gene expression boolean network is built using reverse engineering from publicly available microarray data. These data comes from transcriptomic analysis of tissue gene expression of CAV1 KO and WT mice performed in different laboratories and with different technologies that we have integrated in only one network. The obtention of such a network have allowed us to perform “in silico” experiments that have shed some light on CAV1 downstream effects and that are a primary step towards experimental validation.

68

Isabel A. Nepomuceno Chamorro[1], Cankut Cubuk[2], Marta R. Hidalgo[3], Carlos Loucera[2], Kinza Rian[2], Maria Peña-Chilet[2], Matias M. Falco[2], Helena Molina-Abril[4] and Joaquin Dopazo[2]

[1]*Dpto. Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Spain*
[2]*Fundacion Progreso y Salud, FPS, Spain*
[3]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[4]*Dpto. de Matemática Aplicada I. Universidad de Sevilla, Spain*

Identifying pathway associations using neural networks.

Abstract: Motivation: pathway repositories such as KEGG provide annotations of biological processes. However, functional references for global pathway relationships are limited [1]. In this work, we perform a step towards this direction by providing a method to show a matrix of influence that can help the interpretation of functional interaction between pathways by searching the latent space from signaling circuit activities provided by Hipathia [2].

Method: we used a class of deep neural network models named autoencoder to learn a meaningful latent space that could be used in the prognosis of disease. Furthermore, the model was executed in an iterative way: each feature of the dataset (subcircuit) is removed and the autoencoder is generated. In this way, the set of models show the influence of every single signaling circuit in the apoptosis pathway.

Results: We analyzed gene expression values from the TCGA pan-cancer recoded into signalling circuit activities using Hipathia for the Apoptosis pathway. The results show a good separation between classes using the encoding features in the two internal nodes of the middle layer of the neural network. This suggests that the neural network is properly representing biological knowledge.

Conclusions: This is a preliminary result that just scratched the surface of what is possible. Although this analysis seems to be promising, there is an important limitation. It is focused on one pathway. The influence between subcircuits has more interest taking the signaling circuit activities of all the pathways. Furthermore, this work also shows the potential that deep learning have for identifying global pathway relationships.

1. Pita-Juárez Y, Altschuler G, Kariotis S, Wei W, Koler K, Green C, Tanzi R, Hide W: The pathway coexpression network: revealing pathway relationships. PLOS Computational Biology 2018, 14: e1006042

2. Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J: High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget 2017, 8(3):5160-5178.

69

David Gómez-Peregrina[1], José M. Juanes[1,2], Eva Panadero[3], Santiago Gala[3], Vicente Arnau[4], Javier Chaves-Martinez[2,5], Ana Barbara Garcia-Garcia[2,5] and Pablo Marin-Garcia[1]

[1]*MGviz.org, Spain*
[2]*UGDG-INCLIVA, Spain*
[3]*kanteron-systems, Spain*
[4]*University of Valencia, Spain*
[5]*CIBERDEM, Spain*

MGvizPM: Precision Medicine web reports made 'siimple'

Abstract: Medical Genomics Visualization Group (MGviz), Siimple OSS, Seqplexing and Kanteron Systems have jointly developed NGS data analysis workflows that create automatic technical reports for precision medicine with fully integrated QC and LIMS procedures. Our genetic and pharmacogenetic data can be easily integrated in HIS systems and use HL7 standard protocols.

We have developed a full suit of open source tools in Python, R and MERN stack for clinical bioinformatics as a service. These tools include serving variant annotation, interactive selection tools, reannotation and automatic clinical reports generation.

We are doing trials to deploy this service in a cloud platform for creating a service for analyzing NGS gene panels and exomes in a clinical context.

70

José M. Juanes[1,2], Alex Tejada-Cánovas[2], David Gómez-Peregrina[2], Vicente Arnau[3], Felipe J. Chaves[1,4], Ana Bárbara García-García[1,4] and Pablo Marín-García[1]

[1]*UGDG-INCLIVA, Spain*
[2]*MGviz.org, Spain*
[3]*University of Valencia, Spain*
[4]*CIBERDEM, Spain*

SiimpleHub: A centralized system for the authentication and authorization of bioinformatics web applications

Abstract: Genetic data have the potential to be highly sensitive, so security and privacy are extremely important in web applications that manage, collect and share this type of data.

As part of MGviz and Siimple, and with the collaboration of the “Unidad de Genómica y Diagnóstico Genético” of INCLIVA, the TBC Group of the i2Sysbio and Seqplexing, a company dedicated to the development of technology for genetic analysis, we have created a web application focused on offering services for centralized user authentication and authorization management for gathering information between applications from the same organization.

With this system, we achieve a transparent integration of all services and web applications developed within the same organization, which confers a qualitative advance in the centralization and management of the information of the analytical processes, providing an extra value to the whole software suite that the organization offers.

71

Diego Alonso-López[1], Francisco J Campos-Laborie[1], Miguel Angel Gutierrez[2] and Javier De Las Rivas[1]

[1]*Cancer Research Center (CiC-IBMCC), CSIC and University of Salamanca (CSIC/USAL), Salamanca, Spain*
[2]*Area Facultad de Ciencias, Universidad Católica de Avila, Avila, Spain*

Curated resource of protein-protein interactomes to build custom interaction networks.

Abstract: Collection and integration of all known protein physical interactions within a proteome framework is critical to allow adequate exploration of the protein interaction networks that drive the biological processes in the cell of any given model organism. APID Interactomes (apid.dep.usal.es) is a web server of biological data that provides a comprehensive and curated collection of "protein interactomes" for more than four hundred organisms derived from the exclusive integration of known experimentally validated protein-to-protein physical interactions (PPIs). This year, in September 2018, we completed a comprehensive review of the resource to include several new utilities and quality controls. In this way, we have integrated and unified the PPIs from five primary databases of molecular interactions (BioGRID, DIP, HPRD, IntAct, MINT), from two original systematic resources for human data (BioPlex and HuRI) and also from experimentally resolved 3D structures (PDB) where more than two distinct proteins have been identified. As part of the new quality control strategy, we have included only PPIs that are clearly reported in an experimental research

publication (with an identified PMID), and we have included the details about the specific "experimental interaction detection methods" that are reported for each protein pair (following the compendium of methods that are included in the latest version of HUPO PSI-MI Ontology and Controlled Vocabulary: www.ebi.ac.uk/ols/ontologies/mi). Thanks to this approach we have also classified all experimental interaction detection methods into two main categories that provide very different types of protein networks: (i) "binary" methods and (ii) "co-complex" methods. All these data can be downloaded by the users in a custom way. We are also developing an application ("app") to query and use our resource directly within Cytoscape.

72

Daniel Toro-Domínguez[1], Jordi Martorell-Marugán[1], Raúl López-Domínguez[1], Adrián García-Moreno[1], Víctor González-Rumayor[2], Marta Alarcón-Riquelme[1] and Pedro Carmona-Saez[1]

[1]*Pfizer-University of Granada-Junta de Andalucía Centre for Genomics and Oncological Research (GENYO), Spain*
[2]*Atryshealth, Spain*

ImaGEO: Integrative Gene Expression Meta-Analysis from GEO database

Abstract: The Gene Expression Omnibus (GEO) database provides an invaluable resource of publicly available gene expression data that can be integrated and analyzed to derive new hypothesis and knowledge. In this context, gene expression meta-analysis is increasingly used in several fields to improve study reproducibility and discovering robust biomarkers. Nevertheless, integrating data is not straightforward without bioinformatics expertise. Here, we present ImaGEO, a web tool for gene expression meta-analysis that implements a complete and comprehensive meta-analysis workflow starting from GEO dataset identifiers. The application integrates GEO datasets, applies different meta-analysis techniques and provides functional analysis results in an easy-to-use environment. ImaGEO is a powerful and useful resource that allows researchers to integrate and perform meta-analysis of GEO datasets to lead robust findings for biomarker discovery studies.

Availability:

ImaGEO is accessible at http://bioinfo.genyo.es/imageo

73

Raul Lopez-Dominguez[1], Daniel Toro-Domínguez[1], Jordi Martorell-Marugan[1], Christian Holland[2], Guillermo Barturen[1], Julio Saez-Rodriguez[3], Marta Alarcon-Riquelme[1] and Pedro Carmona-Saez[1]

[1]*Centre for Genomics and Oncological Research (GENYO), Spain*
[2]*Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH Aachen University, Germany*
[3]*Institute of Computational Biomedicine, Heidelberg University*

Analysis of Transcription Factor activity patterns in Systemic Lupus Erythematosus

Abstract: Systemic Lupus Erythematosus (SLE) is a complex and heterogeneous autoimmune diseasewhere the interaction between genetics and environment factors play an important role in its development (Delgado-Vega et al. 2010). Patients with lupus can suffer from periods of disease activity that may remit spontaneously, or in most cases, require treatment to be controlled.

During disease activity tissue damage occurs, and aggressiveness and duration of these activity periods, as well as response to treatment, are highly heterogeneous across different patients.

During the last two decades, many researchers have focused their efforts towards thecharacterization of the genetic determinants that are associated with SLE pathogenesis. These efforts have established more than fifty SLE-associated loci and the finding that type I interferon-inducible gene expression signature is commonly deregulated in SLE patients.

Nevertheless, although it is has been shown that most SLE loci occur in regulatory regions of susceptibility genes (Maurano et al. 2012), there is a lack of systematic analyses that explore the gene regulatory network associated with Lupus pathogenesis. Some previously published works (Harley et al. 2018; Dozmorov, Wren, and Alarcón-Riquelme 2014) have analyzed the enrichment of Transcription Factor (TF) binding sites in regulatory regions of SLE-associated loci. Nevertheless, these studies provide evidences about physical interactions among TFs and genomic regions of SLE-loci, but they do not provide information about the TF activity or the deregulation of gene expression programs.

In this work we have explored activity patterns of all human TFs from TFclass database in SLE samples by analyzing expression levels of their direct target genes (the so-called TF regulon)(Garcia-Alonso et al. 2018). TF activity scores can be estimated based on the levels of its associated target genes. The analysis of the TFs by sample activity matrix has allowed us to establish sets of TFs with differential behaviour among SLE and healthy controls, as well as correlations with disease activity patterns. We have analyzed two large independent cohorts finding a set of TFs that show consistent patterns in both sets. Some of these TFs have beenpreviously described in the context of SLE but there are others that provide new insights into molecular mechanisms that might be important contributors for the understanding of SLE pathogenesis.

74

Kinza Rian[1], Marta R. Hidalgo[2], José Carbonell-Caballero[3], Amal Maurady[4], Mohammed Reda Britel[5], Miguel A. Blázquez[6], Jose Gadea[6] and Joaquin Dopazo[1]

[1]*Fundación Progreso y Salud, Spain*
[2]*Centro de Investigación Príncipe Felipe, CIPF, Spain*
[3]*Center for Genomic Regulation, CRG, Spain*
[4]*Faculty of Science and Technology (FST-T), 90000, Tangier, Morocco*
[5]*Laboratory of innovative technologies (LTI), National School of Applied Sciences in Tangier, Morocco*
[6]*Instituto de Biología Molecular y Celular de Plantas, Universidad Politécnica de Valencia, Spain*

Arapathia: A mechanistic approach based on models of signalling pathways to understand plant physiology

Abstract: One of the main challenges in the analysis of genomic data for agricultural developments is understanding the aspects of the cell functionality that account for plants diseases, stress effect or fertilizers action mechanisms. Here we propose a new approach based on mathematical modelling of signalling pathways successfully used in cancer research [1],adapted to the Arabidopsis thaliana species pathways. Additionally, some statistical tests have been modified in the method in order to overcome the challenges derived from the small number of samples typically available in these plants. In this method, pathways are decomposed into elementary sub-pathways or signal transmission circuits (which ultimately trigger cell functions). Then, a probabilistic model is used to estimate which is the intensity of signal that arrives to the final effector proteins from the initial receptor proteins across each signalling circuit. Individual gene expression values are used to estimate protein activities in the circuits that are used to calculate circuit activities. Differential activation of such circuits is then estimated by comparing circuit activation profiles between biological conditions. Thus, it can be considered that this method provides high-throughput estimations of cell functionalities caused by the interaction of individual gene activities within the pathway. These circuit activity measurements can be used to discover biomarkers that discriminate among the compared conditions. Also, circuit activities can help in the design of gene-targeted interventions for genetic improvement and to increase agricultural productivity.

References:

1. Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J: High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget 2017, 8(3):5160-5178

75

Ruben Sanchez-Garcia[1], Joan Segura[1], David Maluenda[1], Jose M. Carazo[1] and Carlos Oscar Sánchez Sorzano[1]

[1]*CNB/CSIC, Spain*

Deep consensus: deep learning for particle pruning in cryo-electron microscopy

Abstract: The advent of direct electron detectors and the improvements in image processing algorithms have revolutionized the field of cryo electron microscopy (cryo-EM). As a consequence, the 3D structure of many protein complexes can now be determined using cyo-EM as a routine technique. In order to obtain high resolution data, single particle (SP) cryo-EM workflows require to collect tens of thousands of projections of the complexes. These projections, generally known as particles, are selected from the micrographs using many different automatic or semi-automatic particle picking algorithms. However, due to several factors such as the low signal-to-noise ratio of the micrographs or the presence of different types of contaminants, the fraction of false positive and false negative particles picked by all those methods is not negligible, ranging from 10% to more than 25% in many cases. Thus, cleaning steps known also as pruning, are required in order to decrease impurity levels. Although typical pruning approaches, that combine 2D classification algorithms or particle sorting with human intervention, have proven successful in many situations, they are also time consuming and difficult to reproduce.

76

Janet Piñero[1], Josep Saüch[1], Emilio Centeno[1], Javier García-García[1] and Laura I. Furlong[1]

[1]*GRIB (IMIM-UPF)*

The DisGeNET Cytoscape App: enabling network biology for human diseases

Abstract: Network based approaches have proven to be essential to understand the molecular mechanisms underlying human diseases. The use of these methods has been boosted by the abundance of information about the genetic determinants of human diseases and of high quality human interactomics data. Here, we present the DisGeNET Cytoscape App, designed to facilitate the exploration of the genetic basis of human diseases in the context of different types of biological networks.The DisGeNET Cytoscape App offers a variety of functions to query, analyze, and visualize gene-disease and variant-disease associations networks obtained from DisGeNET. It allows queries the data for specific diseases, genes, and variants, and their combinations, or to perform queries restricted by database source or disease class. Moreover, the metrics offered by DisGeNET, such as the DisGeNET score, the DisGeNET association type, and the Evidence Index can be used to filter the networks. Moreover, the DisGeNET App enables annotating foreign networks generated by other applications with DisGeNET diseases. Finally, we have implemented an API to access the data using different scripting languages via the REST protocol. In this way, different types of networks can be automatically created by external workflows in R, and Python. The DisGeNET App is freely available for download at the Cytoscape App store (https://apps.cytoscape.org/apps/disgenet).

FUNDING: We received support from ISCIII-FEDER (CP10/00524, CPII16/00026), the EU H2020 Programme 2014-2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and is

supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. The DCEXS is a "Unidad de Excelencia María de Maeztu", funded by the MINECO (ref: MDM-2014-0370).

77

Irene Pérez Díez[1], Marta R Hidalgo[1], Miguel Barquín[2], Atocha Romero[2], Rosa Farràs[3] and Francisco García-García[1]

[1]*Bioinformatics and Biostatistics Unit, Principe Felipe Research Center, CIPF, Spain*
[2]*Medical Oncology Department, Hospital Universitario Puerta de Hierro-Majadahonda, Majadahonda, Spain*
[3]*Department of Oncogenic Signalling, Principe Felipe Research Center, CIPF, Spain*

A Gender Point Of View in Non Small Cell Lung Cancer: Systematic Review and Meta-Analysis of Omics Studies

Abstract: Lung cancer is the most common cause of cancer death worldwide, and 85% of patients belongs to a subtype known as non small cell lung cancer. Although it is greatly associated with smoking, lung cancer in never smokers is more common in women than men. Understanding its biology and molecular mechanisms is crucial for the development of effective therapies and the improvement of its diagnosis.

Our approach starts with a systematic review and selection of omics studies available in public repositories such as Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA), focusing on microarray and RNA-seq datasets. Then, each study is analyzed in a signaling pathways and molecular functions context to highlight their alterations. Finally, the application of a functional meta-analysis to all results provides a better interpretation in a Systems Biology approach, making it possible to combine the expression meta-analysis with epigenomics data.

The results obtained will allow us to know the altered functions in non small cell lung cancer and its gender differences, providing us with the opportunity to determine new and more effective therapeutic targets within the framework of Precision Medicine.

78

Raúl Pérez Moraga[1,2], María José Escartí Fabra[3], Julio Sanjuán[3], Francisco Garcia-Garcia[1] and María de la Iglesia-Vayá[2,3]

[1]*Bioinformatics and Biostatistics Unit, CIPF, Spain*
[2]*Joint Unit FISABIO & CIPF, Spain*
[3]*Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), ISCIII, Spain*

Study of alterations of brain functional connectivity in schizophrenia from image analysis by functional magnetic resonance imaging: a gender perspective

Abstract: Schizophrenia is a complex disease, that affects cognitive syndrome that seem to originate from disruption of brain connectivity caused by genetic or environmental factors, or both, also, one key part of schizophrenia are the auditory verbal hallucinations (AVH). This extreme phenotypic symptom is the one most frequent positive signs in patients with schizophrenia, therefore studying the functional connectivity of this symptom is key point to better understand a syndrome as complex as schizophrenia.

In this project we approach the problem from the perspective of bioinformatics, neuroimaging and data analysis. Our project are composed by 65 patients divided into 3 groups, one control group of healthy patients (n = 22), one group of schizophrenic patients without AVH (n = 26) and the last group of schizophrenic patients with AVH (n = 20). The cerebral activity has registered with a 3 Tesla magnetic resonance Imaging (MRI) scanner. During the scanning the patients have

listened records of words with neutral and emotional charge sessions with the objective to register the cerebral activity under these two different circumstances.

The results obtained in the scanner are treated with different bioinformatic tools, where the images obtained in the scanner will be treated by different scripts in Python and in R to perform an analysis of the functional connectivity between the different groups, following a data driving methodology and also taking into account the gender of the patients, since schizophrenia is significantly more prevalent in men than in women.

79

Miguel Rodriguez Galindo[1], Sonia Casillas Viladerrams[2] and Antonio Barbadilla Prados[2]

[1]*Center for Genomic Regulation, CRG, Spain*
[2]*Institute of Biotechnology and Biomedicine, IBB, Spain*

Analysis of germline de novo mutation rates on exons and introns

Abstract: While recent studies have found that the mutation rates greatly vary across the human genomes, this rate is generally assumed to be equal across close genic regions. If this condition –part of the evolutionary biology theoretical core– is not fulfilled on natural populations, some of the effort regarding the study of natural selection will be compromised.

A recent paper claims that there are fewer mutations in exons than in introns at somatic cells due to an enhanced exonic mismatch repair system activity, arising from a specific epigenomic context, and provides evidences that this phenomena should undergo also on germline cells, with a clear impact on evolutionary biology. In this work we conduct several bioinformatic analysis to shed light on the extrapolation of the somatic phenomena into germinal cells.

By means of a bibliographic review we found that the distribution, spectra and determinants of mutation rates is quite different between human germline and soma. Moreover, we have found that the epigenomic context seems to greatly differ between somatic and germline cells. Based on all this information, we do not expect a different mutation rate on exons and introns at germ cells given the privileged status of the germline genome with respect to the somatic one.

In the line of our hypothesis, we empirically find nearly the same number of germinal mutations in exons than the expected from their sequence content. Therefore, we find no evidence of an enhanced mismatch repair system activity in exons with respect to adjacent introns at the germline, in contrast to what has been previously described in somatic cells.

Our findings have important implications for the understanding of the dichotomous nature between germline and somatic genomes with respect to mutational and DNA repair processes. Moreover, they advance knowledge of the fundamentals of evolutionary biology by supporting different, but not mutually excluding, theoretical frameworks.

80

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

[1]*Research Group on Integrative Biomedical Informatics, Institut Hospital del Mar d'Investigacions Mèdiques, UPF, Spain*

ResMarkerDB: a database of biomarkers of response to antibody therapy in breast and colorectal cancer

Abstract: The development and clinical efficacy of therapeutic monoclonal antibodies for breast and colorectal cancer have contributed greatly to the improvement of patients' outcomes by individualizing treatments according to their genetic background [1]. Responding patients, however, may become resistant to treatment in advanced stages. In other cases, patients may be resistant to treatment even though they are molecularly characterized to be responsive [2,3]. Although several databases characterize biomarkers of drug response, there is a need of resources that offer this information to the user in a harmonized manner.

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

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

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

REFERENCES

[1] Chiavenna,S.M. et al. (2017) State of the art in anti-cancer mAbs. J. Biomed. Sci., 24, 15.

[2] Pruneri,G. et al. (2016) Biomarkers for the identification of recurrence in human epidermal growth factor receptor 2-positive breast cancer patients. Curr. Opin. Oncol., 28, 476–483.

[3] Bronte,G. et al. (2015) New findings on primary and acquired resistance to anti-EGFR therapy in metastatic colorectal cancer: do all roads lead to RAS? Oncotarget, 6.

81

Joan Segura[1], Ruben Sanchez-Garcia[1], Carlos Oscar Sánchez Sorzano[1] and Jose Maria Carazo[1]

[1]*National Center for Biotechnology, CNB-CSIC, Spain*

3DBIONOTES: crossing molecular biology

Abstract: Next generation sequencing methods are producing a vast amount of proteomic and genomic information. As consequence, most relevant biological databases as UNIPROT (2015) and ENSEMBL (Cunningham, et al., 2015) are flooded with an exponential growing amount of data extending the number of annotations for genes and proteins and in particular variants associated to diseases.

Genomic and proteomic annotations are a valuable contribution in the study of protein and gene functions. However, structural information is an essential key for a deeper understanding of the molecular properties that allow proteins and genes to perform specific tasks. Therefore, depicting genomic and proteomic information over structural data would offer a very complete picture in order to understand how proteins and genes behave in the different cellular processes.

In this work we present a web platform, 3DBIONOTES (Segura, et al., 2017), that aims to merge the different levels of molecular biology information, including genomics, proteomics and interactomics data into a unique graphical environment. 3DBIONOTES integrates proteomic, genomic and functional annotations with structural data, providing a unified and interactive view of the different sources of information. Current development offers a unified view of four of the most relevant biological databases: ENSEMBL (Cunningham, et al., 2015), UniProt (2015), PDB (Berman, et al., 2014) and EMDB (Lawson, et al., 2011) that comprises the core of the framework.

3DBIONOTES core is enriched with biomedical and biochemical annotations from different sources. This information can be interactively displayed at any molecular level: gene sequence, protein sequence, protein structure and protein-protein interaction level. Finally, a statistical analysis in order to find cooccurrence between genomic variants associated to particular diseases and other structural and biochemical features is computed by a means of a Fisher test. In that way, 3DBIONOTES finds relations between diseases and biochemical or structural features of proteins.

3DBIONOTES web application is accessible at http://3dbionotes.cnb.csic.es

---

82

Juan Luis Trincado Alonso[1], Marina Reixachs[1], Eduardo Eyras[1] and Jun Yokota[2]

[1]*Universitat Pompeu Fabra, UPF, Spain*
[2]*IMPPC, Japan*

The immunogenic impacts of splicing alterations in small cell lung cancer

Abstract: We describe a novel approach for the exhaustive identification of neo-epitopes from tumor-specific splicing alterations, including aberrant spliced junctions, retained introns and exonizations. Using mass spectrometry for MHC-I associated proteins, we show that splicing derived neo-epitopes are processed and presented by MHC-I complexes. We applied this method to a cohort of 123 small cell lung cancer patients (SCLC) and found that tumor-specific splicing alterations more frequently eliminate than create epitopes, hence uncovering a new mechanism of immune escape in SCLC.

83

Francesco Ronzano[1], Alba Gutiérrez-Sacristán[2] and Laura I. Furlong[1]

[1]*GRIB (IMIM-UPF), Spain*
[2]*Department of Biomedical Informatics, Harvard Medical School, United States*

Comorbidity4j: a Web tool for interactive comorbidity analysis

Abstract: Comorbidity analyses are aimed at the identification of relevant patterns of co-occurring diseases, given a population of individuals [1]. Pushed by the growing availability of Electronic Health Records for data mining, nowadays the outcome of comorbidity analyses has a substantial impact on the estimation of life expectancy, quality of life and healthcare costs [2-4]. In this context, the availability of analytical tools to easily and interactively explore disease comorbidities over large datasets of patients is fundamental. Nevertheless, few comorbidity analysis frameworks are currently available, most of them tailored to specific sets to disease. Among them there are the R packages comoRbidity [5], MedicalRisk [6], Coxnet [7] and Icd R, and the SAS package Elixhauser [8].

We present Comorbidity4j, a Java tool that supports the systematic analysis of comorbidities generating interactive Web-based visualizations of results. Patient demographic data, stratification facets (e.g. gender, race, etc.) and disease history are provided as input to Comorbidity4j by a set of CSV files, leaving users total freedom with respect to the choice of disease identifiers and the name and order of CSV columns. Filters can be defined to limit comorbidity analyses to subgroups of patients or diseases, to select the most relevant pairs of comorbid diseases or to consider temporal diseases directionality. Comorbidity4j computes several disease association statistics including Relative Risk, Odds Ratio, Comorbidity Score [9] and Pearson correlation coefficient, performing an exact Fisher Test to verify the statistical significance of disease co-occurrences.

Comorbidity4j generates interactive Web-based visualizations of comorbidity analyses, including: (i) a summary of the input parameters and the processing log; (ii) a rich collection of charts to provide an overview of the input patient dataset (patient distribution by age, gender, etc.); (iii) an interactive interface where disease pairs can be filtered by combining comorbidity scores and disease names and then visualized by heatmaps and disease networks.

Comorbidity4 supports multi-threading to efficiently process big patient datasets. The tool can be executed on users' private workstations to guarantee data privacy. We also provide comorbidity4j as an online service. A detailed documentation of comorbidity4j is available at: http://comorbidity4j.readthedocs.io/.

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

REFERENCES

[1] Valderas, Jose M., et al. (2009) Defining comorbidity: implications for understanding health and health services. The Annals of Family Medicine 7.4, 357-363.

[2] Gijsen, Ronald, et al. (2001) Causes and consequences of comorbidity: a review. Journal of clinical epidemiology 54.7, 661-674.

[3] Backenroth, Daniel, et al. (2016) Using rich data on comorbidities in case-control study design with electronic health record data improves control of confounding in the detection of adverse drug reactions. PloS one 11.10, e0164304.

[4] Marengoni, Alessandra, et al. (2011) Aging with multimorbidity: a systematic review of the literature. Ageing research reviews10.4, 430-439.

[5] Gutiérrez-Sacristán, A. et al. (2018). comoRbidity: an R package for the systematic analysis of disease comorbidities. Bioinformatics, 34(18), 3228-3230.

[6] McCormick, Patrick, and Thomas Joseph. (2015) Medicalrisk: Medical Risk and Comorbidity Tools for ICD-9-CM Data.

[7] SUN, Hokeun, et al. (2014) Network-regularized high-dimensional Cox regression for analysis of genomic data. Statistica Sinica,vol. 24, no 3, p. 1433.

[8] Elixhauser, Anne, et al. (1998) Comorbidity measures for use with administrative data." Medical care, 36.1: 8-27.

[9] Roque, Francisco S., et al. (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS computational biology 7.8, e1002141.

84

Janet Piñero[1], Abel Gonzalez-Perez[2], Emre Guney[1], Joaquim Aguirre-Plans[1], Ferran Sanz[1], Baldo Oliva[1] and Laura I. Furlong[1]

[1]*GRIB (IMIM-UPF), Spain*
[2]*Institute for Research in Biomedicine, IRB Barcelona, The Barcelona Institute of Science and Technology, BIST, Barcelona, Spain*

Network, Transcriptomic and Genomic Features Differentiate Genes Relevant for Drug Response

Abstract: Understanding the mechanisms underlying drug therapeutic action and toxicity is crucial for the prevention and management of drug adverse reactions, and paves the way for a more efficient and rational drug design. The characterization of drug targets, drug metabolism proteins, and proteins associated to side effects according to their expression patterns, their tolerance to genomic variation and their role in cellular networks, is a necessary step in this direction. In this contribution, we hypothesize that different classes of proteins involved in the therapeutic effect of drugs and in their adverse effects have distinctive transcriptomics, genomics and network features. We explored the properties of these proteins within global and organ-specific interactomes, using multi-scale network features, evaluated their gene expression profiles in different organs and tissues, and assessed their tolerance to loss-of-function variants leveraging data from 60K subjects. We found that drug targets that mediate side effects are more central in cellular networks, more intolerant to loss-of-function variation, and show a wider breadth of tissue expression than targets not mediating side effects. In contrast, drug metabolizing enzymes and transporters are less central in the interactome, more tolerant to deleterious variants, and are more constrained in their tissue expression pattern. Our findings highlight distinctive features of proteins related to drug action, which could be applied to prioritize drugs with fewer probabilities of causing side effects.

FUNDING: We received support from ISCIII-FEDER (CP10/00524, and CPII16/00026), IMI-JU under grant agreements no. 116030 (TransQST) and no. 777365 (eTRANSAFE) resources of which are composed of financial contribution from the EU-FP7 (FP7/2007-2013) and EFPIA companies in kind contribution, and the EU H2020 Program 2014–2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and was supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. The DCEXS is a "Unidad de Excelencia María de Maeztu", funded by the MINECO (ref: MDM-2014-0370).

85

Cinta Pegueroles[1], Susana Iraola-Guzmán[1], Uciel P. Chorostecki[1], Ewa Ksiezopolska[1], Ester Saus[1] and Toni Gabaldón[1]

[1]*CRG-Centre for Genomic Regulation, Spain*

Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus Caenorhabditis

Abstract: Long non-coding RNAs (lncRNAs) are a heterogeneous class of genes that are non-coding and longer than 200 nt. Since lncRNAs (or at least a fraction of them) are expected to be functional, many efforts have been invested to screen for the presence of lncRNAs in several species. However, our knowledge of lncRNAs in non vertebrate species is very limited, with insects being the only exception. To gain insights into lncRNAs composition and evolution in nematodes, we catalogued lncRNAs using RNAseq data from the same larval stage in four species of Caenorhabditis and we used a combination of parameters to identify homology between lncRNAs: syntenic relationships, sequence conservation and structural conservation. We classified a total of 1,532 out of 7,635 genes into groups of co-expressed syntenic genes, suggesting that a large fraction of the predicted lncRNAs may be species specific. Genes within families were more likely to produce blastn hits and to share motifs as compared to random pairs of unrelated genes, suggesting that some stretches of sequence conservation are at least preserved within groups. Interestingly, lncRNAs regions covered with motifs are more likely to be unpaired, indicating that these regions may be bound to other molecules. In contrast, we did not detect a significant enrichment for local secondary structure conservation in genes within families. Finally, we show that the reduction in the genome size of the selfertile nematodes also affects the lncRNA class, since in these species the number of annotated lncRNAs was lower than in out-crossing species. We provide the first catalog of lncRNAs in nematodes, which may be a potential source for novel functions, and genes sharing homology between species, which may be important for their development.

86

Marina Marcet-Houben[1] and Toni Gabaldón[1]

[1]*CRG-Centre for Genomic Regulation, Spain*

Evolutionary and functional patterns of conserved gene neighborhood in fungi

Abstract: Gene clusters comprise genomically co-localized and potentially co-regulated genes that tend to be conserved across species. In eukaryotes, multiple examples of metabolic gene clusters are known, particularly among fungi and plants. However, little is known about how gene clustering varies among taxa, or among genes involved in different functional classes. Furthermore, mechanisms involved in the formation, maintenance and evolution of such clusters remain a puzzle.

We have designed Evolclust, a python-based tool for the inference of evolutionary conserved gene clusters from genome comparisons, independently of the function or gene composition of the cluster. Evolclust predicts conserved gene clusters from pairwise genome comparisons and infers families of related clusters from multiple (all vs all) genome comparisons.

Using Evolclust, we surveyed 341 completely-sequenced fungal genomes to discover evolutionary conserved families of gene clusters. We then used a phylogenomics approach to dissect the evolution of the 12,124 cluster families. Our results show that most clusters have a single origin, and evolve by vertical evolution coupled to differential loss. However, convergent evolution -i.e. independent appearance of the same cluster- plays a more important role than anticipated. Horizontal gene transfer of gene clusters seems to be rather restricted, and particularly affects gene clusters involved in secondary metabolism. Altogether our results provide insights on the evolution of gene clustering as well as a broad catalog of evolutionary conserved gene clusters whose function remains to be elucidated.

87

Eva Vargas[1], Signe Altmäe[2] and Francisco J Esteban[1]

[1]*Department of Experimental Biology, University of Jaen, Spain*
[2]*Department of Biochemistry and Molecular Biology, University of Granada, Spain*

New insights into the molecular basis of endometriosis through the study of its comorbid disorders: a Systems Biology approach

Abstract: Systems Biology approaches have shown to be useful to address the study of multifactorial diseases [1]. Endometriosis is a complex gynecological condition characterized by the presence of endometrial-like tissue outside the uterus [2]. Although it is the most studied gynecological pathology, its pathogenesis remains still obscure. Women with endometriosis are at high risk of several chronic diseases, such as autoimmune diseases and cancer, among others [3]. Our aim was to identify genes and pathways in common with endometriosis and different co-occurring diseases in order to find potential biomarkers and targeting pathways of endometriosis. First, we performed a systematic review in order to identify all possible endometriosis-associated comorbidities. Next, we coded this list into medical terms whose gene expression profiles were downloaded from Phenopedia database and subsequently analyzed following a Systems Biology approach. Our results highlight a group of 131 candidate genes which were recurrently expressed among endometriosis and its comorbidities. We were able to validate on independent sample sets the expression of 18 genes: AGTR1, BDNF, C3, CCL2, CD40, CDH1, CYP17A1, ESR1, IGF1, IGF2, IL10, IRS2, MMP1, MMP7, MMP9, PGR, SERPINE1, and TIMP2, which were mainly involved in immune and drug response, steroid hormones metabolism and cell proliferation. This results shed some light on the molecular processes underlying the etiopathogenesis of endometriosis and its comorbid conditions.

References:

[1] Díaz-Beltrán L, Esteban FJ, Varma M, Ortuzk A, David M, Wall DP. Cross-disorder comparative analysis of comorbid conditions reveals novel autism candidate genes. BMC Genomics. 2017; 18(1): 315.

[2] Baranov VS, Ivaschenko TE, Liehr T, Yarmolinskaya MI. Systems genetics view of endometriosis: a common complex disorder. Eur J Obstet Gynecol Reprod Biol. 2015; 185: 59-65.

[3] Yuan M, Li D, Zhang Z, Sun H, An M, Wang G. Endometriosis induces gut microbiota alterations in mice. Hum Reprod. 2018; 33(4): 607-16.

88

Beatriz García-Jiménez[1], Tomas De La Rosa[2] and Mark D. Wilkinson[1]

[1]*Center for Plant Biotechnology and Genomics UPM - INIA, Universidad Politecnica de Madrid, Madrid, Spain*
[2]*Universidad Carlos III de Madrid, Spain*

MDPbiome: microbiome engineering through prescriptive perturbations

Abstract: Motivation:

Recent microbiome dynamics studies highlight the current inability to predict the effects of external perturbations on complex microbial populations. To do so would be particularly advantageous in fields such as medicine, bioremediation or industrial scenarios.

Results:

MDPbiome statistically models longitudinal metagenomics samples undergoing perturbations as a Markov Decision Process (MDP). Given a starting microbial composition, our MDPbiome system suggests the sequence of external perturbation(s) that will engineer that microbiome to a goal state, for example, a healthier or more performant composition. It also estimates intermediate microbiome states along the path, thus making it possible to avoid particularly undesirable/unhealthy states. We demonstrate MDPbiome performance over three real and distinct datasets, proving its flexibility, and the reliability and universality of its output 'optimal perturbation policy'. For example, an MDP created using a vaginal microbiome time series, with a goal of recovering from bacterial vaginosis,

suggested avoidance of perturbations such as lubricants or sex toys; while another MDP provided a quantitative explanation for why salmonella vaccine accelerates gut microbiome maturation in chicks. This novel analytical approach has clear applications in medicine, where it could suggest low-impact clinical interventions that will lead to achievement or maintenance of a healthy microbial population, or alternately, the sequence of interventions necessary to avoid strongly negative microbiome states.

Availability:

Code (https://github.com/beatrizgj/MDPbiome) and result files (https://tomdelarosa.shinyapps.io/MDPbiome/) are available online.

89

Jon Sánchez-Valle[1], Vera Pancaldi[2] and Alfonso Valencia[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Cancer Research Center of Toulouse, France*

Unveiling the molecular basis of disease co-occurrence: towards personalized comorbidity profiles

Abstract: Comorbidity is a medical problem that is attracting increasing attention in healthcare and biomedical research. However, little is known about the molecular processes leading to the development of a specific disease in patients affected by other conditions. Moreover, despite the general tendencies, not all the patients with a disease are at the same risk of developing secondary diseases. In this study, we present the Diseases' Molecular Comorbidity Network inferred from similarities in patients' molecular profiles, which significantly recapitulates epidemiologically documented comorbidities, providing the basis for their interpretation at a molecular level. Furthermore, expanding on the analysis of subgroups of patients with similar molecular profiles, our approach extracts patient subgroups with specific comorbidity risks (not previously described at the disease level), implicates distinct genes in such relations, and identifies drugs whose side effects are potentially associated to the observed comorbidities. The results are accessible through the Disease PERCEPTION portal (http://disease-perception.bsc.es/), which allows exploration of the Diseases' Molecular Comorbidity Network and the Stratified Comorbidity Network.

90

Salvador González Gordo[1]

[1]*Estación Experimental del Zaidín, CSIC, Spain*

RNA-Seq analysis unravels the modulation of gene expression by nitric oxide (NO) during ripening of sweet pepper (Capsicum annuum L.) fruits

Abstract: Nitric oxide (NO) is a signaling molecule which regulates a wide range of plantphysiological processes. Nevertheless, the role of NO in the regulation of fruit ripening is still unclear. In previous studies, it was developed a non-invasive method to apply exogenous NO gas to pepper fruits at breaking point stage [1]. In this work, (Capsicum annuum L.) fruits at four different ripening stages were selected: immature (green), breaking point (with and without NO treatment, 5 ppm, 1 h) and ripe (red).

Using these samples, RNA-sequencing was performed using an Illumina platform generating 223,079,396 reads. Raw reads were processed and assembled using an accuratepredesigned workflow [2]. Thus, de novo sweet pepper transcriptome assemblyprovided a total of 63,359 tentative transcripts (TTs). TTs were annotated againstArabidopsis

thaliana, Solanum lycopersicum and Solanum tuberosum resulting in29,716, 28,333 and 30,169 unique orthologues, respectively. NO-treatment causeddifferential expression of 2,436 genes (1,912 up-regulated and 524 down-regulated).

Moreover, a time-course analysis of gene expression was carried out using maSigPro Rpackage providing 8,805 genes that could be grouped according to their expressionpattern. Finally, gene datasets were used for functional enrichment analyses usingdifferent web tools (AgriGO, PlantRegMap and KOBAS 3.0). Our design proved thattranscriptomics and further bioinformatics analysis could be useful tools to unravel the involvement of NO in the ripening process in agricultural crops.

91

Adria Closa[1], Antonio Agraz-Doblas[2,3], Ignacio Varela[3], Pablo Menendez[2,4,5] and Eduardo Eyras[1,5]

[1]*Universitat Pompeu Fabra, UPF, Spain*
[2]*Josep Carreras Leukemia Research Institute, Spain*
[3]*Instituto de Biomedicina y Biotecnología de Cantabria, Spain*
[4]*CIBERONC, Spain*
[5]*Catalan Institution for Research and Advanced Studies, ICREA, Spain*

Identification of RNA-processing prognostic and therapeutic markers in MLL-rearranged infant acute leukemia

Abstract: Infant acute lymphoblastic leukemia (ALL) has a poor prognosis, especially with MLL gene rearrangements (MLL-r), which occur in ~80% of the patients during embryonic/fetal hematopoiesis. Genome-sequencing studies of MLL-r ALL patients have shown a very low frequency of somatic mutations, indicating that MLL may not require additional alterations to induce full transformation. However, ALL cannot be recapitulated in a mouse model by only integrating the fusion, suggesting that additional alterations are necessary for leukemogenesis.

MLL fusions have the potential to impact the RNA processing of genes at genome scale through changes in transcriptional elongation, thereby providing a new layer of molecular variation that has remained undetected so far, and which could lead to new prognostic markers and therapeutic strategies. We present here an exhaustive analysis of the RNA-processing alterations in infant ALL samples in relation to the MLL fusions. We have analyzed RNA-sequencing on a cohort of 32 MLL-r and 10 non-MLL infant ALL cases, plus 5 normal B-cell progenitor samples.

In a preliminary analysis we detect a clear differential expression pattern between the different types of MLL-r ALL (MLL-AF4 and MLL-AF9) and non-MLL samples indicating an impact of MLL fusions on gene expression. However, we found a low level of overlap in the differential expressed genes between the fusion groups and non-MLL samples, pointing towards fusion-specific molecular alterations. In particular, we found a pattern of expression alteration in several splicing factors. In agreement with this finding, we found common and specific events that are differentially spliced between MLL-r samples and controls, with a high proportion of skipping exons and alternative first exons specific for both types of MLL fusion. We further present an analysis of the potential functional impacts and the phenotypic convergence of these alterations across patients. This is the first study of RNA-processing alterations in association to MLL-AF4 and MLL-AF9 fusions in ALL and of their role in leukemogenesis.

92

Davide Cirillo[1], Andres Cañada[2], Javier Omar Corvi[1], José M. Fernández[1], Jose Antonio Lopez-Martin[3], Salvador Capella-Gutiérrez[1], Martin Krallinger[1] and Alfonso Valencia[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*

[2]*Spanish National Cancer Research Center, CNIO, Spain*
[3]*Research Institute Hospital 12 de Octubre, i+12, Spain*

Training IBM Watson with MelanomaMine

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

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

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

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

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

93

Sheila Zuniga[1], Noelia Tarazona[1], Francisco Gimeno-Valiente[1], Pilar Rentero-Garrido[1], Valentina Gambardela[1], Marisol Huerta[1], Desamparados Roda[1], Susana Rosello[1], Josefa Castillo[1] and Andres Cervantes[1]

[1]*Instituto de Investigacion Sanitaria INCLIVA, Spain*

Mutational screening of colorectal cancer patients in phase I clinical trials using unique molecular identifiers in gene panels to improve variant calling

Abstract: One of the key factors that slows down drug development is the genetic difference between individuals. With the arrival of new sequencing techniques, mutational screening in a clinical context is now feasible. In contrast to classical approaches in which the same treatment is provided for all patients with the same tumor type, current clinical trial protocols have started to include a mutational screening before a patient is enrolled in the first phases of the drug trial in order to identify population subgroups that clearly can benefit from that particular treatment.

Solid tumors are highly heterogeneous both tissular and genetically. The presence of multiple subclones within the same tumor mass, some of them in a very low frequency that usually are responsible for relapse in later disease stages, as well as the inclusion of normal tissue within the pathological area complicate the identification of somatic variants. Moreover, it is well-known that PCR artifacts introduce noise that imply an increase in the number of final identified variants making result interpretation harder.

Recently, a new technical approach that incorporates Unique Molecular Identifiers (UMI) to tag each DNA molecule for Next-Generation Sequencing (NGS) library construction in targeted DNASeq assays, promises to reduce sequencing background to improve variant detection.

Here we present our pipeline and conclusions about the key points in the analysis and use of UMI-based techniques compared to more traditional approaches for NGS library generation based on the mutational screening of colorectal cancer patients who are potential candidates to receive experimental therapy in phase I clinical trials.

94

Uciel Chorostecki[1] and Toni Gabaldón[1]

[1]*CRG, Spain*

Adopting structural flexibility to fill the gap between structure and function in lncRNAs.

Abstract: Long non-coding RNAs (lncRNAs) are abundant in mammalian transcriptomes. However, it remains unclear how many of them are functional, and how their functions are performed. LncRNAs seem to be poorly conserved at the sequence level, but some of them share conserved structural elements and are present at syntenic genomic positions in different species. A recent study revealed that secondary structure constrains sequence variation in lncRNAs, so that polymorphisms are depleted in low accessibility regions and tend to be neutral with respect to structural stability. This is in contrast with previous analyses that dismissed relationships between structure and sequence evolution in lncRNAs. A crucial difference in the former study is that the considered structural feature, accessibility, is computed from an ensemble of thermodynamically stable structures.

It has been recently developed in our group a novel Illumina-based implementation of in-vitro parallel probing of RNA structures called nextPARS. This approach achieves comparable accuracy to previous implementations, while enabling higher throughput and sample multiplexing. Using this approach we observed that many lncRNA sites exhibit positive signals for both single- and double-strand specific enzymes, suggesting several structures may coexist. Based on this, we argue that the difficulty of identifying links between sequence and structure in lncRNAs results in part from limitations imposed by assuming a single, stable structure. Thus, we considered ensembles of co-existing structures in lncRNAs, and we developed a computational framework that enables this.

Using this approach, we studied lncRNAs from animals and fungi using experimental data from RNA structure probing. This novel multidisciplinary approach establishes a framework for understanding the evolution of lncRNAs and should help to fill the gap between structure and function of lncRNAs in different species.

95

Juan Rodriguez-Rivas[1], Simone Marsili[2], David Juan[3] and Alfonso Valencia[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Independent scholar, Spain*

[3]*Institute of Evolutionary Biology - Pompeu Fabra University, Spain*

Reliable coevolution-based contact prediction between proteins with limited sequence data

Abstract: Protein interaction surfaces are subject to strong evolutionary constraints to maintain functional complexes. The evolutionary interdependence, or coevolution, of physically interacting residues between proteins can be detected through covariations in multiple sequence alignments (MSA). However, sequences in MSAs are not independent since they are evolutionarily related, generating covariations not associate with physical contacts. Current coevolution-based contact prediction methodologies are not explicitly designed to deal with this source of noise that hinders the identification of structural constraints. This issue is even more severe in the common scenario of limited sequence data. We conduct a large computational experiment in a dataset of 490 inter-protein domain-domain interactions to study the effect of the phylogenetic signal and potential corrections. We find that maximums from empirical null distributions provide an estimation of the phylogenetic signal in a case-specific manner. Based on this observation, we devise a method that implements case-specific phylogenetic corrections. Our approach improves the overall quality of contact predictions, especially for MSAs with fewer sequences. Additionally, we show that the distribution of maximums from empirical null distributions can be modelled by a Generalized Extreme Value (GEV) distribution that, in turn, can be used to estimate the statistical significance of contact predictions. We discuss representative examples of the strengths and limitations of our approach. In summary, we propose and validate a new methodology that provides reliable contact predictions for a wider range of protein-protein interactions.

96

Matteo Schiavinato[1], Juliane C. Dohm[1], Toni Gabaldón[2] and Heinz Himmelbauer[1]

[1]*University of Natural Resources and Life Sciences, BOKU, Austria*
[2]*Centre for Genomic Regulation, CRG, Spain*

A phylogenomic strategy for subgenome separation

Abstract: Nicotiana benthamiana is an Australian tobacco plant, and is a popular host for recombinant protein production. Despite its growing usage, the origin of its allo-tetraploid genome is still debated. As for many other plants, its genome is the result of a hybridization process. One of the two candidate hybridization parents is thought to be an ancestor of Nicotiana section Sylvestres, while the other is less certain, even though a preference for Nicotiana section Noctiflorae is found in literature. Identifying the second parent would improve the genomic and metabolic understanding of this plant, as well as the knowledge on the dynamics that shaped the Nicotiana genus. Here we showcase a workflow to address this matter using a combination of sequencing data and phylogenomic tools. We perform a gene prediction on the N. benthamiana Nb-1 draft genome assembly, identifying 50,516 genes and their coding regions (CDS). We then map mRNA-seq from six parental Nicotiana candidates against these CDS and quantify covered positions for each species. We form pairs of candidate parents and we join their mapping results to elect the best performing parental pair. We then generate a phylome for N. benthamiana using the PhylomeDB pipeline and use all its gene trees to count occurrences as closest sister taxon for each of the tested Nicotiana species. Our results allow us to speculate that the second parental genome of N. benthamiana might be an ancestor of section Noctiflorae.

97

Adolfo López Cerdán[1,2], Francisco García-García[1] and María de la Iglesia-Vayá[2]

[1]*Bioinformatics and Biostatistics Unit, CIPF, Spain*

[2]*Joint Unit FISABIO & Prince Felipe Research Center, CIPF, Spain*

Radiomics: new approaches in Biomedicine combining biomedical image and omic data

Abstract: Medical Imaging arises as one of the most effective innovations in cancer diagnosis and treatment. Its noninvassiveness allows the visualization of the radiographic phenotype of a tumor across its different stages.

Radiomics is an emerging field that can turn this phenotype into a huge set of quantitative phenotypic features using a large number of automatically applied data-characterization algorithms. The potential of this field has widely shown across multiple tumour types.

In this study, the main goal will be the integration of Radiomics phenotypic features with clinical and omics (genomics, transcriptomics,…) data in order to build descriptive and predictive models of some types of tumors. To achieve this, we will make use of the open-source platform PyRadiomics, implemented in Python, and several multidimensional approaches from R software.

The models obtained will provide us valuable information about diagnosis, prognosis and prediction of cancer within the framework of Precision Medicine.

98

Martín Garrido-Rodríguez[1,2,3], Rosario Morrugares[2,3], Eduardo Muñoz[2,3], Joaquín Dopazo[4] and Marco A Calzado[2,3]

[1]*InnohealthGroup, Spain*
[2]*Instituto Maimónides de Investigación Biomédica de Córdoba, IMIBIC, Spain*
[3]*Universidad de Córdoba, UCO, Spain*
[4]*Clinical Bioinformatics Area, Fundación Progreso y Salud, FPS, Spain*

Integrative analysis of Transcriptomic, Proteomic and Phosphoproteomic data in DNA damage signaling pathways

Abstract: Until recently, the analyses carried out to understand the cellular behavior underlying a response to different stimuli or situations were performed using transcriptomic or proteomic data in most cases, independently. However, in the recent years and mainly due to the instrumental breakthrough and cost reduction, more and more laboratories have started to apply different omic techniques over the same studies. In this project, our aim was to characterize the response of human fibroblasts to ionizing radiation in order to identify pathways, biomarkers and druggable targets to prevent or treat radiodermatitis using a multi-omic approach. For this, primary fibroblasts were cultured and exposed to two doses of X-ray radiation. While the transcriptomic assay was carried out using RNA-Seq, the proteomic and phosphoproteomic assays were performed quantitatively using a SILAC approach. Through different bioinformatic resources, we processed, analyzed and integrated the high amount of biological information contained in this data set. In the proteomic and phosphoproteomic datasets, we performed a quality control of the protein extraction and analyzed the significance of the changes detected by SILAC. For the RNA-Seq analysis, quality of the reads was examined using FastQC and FastQScreen, and after this the differential expression analysis was carried out using a HISAT2-featureCounts-DESeq2 pipeline. Changes at the different omic levels were related using a functional approach by classifying the genes, proteins and phosphopeptides in different categories (GO Terms, KEGG signaling circuits, KSEA activities…etc). Although we found similarities between the different omic levels at functional categories that are related with the biological problem, as the RNA splicing machinery or inflammatory processes, we did not find such similarities in other pathways that we would expect to have more relevance in the analysis. These results highlight the need for the development of bioinformatic integrative models for multi-omics data from different high-throughput experimental platforms to better understand molecular mechanisms of complex biological processes.

99

Manuel Molina Marín[1], Leszek P. Pryszcz[2], Uciel Chorostecki[1] and Toni Gabaldón[1]

[1]*Centre for Genomic Regulation, CRG, Barcelona, Spain*
[2]*International Institute of Molecular and Cell Biology, Warsaw, Poland*

MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence

Abstract: Orthology and paralogy relationships is the cornerstone of comparative genomics. A growing number of fully-sequenced species are compared using phylogeny based approaches, a information that is currently dispersed across different websites and applications. There is a need to have a structured, simple and agile access to such data. With this in mind, we developed MetaPhOrs, that combines the phylogenetic information of a multitude of databases to predict the orthologs and paralogs of 1,3 million proteins across 3,095 fully sequenced genomes. In addition Metaphors exploits existing redundancy across datasets to develop a confindence score that helps the user to identify the reliability of the sources as well as the level of coherence between the predictions. In this improved version, we have updated the user interface and the underlying database that includes 9 millions of phylogenetic trees from 3,095 species. Furthermore, we have added useful searches and filters, as well the possibility to download the data in several useful formats.

100

Carolina Monzó[1], Azahara Fuentes-Trillo[1], Arrate Pereda[2], Verónica Lendínez[1], Pablo Marín-García[3,4,5], Vicente Arnau[6,7], Guiomar Perez de Nanclares[2] and Felipe Javier Chaves[1,8,9]

[1]*UGDG-INCLIVA, Spain*
[2]*BioAraba National Health Institute, Hospital Universitario Araba-Txagorritxu, Spain*
[3]*UCV, Spain*
[4]*Kanteron Systems, Spain*
[5]*MGviz.org, Spain*
[6]*UV, Spain*
[7]*I2SysBio, Spain*
[8]*CIBERDEM, Spain*
[9]*Sequencing Multiplex, Spain*

Hybridization-based capture data correction of overrepresented sequences bias.

Abstract: Hybridization-based capture sequencing approaches are widely used in exome and gene-panel enrichment for library preparation. They require a targeted probe design and hybridization evaluation, in order to identify possible biases and errors in amplification, that may induce to errors in downstream bioinformatic analysis.

We have developed a method for reducing noise induced by overrepresented repetitive reads resulting from inadequate capture-probes disposition in experimental designs. The method integrates metrics obtained from different tools evaluating mappability, off-target reads and repetitive sequences. It divides the raw data into a curated set of high confidence reads required for single nucleotide and copy number variants, and a second set consisting of low confidence reads for analysis of repetitive elements families.

This methodology is specially useful for Deep-Sequencing data analysis, where intronic and intergenic complex sequences obstruct primer design, and capture sequencing approaches are used instead of more intricate and specific designs, due to their low cost in library preparation.

101

Rubén Grillo Risco[1] and Francisco García-García[1]

[1]*Bioinformatics and Biostatistics Unit, Principe Felipe Research Center, CIPF, Valencia, Spain*

Functional characterization of colorectal cancer by integrative microRNA and mRNA transcriptome analysis

Abstract: Colorectal cancer (CRC) is one of the three most common cancers in humans. High incidence is associated to age, and the correlation between metastases and mortality is known, however the genomic mechanisms involved in the onset and progression of CRC are still not understood.

The enrichment functional analysis provide a direct interpretation of studies with data coming from high throughput technologies as microarrays and massive sequencing. Therefore, the aim of this work is to define an integrative functional characterization of microRNA and mRNA in colorectal patients, what will allow a better knowledge of the regulation mechanisms in this pathology, as well as the biomarkers detection.

This project has a double objective: 1) identification of the functions associated to mRNA and microRNA profiling with a higher or lower differential expression. 2) evaluation of this methodology for functional integration of expression data, comparing complementary scenarios of interest: microRNA, mRNA and both at the same time. These approaches are carried out using different functional terms (Gene Ontology, KEGG and Reactome pathways).

To achieve these objectives we apply several multidimensional models from R, for a microarray dataset of 65 patients in different stages of colon cancer.

102

José María Medina Muñoz[1], Ricardo Lebrón Aguilar[1], Cristina Gómez Martín[1], Michael Hackenberg[1] and José Lutgardo Oliver Jiménez[1]

[1]*UGR, Spain*

The epigenetic clock: computational identification of aging-associated methylation markers

Abstract: DNA methylation is a biochemical process where cytosines are marked with a methyl group. Methylation is often described as a silencing epigenetic mark due to its involvement in gene and transposon silencing. The emergence of Next Generation Sequencing methods has enabled to analyze the methylation at single base resolution (Rajesh & Jaya 2017). Methylation's role depends on the genomic context, e.g. it's associated with inhibition or stabilization of transcription when it acts on promoter or gene body, respectively (Yang et al. 2014). Furthermore, methylation is not static and it varies between physiological and pathological conditions, individuals, tissues, and even cells from the same tissue (Roadmap Epigenomics Consortium et al. 2015). This phenomenon is known as differential methylation and leads to Differentially Methylated Cytosines (DMCs) (Lebrón et al. 2017). Particularly, we studied cytosines whose methylation level varies among individual of different age. Cytosines whose methylation level show a correlation with the age of the individual are known as Age correlated CpGs. The use of this Age correlated CpGs as a biomarker for biological age is very interesting in fields like medicine, anthropology or forensic genetics (Horvath & Raj 2018). We used and developed the bioinformatic tools needed for obtaining and analyzing CpG sites from whole genome bisulfite treated samples. We obtained two sets of Age correlated CpGs from 14 brain samples, using Pearson correlation coefficient and Spearman correlation coefficient, respectively. We developed a TrackHub called AgingHub to visualize our data in the UCSC Genome Browser, then implementing the scripts needed for analyzing the enrichment of genomics elements of interest in these Age correlated CpGs.

Horvath, S. & Raj, K., 2018. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nature Reviews Genetics, 19(6), pp.371–384.

Lebrón, R. et al., 2017. NGSmethDB 2017: enhanced methylomes and differential methylation. Nucleic acids research, 45(D1), pp.D97–D103.

Rajesh, T. & Jaya, M., 2017. Next-Generation Sequencing Methods. Current Developments in Biotechnology and Bioengineering, pp.143–158.

Roadmap Epigenomics Consortium, R.E. et al., 2015. Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), pp.317–30.

Yang, X. et al., 2014. Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. Cancer Cell, 26(4), pp.577–590.

103

Cristina Gómez-Martín[1], Marta Santalla[2], Rafael Lozano[3], Ana María Gonzalez[2], Ricardo Lebrón[1], Michael Hackenberg[1] and Jose L. Oliver[1]

[1]*Department of Genetics, Faculty of Science, University of Granada, Spain*
[2]*Grupo de Genética del Desarrollo de Plantas, Misión Biológica de Galicia-CSIC, Pontevedra, Spain*
[3]*Centro de Investigación en Biotecnología Agroalimentaria (CIAIMBITAL), Universidad de Almería, Spain*

Transcriptomic changes associated with fruit quality and crop productivity in common bean

Abstract: Fruit quality and crop productivity in legume species are critically influenced by maturation and dehiscence developmental processes (Tang et al., 2013). These stages of fruit development are well characterized in the Arabidopsis plant model but not in legumes, as is the case of common bean Phaseolus vulgaris (Li and Olsen, 2016). Here, we used RNA-seq to compare immature and mature stages of pod development in two accessions (PMB0225 and PHA1037) of this species, which also differ in other important agronomical traits such as dehiscence and fiber content. The most recent update of the Phaseolus vulgaris reference genome was used for the alignment of RNA-seq short reads, then computing fold changes among the different developmental stages. In this way, we were able to derive a list of differentially expressed genes (DEG) that are associated with pod maturation in both accessions, mainly involved in transmembrane transporter activity, carbohydrate metabolism and photosynthesis. Furthermore, among DEGs between dehiscent and indehiscent fruits we found genes related to oxidation-reduction and lipid metabolic processes.

Acknowledgements: This work was financially supported by the Ministerio de Economía y Competitividad (AGL2014-51809-R, AGL2015-64991-C3-1-R, AGL2017-88702-C2-2-R and AGL2017-88174-R) and UE-FEDER Program.

References

[1]Tang, H., et al., 2013, Proc. Natl. Acad. Sci. U.S.A. 110, 15824–15829.

[2]Li, L.F., and Olsen, K., 2016, Curr. Top. Dev. Biol. 119, 63–109.

104

Víctor Fernández-Rodríguez[1], Toni Gabaldón[2] and Salvador Capella-Gutiérrez[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Centre for Genomic Regulation, CRG, Spain*

Improving trimAl ability to cope with heterogeneous multiple sequence alignments

Abstract: Alignments of biological sequences, called Multiple Sequence Alignments (MSA), are the entrypoint for many biological applications including evolutionary studies. However, the current algorithms used to reconstruct them tend to minimize (or maximize) mathematical functions rather than truly representing biological events. This is especially relevant for highly variable sequence regions, where the positional homology is difficult to infer. This often produces MSAs with a high noise-to-signal ratio, which will be eventually amplificated during downstream analyses.

Thus, MSAs refinement has become a common practice in many biological domains. However, MSAs refinement algorithms are not except of errors so further investigation is needed making this area a very active research field.

Here we present an improved version of trimAl, a popular resource aiming to improve MSAs using manual and/or automated methods. We will explain why is important to refactor trimAl's source code including issues found and solutions applied. Improvements include 1) a dynamic structure to allow adding and removing format handlers with ease, 2) memory optimization and time reduction, 3) a new reporting system which contributes to handle better errors and exceptions, 4) an internal benchmarking tool, and 5) a new MSA visualization system based on SVG. Finally we will introduce a set of new functionalities enabled by the improvements in the source code. All these efforts make trimAl ready to handle larger and/or more heterogeneous MSAs.

[1] Capella-Gutiérrez, S. & Gabaldón, T. Measuring guide-tree dependency of inferred gaps in progressive aligners. Bioinformatics 29, 1011–1017 (2013).

[2] Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma. Appl. NOTE 25, 1972–197310 (2009).

[3] Dessimoz, C. & Gil, M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11, R37 (2010).

105

Ricardo Lebrón[1], Cristina Gómez-Martín[1], Ernesto Aparicio-Puerta[1], José María Medina-Muñoz[1], Michael Hackenberg[1] and Jose L. Oliver[1]

[1]*University of Granada, Spain*

Single-CpG methylation and gene transcription correlate either positively or negatively even at long distance

Abstract: C5-methylation of cytosine is a widely studied cell-type-specific epigenetic modification that occurs in 60-90% of all CpG sites in differentiated mammalian cells. DNA methylation is generally associated with repression of transcription initiation but it can also be associated with activation of transcription, depending on the specific location where methylation occurs1. Traditionally, the association of methylation with transcription has been studied in regions, mainly in CpG islands. However, the importance of single-CpGs has recently been highlighted with the description of the so-called "CpG traffic lights" (CpG-TLs): single-CpGs whose methylation level correlates with gene-specific transcription2. CpG-TLs were studied near to transcription start site (TSS) and within gene body using the Spearman's rank correlation coefficient3. Here, we extended CpG-TLs detection from 1 Mbp upstream of TSS to 1 Mbp downstream of transcription end site (TES), including gene body. In addition to Spearman's rank correlation coefficient, we used a methylation-state-based Kruskal-Wallis test to increase the reliability of detection. CpG-TLs abundance falls slowly with distance but never becomes zero. Positively-correlated CpG-TLs are slightly more abundant than negatively-correlated and both can be found within gene bodies, upstream of TSS or downstream of TES. Promoters, enhancers, insulators and CpG islands are enriched in CpG-TLs with respect to the background. We also analyzed the influence on gene expression of CpG-TLs located in TFBSs within proximal promoters and distal enhancers. Kind of transcription factor (activator or repressor4) and if it binds either unmethylated or methylated TFBSs5 was considered. Two transcription

factors (Sp1 and relA) seem to be strongly related to CpG-TLs. In conclusion, the CpG-TLs landscape is more complex than previously described and remains largely unexplored.

1.Spruijt, C. G. & Vermeulen, M. DNA methylation: old dog, new tricks? Nat. Struct. Mol. Biol. 21, 949–954 (2014).

2.Medvedeva, Y. A. et al. Effects of cytosine methylation on transcription factor binding sites. BMC Genomics 15, 119 (2014).

3.Khamis, A. M. et al. CpG traffic lights are markers of regulatory regions in humans. bioRxiv (2017).

4.Lambert, S. A. et al. The Human Transcription Factors. Cell 172, 650–665 (2018).

5.Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 356, eaaj2239 (2017).

106

Daniel Gimenez[1], Aleksandar Kojic[1], Marc A. Marti-Renom[2], François Le Dily[3], Ana Cuadrado[1] and Ana Losada[1]

[1]*Spanish National Cancer Research Centre, CNIO, Madrid, Spain*
[2]*CNAG–CRG, Centre for Genomic Regulation, Barcelona, Spain*
[3]*Centre de Regulació Genòmica (CRG), Barcelona, Spain*

Cohesin variants SA1 and SA2 play distinct roles in chromatin organization and gene expression required for embryonic stem cell identity

Abstract: Cohesin is a DNA entrapping complex essential for genome architecture and inheritance. In somatic cells it consists of SMC1A, SMC3, RAD21 and either SA1 (STAG1) or SA2 (STAG2) resulting in two different cohesin complexes that are not functionally redundant although either one is sufficient for cell proliferation. We are aimed to analyse the specific functions of both cohesin variants in mouse embryonic stem cells (mES) by combining the study of genome-wide cohesin variants distribution by ChIP-seq, how this distribution affects specific mES genomic architecture by HiC, and finally how changes in genomic architecture due to loss of each variant might affect gene expression by RNA-seq. In these cells, genes required for pluripotency must be actively transcribed in order to preserve the pluripotent state. Such high levels of expression are reached thanks to the specific conformation of chromatin in superenhancers that ensures efficient binding of active enhancers to the promoters of pluripotency genes. In addition, mES identity requires the repression of lineage specification genes that must be kept in a “poised” state, silent but ready to be rapidly activated. Such poised state depends on the presence of polycomb complexes (PRC1 and PRC2) that define genomic regions with specific and unique structural features. Our data show that while cohesin SA1 always co-localizes with the insulating protein CTCF, there is an important fraction of cohesin SA2 that specifically co-localizes at superenhancers and polycomb repressed promoters in the absence of CTCF. By means of HiC analysis, we have explored the specific roles of each cohesin variant in different levels of chromatin organization, including A/B compartimentalization, TAD integrity, superenhancer looping and polycomb domains compaction. We have seen that cohesins SA1 and SA2 have unique and specialized functions important for mES cells identity.

107

Audald Lloret-Villas[1] and Jordi Rambla[1]

[1]*Center for Genomic Regulation, CRG, Spain*

European Genome-phenome Archive (EGA) - Granular solutions for the next 10 years

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

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

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

108

Michela Verbeni[1], Carlos Cano[1], Paul Lizardi[2] and Armando Blanco[1]

[1]*Department of Computer Science and Artificial Intelligence, University of Granada, Spain*
[2]*GENYO. Centre for Genomics and Oncological Research: Pfizer / University of Granada / Andalusian Regional Government, Spain*

A Tool for Annotation of Genetic Signatures Based on Topologically Associating Domains

Abstract: Study of 3D organization of DNA is attracting increasing interest since it has been shown to be a key player in cell regulatory machinery [1]. Topologically Associating Domains (TADs) are structural units of genomic regions, proven to be highly self-interacting. They can span from hundreds of kilobases to few megabases, potentially including a set of diferent genes together with regulatory regions and, most notably, non-coding RNA. TAD's boundaries highlight genome locations, where not only genes, but also non-coding causal variants will most likely impact regulatory function [2]. Improvements in sequencing technologies are continuously delivering biomarker signatures for many di_erent diseases, involving hundreds or thousands of diferent genomic sites, thus transforming in a challenge interpreting and identifying the underlying regulatory mechanisms for the target condition. We propose a tool for the analysis of genetic signatures and TADs, aimed at determining DNA domain characterized by a signifcant presence of signatures of interest. This approach should simplify the interpretation and further investigation of genetic markers. In particular, we show the performance of this tool employing methylation sequencing data from a pilot study on Lynch Syndrome (LS), which represents between 4% to 5% of all colorectal cancer (CRC) cases and it is caused by di_erent known mutations in DNA repair genes. There is strong evidence for early involvement of CD4+ T-cell responses during the emergence of premalignant lesions in the colon of individuals at risk of LS [3] Therefore we decided to explore whether the appearance of DNA methylation alterations in CD4+ cells in peripheral blood could serve as a very early biomarker of CRC risk. However, we stress that the proposed tool is of general purpose and can be run on any set of genomic sites of interest to identify signifcant TADs.

References

[1] Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. Nature reviews Molecular cell biology. 2015;16(4):245{57.
[2] Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics. 2013;14(6):390{403.
[3] Chang K, Taggart MW, Reyes-Uribe L, Borras E, Riquelme E, Barnett RM, et al. Immune Profling of Premalignant Lesions in Patients With Lynch Syndrome. JAMA oncology.2018;4(8):1085{1092.

109

Laura I. Furlong[1]

[1]*GRIB (IMIM-UPF, Spain*

Enabling comorbidity analyses from real world clinical data

Abstract: Disease comorbidities are a major problem for public health due to their impact on quality of life, the management of patients, and healthcare cost. The increasing availability of clinical data for research (from disease cohorts, surveys and electronic health records) offers the opportunity to discover disease comorbidity and multimorbidity patterns from the clinical history of patients by data mining approaches. The analysis of data generated during routine medical care could then be used to improve the management of patients. In this context, the availability of analytical tools to ease the study disease comorbidities over real world data is key.

Approaches for the identification and analysis of disease comorbidities, with a special focus on tools that can be used on real world data, will be presented. These tools implement a variety of classical statistical analysis as well as novel machine learning approaches for the identification of disease comorbidities, temporal disease trajectories and disease trajectory clusters [1]. These tools [2,3], are aimed at expediting the study of disease comorbidities by providing several analytical functions and different visualization options to analyse clinical data provided by the user, and can be adapted to different experimental designs. Briefly, users can i) assess the prevalence of comorbidities in specific patients populations, ii) study differences in prevalence due to factors such as age, gender and ethnicity, and iii) assess the temporal ordering of disease co-occurrences. Moreover, they can be applied to any type of disease and are agnostic to different types of disease codification. Finally, in order to gain insight on the mechanisms underlying disease comorbidities, approaches to identify relationships between diseases from the genetic, functional and pharmacological perspective, will be presented.

With this suite of tools we aim to foster a holistic analysis of disease comorbidities, from the estimation of prevalence and the dynamics of progression of diseases, to the exploration of the underlying biological mechanisms of disease associations.

FUNDING: We received support from ISCIII-FEDER (PI13/00082, CP10/00524, and CPII16/00026), the EU H2020 Program 2014–2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and was supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. The DCEXS is a “Unidad de Excelencia María de Maeztu”, funded by the MINECO (ref: MDM-2014-0370).

REFERENCES

1. Giannoula A, Gutierrez-Sacristán A, Bravo Á, Sanz F, Furlong LI. Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study. Sci Rep. 2018 Mar 9;8(1):4216.

2. Gutiérrez-Sacristán A, Bravo À, Giannoula A, Mayer MA, Sanz F, Furlong LI. comoRbidity: an R package for the systematic analysis of disease comorbidities. Bioinformatics. 2018 Sep 15;34(18):3228-3230.

3. https://comorbidity4j.readthedocs.io/

110

Carlos Menor Ferrándiz[1], Stefan Götz[1], David Seide[1], Robert Nica[1], Mariana Monteiro[1], Alejandro Rocamonde[1] and Francisco Salavert[1]

[1]*Biobam Bioinformatic Solutions S. L., Spain*

Functional Genomics Analysis of Newly Sequenced Genomes From Scratch with Blast2GO

Abstract: Functional genomics attempts to describe functions of genes and proteins by making use of data derived from genomic and transcriptomic experiments. The combination of omics data allows to better understand the relationship between geno- and phenotype on a system-wide scale.

Blast2GO is a bioinformatics platform which offers an integrated solution for functional genomics analysis of novel genomes. It offers biologists to get from raw NGS data through all steps up to the generation of biological insights for a species without any previous genome resources. The software runs out of the box, is biologist-oriented and has an intuitive design. The platform combines a rich graphical user interface with high-performance cloud computing. This allows high-throughput as well as exploratory analysis in just one place.

Most important features include:

- Structural characterizations with RNA-seq supported gene predictions

- RNA-seq features like assembly, quantification and pair- and time-course differential expressions analysis

- High-throughput functional annotation predictions

- Functional enrichment and pathway analysis

- Track-based genome browser for GFF, VCF, BAM and FASTA files

This poster provides an overview of the different analysis strategies, provides details about new features as well as describes the use-case of a non-model de-novo transcriptomics analysis covering the whole pipeline - from read-assembly up the functional analysis and generation of biological insights.

111

Maria Rigau[1], David Juan[2], Alfonso Valencia[1] and Daniel Rico[3]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Institut de Biologia Evolutiva, CSIC-UPF, Spain*
[3]*Institute oc Cellular Medicine, United Kingdom*

Intronic CNVs cause gene expression variation in human populations

Abstract: Introns comprise about half of the human non-coding genome and they can have important regulatory roles. However, mutations in introns are usually ignored when looking for normal or pathogenic genomic variation and little is known about their population patterns of structural variation and their functional implication. By combining the most extensive maps of CNVs in human populations, we have found that intronic losses are the most frequent copy number variants (CNVs) in protein-coding genes in human, affecting 4,147 genes (including 1,154 essential genes and 1,638 disease-related genes). This intronic length variation results in dozens of genes showing extreme population variability in size, with 40 genes with 10 or more different sizes and up to 150 allelic sizes. Intronic losses are frequent in evolutionarily ancient genes that are highly conserved at the protein sequence level. This result contrasts with losses overlapping exons, which are observed less often than expected by chance and almost exclusively affect primate-specific genes. By integrating CNV and RNA-seq data, we have showed that intronic loss can be associated with significant differences in gene expression levels in the population (CNV-eQTLs). These intronic CNV-eQTLs regions are enriched for intronic enhancers and can be associated with expression differences of other genes showing long distance intron-promoter 3D interactions. Our data suggests that the frequent gene length variation in protein-coding genes resulting from intronic CNVs might influence their regulation in different individuals.

112

Dmitry Repchevsky[1], Daniel Naro[2], Romina Royo[1], Silvia Llorente[3], Josep Lluís Gelpí[2] and Jaime Delgado[3]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*University of Barcelona, UB, Spain*
[3]*Universitat Politècnica de Catalunya, UPC, Spain*

GENCOM: Contributing to MPEG-G data format creation and adoption.

Abstract: Advances in DNA sequencing technologies lead to an increase in the amount of sequenced genomic data, requiring new approaches for how this information is stored and processed. The most popular formats for storing genomic data - Sequence Alignment/Map Format (SAM)[1] and its binary counterpart (BAM)[1] do not provide an acceptable level of compression. The latter was improved in the CRAM [2] format.

Alongside the need for better compression, a modern genomic storage format should also integrate metadata representation, security strategies, and standardized accesses to improve interoperability. These features are not contemplated in the CRAM specification. Furthermore, there is no standardization body guaranteeing the perennity of CRAM. This motivated MPEG standardization group to start working on the new MPEG-G ISO/IEC 23092 Genomic Information Representation standard. Countries contributing to the MPEG-G standard are involved in the development via local research efforts. Spanish research project - Secure GENomic information COMpression (GENCOM) contributed with benchmarking of compression methods, file format definition, security and information metadata. GENCOM is a collaboration project between Polytechnical University of Catalonia, University of Barcelona and Barcelona Supercomputing Center.

Here we present the MPEG-G format, going over the three main parts which composes it: file format, genomic representation and compression, metadata and security strategies. In addition to the participation in the implementation of the reference software, we developed pure Java implementation of the standard to be included in the HTSJdk library.

This integration will allow MPEG-G file format usage in modern genome analysis pipelines that include tools like GATK[4] or Picard [5]. MPEG-G opens new uses cases which were not supported by SAM/BAM or CRAM. The ISO/IEC standardization body is also a guarantee for users that the format will be supported in the future.

[1] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Durbin, R. (2009, 6). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078-2079. doi:10.1093/bioinformatics/btp352

[2] Fritz, M. H.-Y., Leinonen, R., Cochrane, G., & Birney, E. (2011, 1). Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Research, 21, 734-740. doi:10.1101/gr.114819.110

[3] ISO/IEC JTC1/SC29/WG11 MPEG, White paper on the objectives and benefits of the MPEG-G standard, January 2018, Gwangju, Korea

[4] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010, 7). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20, 1297-1303. doi:10.1101/gr.107524.110

[5] Broad Institute. "Picard Tools." http://broadinstitute.github.io/picard

113

Javier Garrayo[1], Vicky Sundesha[1], Eduard Porta-Pardo[1], José María Fernández González[1], Laia Codó[1], Alfonso Valencia[1], Josep Ll Gelpi[2] and Salvador Capella-Gutiérrez[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*

[2]*Dept. Bioquimica i Biologia Molecular. Univ. Barcelona, UB, Spain*

OpenEBench. TCGA: A community-led benchmarking use case.

Abstract: The Cancer Genome Atlas (TCGA) community is a joint effort to characterize cancer driver genes in 33 different cancer types from nearly 10,000 exomes [1]. In such an effort, several methods for predicting cancer genes from genomic data are constantly being developed and improved.

OpenEBench (https://openebench.bsc.es), is the ELIXIR benchmarking and technical monitoring platform for bioinformatics tools, web servers and workflows. OpenEBench is part of the ELIXIR tools platform ecosystem (https://www.elixir-europe.org/platforms/tools). Its development is led by the INB/ELIXIR-ES (https://inb-elixir.es) Central node at the Barcelona Supercomputing Center (BSC) in collaboration with different partners across Europe. It is focused on providing an infrastructure where end-users can learn from different available options and select the one best fitting their scientific needs; bioinformatics software developers can find relevant datasets and meaningful scientific challenges to evaluate their own developments, and communities interested in a particular scientific domain can easily define which datasets and metrics are relevant for developers to work on, which in turn will allow the field to move ahead.

TCGA, as an engaged community in OpenEbench, provides an ideal use-case to demonstrate the latest developments at OpenEBench including the deployment of use of a Virtual Research Environment (VRE). End-users can have access to the software tools who has participated in the TCGA effort to identify cancer driver genes across 33 different cancer types while developers have the opportunity to evaluate their latest developments by submitting their predictions to the OpenEBench VRE. A predefined workflow including format and data validation, metrics computation, results consolidation and visualization is executed for each submission. Results can be visualized at the VRE, downloaded and/or extended by using a number of scripts provided via the OpenEBench repository [2].

In the case of TCGA, users can test their performance against 8 different tools. A list of gene identifiers should be submitted for doing so. They are then compared against several reference datasets defined by the TCGA community depending on the different cancer types, which results in two assessment metrics (true positive rate and precision). This allows the submitted predictions to be benchmarked against the rest of the participants currently present within the TCGA community. The benchmark results are provided in a compressed file (containing the assessment data for all participants and an SVG image with the visualization) and can also be visualized in a scatter-plot where the user can then apply several classification methods implemented in OpenEBench in order to transform the results to table format, which are easier to interpret by non-expert users.

[1] Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., … Mariamidze, A. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell, 173(2), 371–385.e18.

[2] https://github.com/inab/TCGA_visualizer

114

Touria Derkaoui[1], Mohamed Mansouri[2], Ali Loudiyi[2], Amina Barakat[1], Naima Ghailani Nourouti[1], Jaime Martínez de Villarreal[3], Carlos Cortijo Bringas[3] and Mohcine Bennani Mechita[1]

[1]*Laboratory of Biomedical Genomics and Oncogenetics, Faculty of Sciences and Techniques of Tangier, Morocco*
[2]*Oncology Clinic Al Amal of Tangier, Morocco*
[3]*Genetracer Biotech Laboratory, Cantabria, Spain*

Should pre-analytical tissue handling be considered in Next Generation Sequencing?

Abstract: Introduction: Targeted therapy is an evolving approach to cancer treatment. Next Generation Sequencing (NGS) is considered as the most powerful technique for genomic profiling and have a potential utility for selecting appropriate therapy. High DNA quality is required for a successful sequencing. However, the pre-analytical treatment of paraffin tissue blocks, the most available material in oncology, can greatly affect the quality of the extracted DNA. Cytosine deamination is a major artifact observed in these samples, which alters sequencing results.

Methods: In this study, we have included 24 formalin-fixed paraffin-embedded (FFPE) samples from Triple Negative Breast Cancer (TNBC). Mutations sequencing was performed by NGS at Gene Tracer Biotech laboratory in Spain. DECODER panel for detection of hotspot somatic mutations of 70 genes related to various tumorigenesis processes was used. The committee for biomedical research in the faculty of medicine and pharmacy of Rabat approved this study.

Results: After extraction from FFPE samples, we have observed low amounts of DNA, concentrations vary between 13.08 and 219 ng, and DNA was highly fragmented and has a poor quality. In addition, analysis of sequencing results have demonstrated a huge proportion of mutations C>T and G>A. Those mutations were interpreted as sequence artifacts. The artifact observed in the paraffin samples is certainly due to cytosine deamination. To distinguish between true somatic mutations and spontaneous cytosine deamination was very difficult, and can cause false positive results.

Conclusion: It is a priority to improve pre-analytical conditions and storage of tissue samples in order to ensure the reliability of molecular analysis, which can be relied upon to guide therapeutic decision. In this sense, the uracil-DNA glycosylase enzyme have shown in vitro efficacy in the elimination of cytosine deamination at tissue samples for NGS sequencing.

115

Carla Giner-Delgado[1], Isaac Noguera[1], Paul F. O'Reilly[2] and Mario Cáceres[1,3]

[1]*Universitat Autònoma de Barcelona, UAB, Spain*
[2]*King's College London, United Kingdom*
[3]*Catalan Institution for Research and Advanced Studies, ICREA, Spain*

The effect of human polymorphic inversions on genomic variation

Abstract: Chromosomal inversions have been studied for over a century in different species for their role in processes of adaptation and speciation. Despite not changing the total amount of sequence, they can alter local genome organization and inhibit recombination between sequences in opposite orientations, affecting the nucleotide variation in the region. However, in humans, their importance as genomic modifiers remains largely unknown, mainly because their detection is technically challenging. Here, we take advantage of a unique data set of common inversions that have been experimentally genotyped in seven populations from the 1000 Genomes Project to investigate the relationship between human inversions and the surrounding nucleotide diversity. We found that inversions originated by non-homologous mechanisms follow the patterns expected from simulations. First, the variation within the inverted haplotypes appears strongly reduced from the absence of recombination with variation-rich non-inverted sequences. Second, we observe a variation increase in regions with inversions at intermediate frequencies, that mimics balancing selection signals detected by statistics such as Tajima's D and is stronger in old inversions. Third, inversions nearly fixed in the population exhibit a strong reduction of variation similar to that of a selective sweep. These effects on variation are visible in the entire length of the inverted sequence and up to 20 kb of the flanking regions, indicating a total

suppression of recombination in the central region which possibly extends past the inversion breakpoints. In contrast, inversions flanked by inverted repeats (created by non-allelic homologous recombination) show no clear signs of inhibition of recombination and their haplotype diversity indicates multiple inversion events on independent sequences. In conclusion, we have shown that inversions in the human genome have a strong role as modifiers of local genomic variation and offer the first detailed description of these effects in multiple cases. Our results can therefore have important implications for applications based on nucleotide variation levels, such as genome-wide association studies or selection scans.

116

Mattia Bosio[1], Alfonso Valencia[1,2,3] and Salvador Capella-Gutiérrez[1,2]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Spanish National Bioinformatics Institute (INB), ELIXIR-ES, Spain*
[3]*Catalan Institution for Research and Advanced Studies, ICREA, Spain*

Improving RNA-Seq germline variant calling within RD-connect consortium

Abstract: Within the RD-Connect consortium we developed an RNASeq variant calling pipeline, automating sample analysis for the analysis platform. We integrated best practices from ENCODE Consortium, GATK toolkit, and custom filtering to produce reliable variant calls.

Analysing matching WGS and RNA-seq data, we investigated similarities i.e. genotype concordance and overlap, which are in line with the literature, and differences between variant datasets (high number of RNA false positives). We then developed a post-processing strategy for RNA-seq variants to produce more reliable calls and characterize the substantial fraction of novel variants compared to WGS (i.e. which variants are of biological origin and which of technical errors). For this task, we integrated relevant features from best practices and RNA-seq calling protocols e.g. [1-3], and evaluated their impact on concordance, sensitivity, and precision against WGS callset.

Aiming to a better characterization of variants, not limited to known sites, we jointly processed a set of 20 samples with matching DNA and RNA-seq data, modeling how different features can help discriminating variants falling in two sets i.e. RNA private variants like RNA-editing and/or false positives, and shared with WGS; via a random forest estimation. With this, we can estimate each variant likelihood of being a false positive, enabling a deeper characterization of RNA-seq variants, and easing the downstream integration efforts and/or the interpretation process.

The developed classification framework showed to improve the precision-recall performances of the standard variant calling pipeline with respect to finding germline variants. Comparing our framework with similar approaches like [1], on independent samples from RD-Connect and from public gold-standard resources (Gene in a bottle consortium, NA12878 sample), we showed that the random forest classification achieves superior results than hard-filtering strategies.

[1] Piskol et al. Am. J. Hum. Genet. 2013
[2] Oikkonen el al. Wellcome Open Res. 2017
[3] Xu. Computational and Structural Biotechnology Journal, 2018

117

Pablo Minguez[1] and Perceval Vellosillo[1]

[1]*IIS-Fundacion Jiménez Díaz, Spain*

Resources for the functional characterization of post-translational modifications

Abstract: The structural stabilization and functional regulation of proteins are partially controlled by means of post-translational modifications (PTMs). There are more than 200 PTM types already described that act at proteome level and are conserved over evolution. Thanks to the new advances in Mass Spectrometry techniques, large datasets have been produced for a selection of PTM types under several conditions and for main model organisms. However, the functional meaning of individual PTMs and their fitting into the post-translational regulatory puzzel needs still of methods and resources that help to distinguish between functional and non-functional modifications as well as providing specific information about their role. Based on the developments and resources available at the PTMcode database (http://ptmcode.embl.de) which include methods based on co-evolution, structural proximity and manual annotation to catch crosstalks between PTMs that regulate proteins and their interactions, we are now using genomic information stored in databases of human variants in order to add new insights to PTM function. New data will be incorporated into the PTMcode database

---

118

Ana Dueso-Barroso[1], Montserrat Puiggròs[1] and David Torrents[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*

Identification of expressed processed pseudogenes across multiple cancer types.

Abstract: (SCIENTIFIC POSTER ONLY)

The identification and characterization of somatic events associated with cancer is key to understand the biology behind tumors formation and progression, and might provide with new clinical markers. The formation of processed pseudogenes (PPs) in tumor cells is considered a somatic event with potentially diverse functional consequences. In the context of Pancancer consortium, and in collaboration with several groups, we have characterized the landscape of somatic processed pseudogenes across more than twenty tumor types (Bernardo Rodriguez-Martin et al. Nature Genetics, in press).

We have first developed a protocol to identify PPs through the analysis of the landscape of somatic variation obtained from ICGC-PanCancer variant calling pipelines and SMUFIN over 2859 tumor-normal whole genome pairs coming from 34 different tumor types and subtypes. Our methods combine automatic approaches with manual inspection of the sequences to verify the candidate pseudogenes.

In total, we found evidences for 433 somatic retrotranscription and integration of coding mRNAs across 251 tumor genomes, ranging from complete copies to different fragment sizes. We observed that somatic PPs are not equally distributed across the different tumor types analyzed. In particular, Lung Squamous Cell Carcinoma, Head and Neck Squamous Cell Carcinoma and Esophageal Adenocarcinoma show the highest frequency among patients. Moreover, this distribution of PPs is positively correlated with the activity of L1 retrotransposon, which is also predominant in the same tumor types.

The integration of somatic PPs found across all tumor types, also appears to be enriched in more accessible parts of the chromatin, like gene regions, where we identified 49% of the cases. It is among these cases, where we have explored evidences for expression using RNA-seq data. We identified read support for the expression of 17 PPs, across 14 different samples and 6 different tumor types. Among these 17 expressed PPs, 14 of them were inserted in different gene parts, generating diverse forms of PP-host gene fusion transcripts with different potential forms of functional interactions. The reconstruction of the potential PP-host gene fusion transcript predicts that the major form of PP

insertion generates a premature stop codon within the coding region of the host transcript. Taken together, these results show that processed pseudogene formation can alter the transcription on cells.

119

Pablo Rodriguez-Brazzarola[1], Sergio Díaz Del Pino[1], Esteban Pérez Wohlfeil[1] and Oswaldo Trelles[1]

[1]*Universidad de Málaga, Spain*

Saving time for visual analytics multi-genome comparisons

Abstract: Due to major breakthroughs in sequencing technologies throughout the last decades, the time and cost per sequencing experiment has reduced drastically, overcoming the data generation barrier during the early genomic era. This encouraged the scientific community to develop computationally methods to compare large genomic sequences between them.

For instance, a full genome comparison study between two species would require over 200 large sequence (chromosome) comparisons. Because of this, comparing multiple genomes of interest implies and exponentially increasing of the number of comparisons. However, most of these comparisons are not strictly necessary since the vast majority of them do not contain relevant information. As a result of this issue, computational methods to perform pairwise comparisons between large sequence and determine whether there are significant sets of fragments that conserve strand and collinearity have been developed by the scientific community. Hence, data generation problem is no longer an issue from a computational point of view, yet the limitations have shifted towards the analysis of such huge amount of data.

In this work we present XCout, a web-based visual analytics application for multiple genomic comparisons designed to improve the workflow of analyzing large amounts of genomic data by using novel technologies in web visualization. XCout enables the possibility to check the results of hundreds of huge comparisons further reducing the time of the analysis by identifying pairwise comparison similarities with a significant signal of interest between different chromosomes across multiple species. The comparisons are presented to the user through an approach similar to such of a microarray experiment, using a color scale to detect comparisons of interest at first sight. Moreover, the user is also capable to perform an overlap of all the comparisons between all the chromosomes of one species against another chromosome of interest, in order to notice the the origin and contributions of the conserved signals.

120

Adria Caballe[1], Antonio Berenguer Llergo[1] and Camille Stephan-Otto Attolini[1]

[1]*Institute for Research in Biomedicine, IRB Barcelona, Spain*

Adjusting for systematic technical biases in risk assessment of gene signatures in genomic cohorts

Abstract: In recent years, many efforts in clinical and basic research have focused on finding molecular features of tumor samples with prognostic or classification value. Among these, the association of the expression of gene signatures with cancer relapse is of special interest given its relatively direct applicability in the clinic and its power to shed insights into the molecular basis of cancer. Although great efforts have been invested in data processing to control for unknown sources of variability in a gene-wise manner, little is known about the behaviour of gene signatures with respect to the effect of technical variables.

In this work we show that the association estimates of gene signatures with relapse may be biased due to technical sources of variation, and propose a simple and computationally low intensive methodology based on correction by expectation under gene signature randomization. The resulting estimates are centred around zero and ensure correct asymptotic inference. Moreover, our methodology is robust against spurious correlations possibly driven by general tendencies present in the data.

121

Ricardo Holthausen[1], Sergio Díaz-del-Pino[1], Esteban Pérez-Wohlfeil[1], Pablo Rodríguez-Brazzarola[1] and Oswaldo Trelles[1]

[1]*Universidad de Málaga, Spain*

Tracing Computational Synteny Blocks along different species

Abstract: The field of pairwise sequence comparison has witnessed a relevant growth in the last years. Both the increase in computer performance and the development of new tools have allowed the collection of these comparisons in linear time with a controlled memory consumption, thus opening the door to ideas hitherto unfeasible, helping us to acquire a better understanding of the world we live in.

One of the topics that, yet thoroughly studied, needs more grasp, is the evolution of species. Throughout the years, organisms' DNA have replicated, originating mutations with it, and therefore new species. A better understanding of this process could provide us not only with a broader and more profound knowledge about what we are and where are we headed, but also with the possibility of detecting information related to certain characteristics and diseases.

When dealing with genetic data, the level of detail of our approach depends to a large extent on the objective we are focused on. Computational Synteny Blocks (CSB), which are sets of fragments with collinearity and the same strand, can be the proper viewpoint from which to address this problem.

Up until now, few projects have somewhat addressed this problem, being possible the collection of CSBs between species. Yet, no systematic and general approach has been carried out, in the sense that currently there is no way of tracing related CSBs between comparisons of different sets of species nor a standard procedure to assess these blocks' relation.

The tool that is being developed for this work is a 'block tracer', that is, a way for the researchers to select a certain set of comparisons between species' chromosomes and follow the CSBs found that are directly linked between one another. In order to trace these blocks in a set of species, comparisons between these species have to be previously obtained in a linked way.

This work aims to make possible an almost instant tracing of shared CSBs along sets of species. Thus, a more profound understanding of them can unravel new insights with respect to the underpinnings of evolution, which can provide us with valuable information regarding prevention of diseases which etiology is related to DNA mutations.

122

Oscar Reina Garcia[1], Fernando Azorin[1] and Camille Stephan-Otto Attolini[1]

[1]*Institute for Research in Biomedicine, IRB Barcelona, Spain*

chroGPS2: differential analysis of epigenome maps in R.

Abstract: In the last years, after systematic mapping of epigenomics data from multiple organisms, tissues and cell lines, the ability to efficiently integrate, visualize and compare such information remains a challenge.

Here we present chroGPS 2.0, a major update of our previously developed software chroGPS, for visualization and differential analysis of epigenomes. Methods are provided for efficient integration and comparison of data from different conditions or biological backgrounds, accounting and adjusting for systematic biases in order to provide an efficient and statistically robust base for differential analysis. We also include functionalities for general data assessment and quality control prior to comparing maps, such as functions to study chromatin domain conservation between epigenomic backgrounds, to detect gross technical outliers and also to help in the selection of candidate marks for de-novo epigenome mapping.

123

Javier Corvi[1], José M. Fernández[1], Ander Intxaurrondo[1], Martin Krallinger[1], Alfonso Valencia[1] and Salvador Capella-Gutiérrez[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*

Updating the LimTox Content provider workflow

Abstract: Toxicological reports are relevant for the study, assessment and elaboration of chemical compounds and drugs; and from a general point of view for pharmacological and biological research. The available literature in the field is vast and came from a variety of sources including public ones like NCBI PubMed, Europe PMC, European Public Assessment Reports (EPAR) from the European Medical Agency, and private ones from the internal toxicology studies carried on by pharmaceutical companies. It is of great importance and interest to develop text mining tools to automatically analyze relevant toxicology literature to detect potentially adverse effects.

Based on these needs, the LimTox (Literature Mining for Toxicology) system was developed in 2014 and published in 2017 [1]. LimTox is a web-based online biomedical search tool with special focus on adverse hepatobiliary reactions. It integrates a range of text mining, named entity recognition and information extraction components with machine-learning techniques. Although its main focus is on adverse liver events, LimTox also enables basic searches for adverse toxicity events in other organ e.g. nephrotoxicity and/or cardiotoxicity.

Here we present an update of the workflow used to update LimTox content, this new version consists in redeploying and extending the LimTox implementation and functionality by adding features and improvements to the system. One of the major changes is the use of software containers to ease the system deployment anywhere. To facilitate the workflow maintainability, extension and adaptation, the LimTox analytical workflow has been rewritten into modules which are then associated to individual software containers. Indeed this flexible architecture has contributed to an extensive testing by trying different training strategies and corpora, and to develop a generic mechanism for adding new adverse toxicological events endpoints to the systems, and, therefore, to LimTox. Initial results shown an accuracy of >0.95 for abstract classification of adverse liver toxicological events while the sentence classification is >0.85.

[1] Cañada, A., Capella, S., Rabal, O., Valencia, A., and Martin Krallinger, M.* (2017). LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. Nucleic Acids Research, 7 Web Server Issue, doi:10.1093/nar/gkx462.

124

Mario Prieto Godoy[1], Helena Deus[2] and Mark Wilkinson[1]

[1]*Centro de Biotecnología y Genómica de Plantas, Spain*
[2]*Elsevier, Portugal*

Unraveling Certainty in Bio-Scholarly Statements

Abstract: The volume of scholarly articles published every year has doubled in the last two decades, in the biomedical domain rising from ~ 300,000 articles in 1996 to more than 800,000 in 2016. As a consequence, researchers cannot read all articles, even in their own domain, there is the imperative to maintain (the appearance of) "comparative productivity", there is reviewer fatigue, and thus, increasingly, scientists' personal biases are being reflected in scholarly publications (Fanelli 2010) (Sarewitz 2016), and going undetected and uncorrected. Scholars acknowledge one another through citations, yet this raises a similar set of problems - citations have been shown to "drift" in their intensity or their meaning compared to the assertion in the referenced material (De Waard and Maat 2012), likely reflecting the biases of the writer. This phenomenon may be amplified within citation chains, sometimes resulting in near-factual statements which, at the origin, were far more speculative; all of this happening in the absence of any additional evidence. Finally, the concepts within this volume of literature are increasingly being captured via text-mining, where there is no capacity to quality-control for these kinds of phenomena, thus masking the problem further.

Here, we apply questionnaires to measure the ability of researchers to discern various levels of certainty being expressed in the scientific literature, determine their level of agreement, formally define categories of certainty, and then create automated classifiers for scholarly statements. Three Web-based questionnaires were e-mailed to researchers spanning both medical and plant/agricultural biotechnology research, asking them to evaluate scholarly assertions for certainty. Agreement between participants was assessed by Weighted Kappa Cicchetti (Cicchetti, Lord, Koenig, Klin, & Volkmar, 2008). Completed surveys were returned by 270 researchers (Q1-75, Q2-150, Q3-45). Classifiers were created through Machine Learning approaches, using a Neural Network algorithm to automatically assign certainty levels to new statements. Our results showed an >80% of accuracy in the most basal assignment, binary classification problem. Such algorithms can now be used in-tandem with text-mining tools in order to capture the degree of certainty being expressed in original text-mined information. We discuss also how these tools can be used to detect the certainty "drift" problem, as well as pinpoint "certainty inflection points" along a citation chain, which may be associated with data or a dataset that can then be explicitly associated with the increase in certainty of a scholarly assertion.

De Waard, A., and H. P. Maat. 2012. "Epistemic Modality and Knowledge Attribution in Scientific Discourse: A Taxonomy of Types and Overview of Features." Proceedings of the Workshop on Detecting. http://dl.acm.org/citation.cfm?id=2391180.

Fanelli, Daniele. 2010. "Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data." PloS One 5 (4): e10271.

Sarewitz, Daniel. 2016. "The Pressure to Publish Pushes down Quality." Nature 533 (7602): 147–147.

125

Val Fernández Lanza[1], Miguel Díez Fernández de Bobadilla[2], Alba María Talavera Rodríguez[2], Fernando Baquero[2] and Teresa Coque[2]

[1]*Bioinformatics Unit, Ramón y Cajal Health Research Institute, IRYCIS, Spain*
[2]*Department of Microbiology, Ramón y Cajal University Hospital, Ramón y Cajal Health Research Institute, IRYCIS, Spain*

Accnet2: A tool for accessory genome comparison and statistical analysis.

Abstract: Introduction: Accnet2 is a bioinformatics tool to analyze accessory genomes of microbial populations. The new version of Accnet2 includes two new modules: statistical analysis and pan-genome analysis. Moreover, Accnet2 has been rewritten to update the pipeline with more efficient tools.

Methods: Accnet2 builds a bipartite network with genomes (or genomic units) and proteins. Each genome has linked with their corresponding proteins. In this new version of the software, we have included a statistical module that allows inferring a genome clustering and a proteins clustering. The genome clustering groups the genomes into phenotypic units that share specific proteins. On other hands, the proteins cluster shows the predisposition to the co-occurrence of specific proteins. Moreover, the new statistical module performs an enrichment analysis of the proteins present in each cluster (predefined or automatic). Finally, the new pan-genomic module allows to create and study pan-genomes. In this case, we use the frequency of each protein in the pan-genome as edge-weight in the bipartite network. This module accepts predefined pan-genomes or it can create them.

Results: The new version of Accnet improves the performance of the software. Now, we are able to analyze up to 1.000 strains at the same time. Moreover, the tool shows statistical results to infer which proteins seem relevant of each population.

Source code: https://github.com/valflanza/accnet2

126

Ancor Sanz-García[1], Laura Hevia[1], Alejandra Reolid[1], Ester Muñoz-Aceituno[1], Mar Llamas-Velasco[1], Esteban Daudén[1], Francisco Abad-Santos[1]and María Carmen Ovejero-Benito[1]

[1]*Instituto de Investigación Sanitaria La Princesa (IIS-IP), Spain*

Copy number variation associated with psoriasis anti-TNF response and the development of psoriasiform reactions

Abstract: Introduction - Copy number variations (CNVs) have been described to play a role in psoriasis. Anti-TNF drugs (adalimumab, etanercept and infliximab) are used to treat moderate-to-severe psoriasis patients resistant to conventional systemic drugs. These therapies are highly effective and safe. However, around 30-50% of the patients do not present an adequate response to them. Moreover, they are expensive and, rarely, can cause severe adverse effects such as the development of psoriasiform reactions. Purpose - To identify CNVs that could predict anti-TNF drugs response in moderate-to-severe psoriasis patients. To describe CNVs that could anticipate patients who will develop psoriasiform reactions. Methods - DNA was isolated from peripheral blood cells of 70 moderate-to-severe psoriasis patients treated with anti-TNF drugs. Patients exhibiting extreme phenotype response to anti-TNF drugs were selected to enhance differences observed between them. Thus, patients were classified into excellent responders (ER, n=49) and partial responders (PR, n=21) to anti-TNF drugs. Samples were bisulfite converted and hybridized in a 450K methylation microarray. CNVs were detected using diverse R packages such as conumee. Statistical analysis were performed with the CNVs exhibiting the same location and length in the analyzed groups. Afterwards, these results were corrected by the multiple comparison procedure for false discovery rate (FDR). Results - We found 1,406 CNVs between ER and PR. However, this significance was lost with the FDR correction. The analysis of CNVs related to development of psoriasiform reactions, resulted in 1,338 CNVs, from which 24 reached significance after FDR. Conclusions -This is the first pharmacogenomic study that analyzes genome-wide CNVs associated with anti-TNF drugs response in moderate-

to-severe psoriasis. We have shown that CNVs cannot discriminate between ER and PR in moderate-to-severe psoriasis patients. However, 24 CNVs have been associated with the manifestation of psoriasiform reactions. In conclusion, these results could help to anticipate which patients could develop psoriasiform reactions.

127

Gemma Bullich[1], Steven Laurie[1], Jordi Morata[1], David Ovelleiro[1], Sandra Redó[1], Ricky Joshi[1], Cristina Luengo[1], Leslie Matalonga[1], Genís Parra[1], Raul Tonda[1] and Sergi Beltran[1,2]

[1]*CNAG-CRG, Spain*
[2]*Universitat Pompeu Fabra, UPF, Spain*

Implementation of a Copy Number Variant detection workflow within the URD-Cat pilot project on Personalised Medicine.

Abstract: Background

The Undiagnosed Rare Disease Program of Catalonia (URDCat) is a pilot project aiming to provide the Catalan Health System with personalised genomic medicine as a fully integrated service for patients with rare diseases. Genotypic and phenotypic data is collated, integrated and analysed through the RD-Cat platform (rdcat.cnag.crg.eu), based on the RD-Connect platform (platform.rd-connect.eu). The RD-Cat platform already integrates pseudo-anonymised clinical and genomic data of more than 1300 individuals, including 800 RD patients.

RD-Cat has collated 413 pre-existing exomes and genomes, and has sequenced 406 new whole exomes, 94 new genomes, and 23 RNA-Seq experiments. Exome sequencing is widely accepted as a robust and cost-effective approach for single-nucleotide variant identification. However, detection of copy number variants (CNV) is still challenging with low sensitivity and high false positive rates due to the targeted nature of exome-capture protocols. With the aim of implementing a standard workflow for the detection of CNVs within the URDCat project, we have compared the results obtained from the analysis of all available exomes with 3 tools, ExomeDepth, XHMM and Conifer, selected after a preliminary evaluation of 6 tools.

Methods

WES data from 809 samples were split by capture kit and independently analysed with ExomeDepth, XHMM and Conifer using default settings. CNVs overlapping at least 50% with CNVs reported by Conrad et al (2010) cohort were discarded. To compare the results obtained by the 3 programs, only CNVs with an observed frequency <1% within our cohort were considered.

Results

Analysis of the output of the 3 tools showed at least one CNV was identified in approximately 90% of the individuals according to ExomeDepth (717 out of 809) and XHMM (710 out of 809), whereas in only 12% of the cohort according to Conifer (100 out of 809). The mean number of CNV calls per sample was 33 for ExomeDepth, 48 for XHMM and 22 for Conifer, considering only those samples with at least 1 CNV identified. The median size of the predicted CNVs was 4.4 Kb for ExomeDepth (ranging from 1 bp to 15863 Kb), 12 Kb for XHMM (ranging from 57 bp to 17966 Kb) and 57 Kb for Conifer (ranging from 392 bp to 23603 Kb). Altogether, only 4% of all CNV's were called by all three tools. The highest overlap (18%) was between ExomeDepth and XHMM, while ExomeDepth and Conifer shared the lowest number of CNVs (6%).

Conclusions

Our preliminary results showed that CNV detection using 3 different tools in parallel might increase sensitivity as it detects a wide range of CNV sizes. However, low concordance exists among the 3 different programs. A confidence score based on the degree of overlap might be useful to stratify the CNVs according to their probability to be real.

129

Laia Codó[1], José M. Fernández[1], Dmitry Repchevski[1], Genís Barrayi[2], Vicky Sundesha[1], Alba Serrano[1], Alfonso Valencia[1], Salvador Capella-Gutiérrez[1] and Josep Ll. Gelpí[1,3]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Institute for Research in Biomedicine, IRB Barcelona, Spain*
[3]*Department de Bioquímica i Biomedicina Molecular, Universitat de Barcelona, UB, Spain*

OpenEBench virtual research environment

Abstract: OpenEBench (https://openebench.bsc.es), is the ELIXIR benchmarking and technical monitoring platform for bioinformatics tools, web servers and workflows. OpenEBench is part of the ELIXIR tools platform ecosystem (https://www.elixir-europe.org/platforms/tools). Its development is led by the INB/ELIXIR-ES (https://inb-elixir.es) Central node at the Barcelona Supercomputing Center (BSC) in collaboration with different partners across Europe. It is focused on providing an infrastructure where end-users can learn from different available options and select the one best fitting their scientific needs; bioinformatics software developers can find relevant datasets and meaningful scientific challenges to evaluate their own developments, and communities interested in a particular scientific domain can easily define which datasets and metrics are relevant for developers to work on, which in turn will allow the field to move ahead.

The OpenEBench Virtual Research Environment (VRE) is the e-infrastructure that integrates OpenEBench resources with the purpose of developing, evaluating, and eventually offering an unified benchmarking service useful for the different scientific benchmarking paradigms.

OpenEBench VRE offers a complete web interface which allows integrating public and/or consolidated benchmarking datasets, private participants data, and the necessary mechanisms to import and execute the benchmarking workflows on top of cloud computing infrastructures, like the ones at the Barcelona Supercomputing Center (BSC) facilities. Given that each community agrees on their own result formats and metrics, the community responsibles have to provide methods to: syntactically validate the participant results; compute the defined metrics over the results; compute the assessed results of the challenge, which combines the computed metrics from all the participants. Also, the community responsibles can provide customized visualization methods to browse participant results, individual metrics and/or assessment.

130

Salvador Capella-Gutiérrez[1], Juergen Haas[2], Vicky Sundesha[1], Dmitry Repchevski[1], Javier Garrayo[1], Víctor Fernández-Rodríguez[1], Miguel Madrid[3], Laia Codo[1], José M. Fernández[1], Anália Lourenço[4], J.L. Gelpí[1] and Alfonso Valencia[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Swiss Institute of Bioinformatics, SIB, Switzerland*
[3]*Centre de Recherches en Cancérologie de Touluse, CRCT, Spain*
[4]*Universidad de Vigo, ESEI., Spain*

OpenEBench. The ELIXIR platform for benchmarking

Abstract: Benchmarking is intrinsically referred to in many aspects of everyday life from assessing the quality of stock market predictions to weather forecasting to predictions in the life sciences, such as 3D protein structure predictions. On an abstract level, benchmarking is comparing the performance of software under controlled conditions. Benchmarking encompasses the technical performance of individual tools, servers and workflows, including software quality metrics, as well as their scientific performance in predefined challenges. Scientific communities are responsible for defining reference datasets and metrics, reflecting those scientific challenges. In the context of the H2020 ELIXIR-EXCELERATE project, we have developed the OpenEBench platform aiming at transparent performance comparisons across the life sciences. OpenEBench supports the scientific communities by assisting in setting up emerging benchmarking efforts, foster exchange between communities and ultimately aims at making benchmarking not only more transparent, but also more efficient.

We will present the current OpenEBench and a preview on the upcoming implementation, which will be strongly focused on assisting communities to join the platform. Current implementation covers the widgets gallery, which has been developed to summarize and export OpenEBench data to other platforms; the experts to non-experts visual transformation of scientific benchmarking results, and the assessment of quality metrics for the technical monitoring of bioinformatics resources. For OpenEBench we are working in three levels of operation: level 1 aims to collect and distribute data from established benchmarking communities, via the OpenEBench API; level 2 is based on computing benchmarking metrics within the platform; while level 3 will extend the existing OpenEBench platform to execute benchmarkable workflows (provided as software containers) using identical conditions to ensure an unbiased technical and scientific assessment. Overall, OpenEBench provides an integrated platform to orchestrate benchmarking activities, from the deposition of reference data to test software tools, to the provision of results employing metrics defined by scientific communities.

131

Salvador Capella-Gutiérrez[1], Vicky Sundesha[1], José M. Fernández[1], Josep Ll Gelpi[2] and Alfonso Valencia[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Dept. Bioquimica i Biologia Molecular. Univ. Barcelona, UB, Spain*

The Spanish National Bioinformatics Institute (INB): 15 years

Abstract: The Spanish National Bioinformatics Institute (INB) is the ELIXIR Node in Spain. The INB was created in 2003 following the Swiss Bioinformatics Institute (SIB) model of a distributed organization of nodes with a central coordination hub. INB is currently one the technological platforms of ISCIII, with two main objectives: maintaining and increasing its alignment with ELIXIR, looking for deeper synergies; and increasing its translational medicine ties with the Spanish National Health System (SNS).

In the recent renewal process (2018 - 2020), INB/ELIXIR-ES has growth from 10 groups to 19 groups to achieve the mentioned strategic objectives. The first objective is demonstrated by the decided support to the European Genome-phenome Archive (EGA) jointly managed with the EMBL-EBI, which was named ELIXIR Core Data Resource and included

in the recommended ELIXIR Deposition databases; the support to the RD-connect as one of the main assets for the rare-diseases community in ELIXIR; the co-leadership of the ELIXIR tools platform; the (co)lead and participation in a number implementation studies including "Development of Architecture for Software Containers at ELIXIR and its use by EXCELERATE use-case communities", "Remote real-time visualization of human rare disease genomics data (RD-Connect) stored at EGA" and different editions of "Beacon".

Other relevant activities in the context of ELIXIR include the promotion of FAIR data principles and scientific software development best practices; the participation in the Staff Exchange programme by hosting colleagues from Sweden, Finland, Italy and Slovenia; the organization of training activities in collaboration with other nodes e.g. High Performance Computing jointly organized with ELIXIR-Slovenia; and projects e.g. the Bring Your Own Workflow workshop in collaboration with the Center of Excellence for Biomolecular Research (BioExcel).

In the translational arena, INB/ELIXIR-ES has designed and promotes TransBioNet (Translational Bioinformatics Network), which aims to bring together bioinformatics groups working at Health Research Institutions of the National Health System. In this way standards, protocols, best-practices and training activities can have a direct impact on researchers working closely to clinicians. Related to these researchers, INB is setting up an identity provider, which is going to be integrated into ELIXIR AAI, in order to allow members from outside RedIRIS to have access to ELIXIR resources.

132

Miguel Ponce de Leon[1], Javier Perales-Patón[2], Héctor Tejero[2], Fatima Al-Shahrour[2] and Alfonso Valencia[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Centro Nacional de Investigaciones Oncológicas, CNIO, Spain*

Contextualizing metabolic models with gene essentiality to predict cancer specific growth requirements

Abstract: Metabolic reprogramming is one of the earliest described hallmarks of transformation of cancer cells transformation. In recent years the study of cancer metabolism has regained attention in order to find cancer specific metabolic vulnerabilities that could be therapeutically exploited. With the increasing availability of large collections of omic data, such as the Cancer Cell Line Encyclopedia or the Cancer Genome Atlas, context-specific models (CSM) have become a powerful tool to integrate heterogeneous source of data, as well as to perform computer simulations such as in-silico gene knockout.

In this communication, we present an iterative pipeline for reconstructing and refining context-specific models of cancer cell metabolism using gene expression profiles, and genome-scale loss-of-functions screening. In order to perform in-silico simulation, we used Constraint-Based modeling as the computational framework. To test our pipeline, we chose Pancreatic Ductal Adenocarcinoma (PDAC) as a case of study. We selected 22 PDAC cell lines which include experimental data such as gene expression, mutations, copy number variation and gene essentiality profiles from genome-scale CRISPR assays. Using these datasets in combination with the Human metabolic model Recon2.2 and two different computational methods, we reconstructed context-specific metabolic models for each cell line. Due to its clinical relevance, we focused on predicting gene essentiality. In-silico predictions were validated using loss-of-function profiles from genome-scale CRISPR assays. Incorrect predictions were used to refine the models.

In order to improve the predictions of classic Flux Balance Analysis (FBA), we extended this approach by including two new features: i) viability of a knockout was evaluated considering the capacity of predicting biomass formation (i.e. classic FBA) and also the energetic capacity of the cell; ii) during the knockout calculation we contextualize the gene-protein-reaction rule using gene expression. This extensions allowed to recover a larger number of essential genes. For instances, many glycolytic genes wrongly predicted as dispensable by the classic FBA were correctly predicted as essential by our extended implementation. Moreover, we developed a reachability analysis that allows to contextualize

the biomass equation. The preliminary results showed that this analysis can yield valuable information on the biomass composition of a given cell type which can be used to improve model formulation.

In conclusion, this study shows that: i) computational modeling can be combined with gene essentiality assays to generate biological hypothesis regarding a tumour metabolic vulnerabilities; ii) the use of ATP production as a predictor of viability increases the number of true positive predictions; iii) context-specific knockout helps to recover new candidate genes; and iv) the combination of wrong predictions with the reachability analysis can be used to to better define the biomass composition of a given cell type.

133

Cecilia Coimbra Klein[1,2,3], Elena Vizcaya-Molina[3], Florenci Serras[3], Roderic Guigó[1,2] and Montserrat Corominas[3]

[1]*Centre for Genomic Regulation, CRG, Spain*
[2]*Universitat Pompeu Fabra, UPF, Spain*
[3]*Departament de Genètica, Microbiologia i Estadística, IBUB, Universitat de Barcelona*

The regulatory genome of drosophila regeneration

Abstract: One of the key questions in regenerative biology is to unveil the regulatory regions capable to trigger tissue recovery. Regeneration is the ability to reconstruct missing parts. The capacity to regenerate varies greatly, not only between species, but also between tissues and organs, as well as from one developmental stage to another in the same species. Drosophila imaginal discs show a high regenerative capacity after genetically induced cell death. We performed genome-wide chromatin landscape analyses (ATAC-Seq and RNA-Seq) at different time points (early, mid and late) of Drosophila imaginal disc regeneration to study the transcriptional programs as well as the regulatory elements responsible for tissue regeneration.

We identified sets of upregulated genes located close to one another in the linear genome (herein called clusters), mostly at early and mid regeneration, indicating that large regions, rather than individual genes, may be controlled by the same regulatory elements. Open chromatin regions that presented higher accessibility in regeneration compared to controls (namely damage-responsive regulatory elements: DRREs), were classified according to their position relative to the closest transcription start site (TSS) of a gene: core promoter (±100 bp of the TSS), in the first intron (FI), proximal (±2 kb from the TSS), and distal (more than ± 2 kb away from the TSS). We distinguished two types of DRREs: emerging, open regions detected only after damage (eDRREs); and increasing, regions already open in control, but displaying increased accessibility after damage (iDRREs). We have also validated several DRREs using enhancer reporter fly lines after inducing apoptosis as well as after physical injury. Since spatial chromatin organization connects active enhancers to target promoters to regulate gene expression, we confirmed individual interactions between eDRREs and clusters of co-regulated genes by Chromatin Conformation Capture analysis. Moreover, DRREs contained conserved binding motifs for transcription factors that are upregulated and required for regenerating organs in fly, zebrafish and mouse.

Our findings indicate there is global co-regulation of gene expression where genes localized in genomic clusters may be regulated by the same elements. Furthermore, we found a regeneration program driven by the cooperation among regulatory elements acting exclusively within damaged tissue, with enhancers co-opted from other tissues and other developmental stages, as well as with endogenous enhancers that show increased activity after injury. Such elements host binding sites for regulatory proteins that include a core set of conserved transcription factors that may control regeneration across metazoans.

134

Angela Del Pozo[1,2], Carlos Rodríguez-Antolín[1], Rubén Martín-Arenas[1], Mario Solís López[1], Beatriz Ruz-Caracuel[1], Álvaro González-Rocafort[1], Gemma García-Cerrato[3], Luis Fernández[1] and Elena Vallespín[1,2]

[1]*INGEMM - Hospital Universitario La Paz Madrid – IdiPaz, Spain*
[2]*CIBERER, Spain*
[3]*Instituto Anatómico Forense, Spain*

Analysis of RNA-Seq data for diagnosis of patients with Congenital Cardiopathies

Abstract: Congenital heart disease (CHD) represents about 1% of live births and encompasses a wide range of malformations. Specifically, the non-syndromic forms of CHD usually present complex genotype-phenotype association due to allelic heterogeneity and high phenotypic overlap. This makes genetic testing a valuable clinical tool to provide information about the etiology of the disease, the pattern of inheritance and the recurrent risk. However, the clinical interpretation of the variants is challenging and a proportion of the studies are inconclusive. Many issues limits the diagnostic; One is due to the incomplete knowledge of the molecular and mutational basis of CHD that hamper the interpretability of data as in many cases the causal mutation affects a gene whose association to phenotype is still unknown. Others, the splicing machinery is altered due to variants in unknown splicing signals and, therefore, difficult to identify with the isolated study of the patient's DNA. So, additional evidences are needed to determine pathogenity is in these inconclusive cases.

The current project aims to explore the utility of transcriptome analysis with RNA-Seq as a complementary diagnostic tool in a cohort of patients with non-syndromic CHD. First, cardiac muscle tissue was obtained from a group of pediatric patients who have been transplanted in cardiac surgery. So, myocardial tissue of the left ventricle has been extracted from the diseased heart to be sequenced in a RNA-Seq experiment. In parallel, tissue of healthy donors (whose cause of death was not associated with cardiovascular disease) was obtained in a collaboration with in Instituto Anatómico Forense. Both sets has been studied and compared in a case-control methodology. This study is based in the hypothesis that abnormal transcripts should be causative of disease. So, novel splicing events that are private to a patient compared with controls should explain the phenotype.

In this work it is presented preliminary results of the analysis of 3 patients. Patient A has a autrioventricular block and a Dilated Miocadiopathy (DMC) and prior of RNA-Seq has been studied in a gene panel being negative in 85 genes associated to CHD. Patient B presents a DMC and phenotype of patient C is unknown. In each patient it has been studied the transcriptome of cardiac tissue and blood before and after the surgery to test blood as a model of CHD with promising results.

As conclusion, the work explores the utility of RNA-Seq to detect causative mutations not only in CHD but other group of diseases, especially Neuromuscular disorders.

135

Sílvia Bonàs-Guarch[1], Marta Guindo-Martínez[1], Irene Miguel-Escalada[2], Niels Grarup[3], David Sebastian[4], Elias Rodriguez-Fos[1], Friman Sánchez[1], Mercè Planas-Fèlix[1], Paula Cortes-Sánchez[1], Santi González[1], Pascal Timshel[3], Tune H. Pers[3], Claire C. Morgan[5], Ignasi Moran[5], Goutham Atla[2], Juan R. González[6], Montserrat Puiggros[1], Jonathan Martí[1], Ehm A. Andersson[3], Carlos Díaz[1], Rosa M. Badia[1], Miriam Udler[7], Aaron Leong[8], Varindepal Kaur[8], Jason Flannick[7], Torben Jørgensen[9], Allan Linneberg[9], Marit E. Jørgensen[10], Daniel R. Witte[11], Cramer Christensen[12], Ivan Brandslund[12], Emil V. Appel[3], Robert A. Scott[13], Jian'an Luan[13], Claudia Langenberg[13], Nicholas J. Wareham[13], Oluf Pedersen[3], Antonio Zorzano[4], Jose C Florez[7], Torben Hansen Hansen[3], Jorge Ferrer[2], Josep Maria Mercader[1] and David Torrents[1]

[1]*Barcelona Supercomputing Center, BSC, Spain*
[2]*Institut d'Investigacions August Pi i Sunyer, IDIBAPS, Spain*
[3]*The Novo Nordisk Foundation Center for Basic Metabolic Research, Denmark*
[4]*Instituto de Salud Carlos III, Spain*
[5]*Imperial College London, United Kingdom*
[6]*ISGlobal, Centre for Research in Environmental Epidemiology, CREAL, Spain*
[7]*Broad Institute of Harvard and MIT, United States*
[8]*Harvard University, United States*
[9]*Research Centre for Prevention and Health, Denmark*
[10]*Steno Diabetes Center, Denmark*
[11]*Aarhus University, Denmark*
[12]*Lillebaelt Hospital, Denmark*
[13]*University of Cambridge School of Clinical Medicine, United Kingdom*

Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes

Abstract: The reanalysis of existing genome-wide association studies (GWAS) data represents a powerful opportunity to gain insights into the genetics of complex diseases. These cost-effective reanalysis strategies are now possible given emerging (i) data-sharing initiatives fed with large amounts of primary genetic data for multiple human genetic diseases, as well as (ii) new and improved GWAS methodologies and resources. Notably, genotype imputation with novel sequence-based reference panels can now substantially increase the genetic resolution of GWAS from previously genotyped data-sets (Huang J. et al. Nat. Commun., 2015). Therefore, we gathered publicly available type 2 diabetes (T2D) GWAS cohorts with European ancestry, comprising a total of 13,857 T2D cases and 62,126 controls, to which we first applied harmonization and quality control protocols. We then performed imputation using 1000 Genomes Project (1000G) (Abecasis G. R. et al. Nature, 2012) and UK10K (Huang J. et al. Nat. Commun., 2015) reference panels followed by association testing (Bonàs-Guarch S. et al. Nat. Commun., 2018). Unlike the vast majority of reported GWAS analyses, we also included the analysis of the X chromosome, which represents 5% of the genome and codes for more than 1,500 genes (Tukiainen T. et al. PLoS Genet., 2014). By using this strategy, we confirmed a large fraction of known T2D loci and we identified six novel associated regions in the autosomes, five driven by common variants (LYPLAL1, NEUROG3, CAMKK2, ABO, and GIP genes) and one by a low-frequency missense variant (EHMT2). Moreover, we also identified one associated region driven by a rare variant in chromosome Xq23, rs146662075, with a two-fold increased risk for T2D in males, which is the association with the largest effect size identified in Europeans to date. Interestingly, rs146662075 is located within an active enhancer associated with the expression of Angiotensin II Receptor type 2 gene (AGTR2), a modulator of insulin sensitivity, and exhibited allelic specific activity in muscle cells. Our observation that 30% of the rs146662075 risk allele carriers developed T2D over 11 years of follow-up, compared to 10% of non-carriers, supported the association of this variant and suggested that an early identification of these subjects through genotyping may be useful for a pharmacological or lifestyle intervention to prevent or delay the onset of T2D. Overall, beyond our contribution towards expanding the number of genetic associations with T2D, our study highlighted the potential of the reanalysis of public data as a complement to large studies that use newly generated data for preventive and therapeutic applications.

136

Mireia Bernabeu-Gimeno[1], José Manuel Martí[1], Wladimiro Diaz-Villanueva[1], Vicente Arnau[1] and Carlos Peña-Garay[1]

[1]*Institute for Integrative Systems Biology, I2SysBio, Spain*

Gollum ecosystem: functional characterization of complex biological systems in extreme environments

Abstract: Gollum[1] is a project of analysis of ecosystems in rock of high depth. The project is coordinated by the Institute of Biology of Systems and the ICTS Underground Laboratory of Canfranc. It takes samples in five different positions of the tunnel of Somport. This tunnel crosses different types of sedimentary rock formed by the accumulation of sediments during the Mesozoic and Cenozoic, which together with its length, depth and ecological diversity make it an ideal place for the ecological study of extremophiles.

We have carried out an exhaustive bioinformatic analysis of the sequenced samples, including data quality controls, and a complete and precise taxonomic classification from several indexed databases. A functional analysis has been carried out to characterize the system dynamics.

6% of the DNA observed corresponds to genera of archaea whose closest taxa are located on sea beds. The functional analysis shows a strong correlation with the metals present in the environment.

[1] Lsc-canfranc.es. (2018). Laboratorio Subterráneo de Canfranc - Memoria Anual. [online] Available at: http://www.lsc-canfranc.es/es/inicio/documentos/memoria-anual.html.

137

Azahara Fuentes[1], Alicia Serrano[2], Blanca Ferrer[2], Verónica Lendínez[1], Carolina Monzó[1], Laura Olivares[1], Carmen Ivorra[3], Maria José Terol[2], Blanca Navarro[2] and F. Javier Chaves[1]

[1]*INCLIVA, Spain*
[2]*Hematology Department HCUV, Spain*
[3]*Sequencing Multiplex, Spain*

Sequencing limitations in b-cell repertoire ngs-data analysis

Abstract: Genetic studies require specific methodologies. Currently, many procedures previously performed by Sanger sequencing are moving to high-throughput technologies, overall to NGS. Many studies are no longer supported by Sanger sequencing because of the increasing needs in regard to sensitivity and automation. B-cell receptor repertoire sequencing always presents limitations in NGS analysis, and require a rigorous design of library preparation and data analysis to fulfill the requirements.

During B cells development, the heavy chain of their surface receptor suffers a recombination process before it is expressed, and only one segment of each of the three regions named Variable, Diversity and Joining (IGHV, IGHD and IGHJ) in the locus is selected to codify the protein, presenting additionally insertions and deletions of random nucleotide sequences of variable length in the junction, known as CDR3 (Complementary Determining Region 3). Subsequently, a hypermutation process occurs mainly on IGHV region, responsible for the high affinity and specificity of the receptor against antigens. In consequence, the repertoire of BcRs is about 10e10. Infections, vaccine studies and some hematological neoplasms such as Chronic Lymphocytic Leukemia require an in depth study of represented B-cell clones, from low to high proportions in the repertoire (2-90%).

The region used up to now for the analysis of these “unique genes” for each B-cell exceed 700bp being an important limitation for implementing these protocols by current NGS systems (second generation). New amplicon design and other strategies are needed to perform Illumina short-read sequencing and data analysis. The variability of the region complicates nested amplification due to the fact that standard primers (Biomed-2) can be only designed in conserved

regions. We have developed a bioinformatics pipeline that covers two approximations: multiplex amplification combining different sets of standard primers and tagmentation.

138

Luis Felipe Arias-Giraldo[1], José Manuel Martí[1], Vicente Arnau[1], Wladimiro Diaz-Villanueva[1] and Carlos Peña-Garay[1]

[1]*Institute for Integrative Systems Biology, I2SysBio, Spain*

Analysis of the temporal evolution of the infection by pathogenic fungis in plants. A new approach to Verticillium wilt of olive tree

Abstract: Most omic studies on pathogen-host interaction in plants focus only on the host and in the infectious agent, considering the last one as the main factor in the infectious process. This approach discards the role of the great variety of microorganisms in an environment as diverse as the soil.

With the development of new software in the field of metagenomics [1], new approaches are possible. In order to take into account all the possible microorganism, we have performed a robust clasification using all the available sequences. Starting from public time series data of RNA-sequences of olive tree infected by Verticillium dahliae [2], we have analized the dynamic of the infection process as a integrated system.

1. Martí, J. M. (2018). Recentrifuge: robust comparative analysis and contamination removal for metagenomic data. bioRxiv, 190934

2. Jiménez-Ruiz, J., Leyva-Pérez, M. D. L. O., Schilirò, E., Barroso, J. B., Bombarely, A., Mueller, L., & Luque, F. (2017). Transcriptomic analysis of L. Roots during the early infection process. The plant genome.

139

Bruna Correa[1], Cecilia Klein[1], Alexandre Esteban[1], Amaya Abad[1], Ramil Nurtdinov[1], Sebastian Ullrich[1], Emilio Palumbo[1], Sílvia Perez-Lluch[1], Rory Johnson[2] and Roderic Guigó[1]

[1]*Centre for Genomic Regulation, CRG, Spain*
[2]*University of Bern, Switzerland*

Dynamics of the pathway from transcription to translation in a transdifferentiation model

Abstract: Current development of massively parallel transcriptomics and proteomics techniques has been providing the throughput and resolution required to understand key regulatory events that compose the pathway that connects transcription to translation. As we know, gene expression is controlled during all steps of this pathway, which includes RNA synthesis, transport, translation and degradation. However, the relative contribution of each regulatory step in different biological contexts remains largely unknown. In order to model the flow of information from RNA to protein and to decode its dynamics, we have used a seven-days model of transdifferentiation of human pre-B cells into macrophages. RNA and protein levels were quantified at different cellular fractions at 12 time points. For each time point, we have analyzed nuclear RNA (Nuc-RNAseq) and cytosolic RNA (Cyto-RNAseq) through RNA sequencing, ribosomal binding to RNA by ribosome profiling (Ribo-Seq), and protein levels through mass spectrometry (Prot-MS). First, we explored the correlation between RNA levels from different cellular fractions and protein levels. As expected, we observed that correlation with protein increases as we go from nuclear RNA, to cytosolic RNA, and finally to ribosome profiling, which showed the highest correlations with protein levels at all time points. Following, by applying a lag penalized weighted correlation technique (LPWC) to pairs of consecutive cellular fractions across time, we were able

to identify groups of "fast" and "slow" genes in terms of their mRNA transport from nucleus to cytoplasm (by comparing Nuc-RNAseq vs. Cyto-RNAseq), binding to ribosomes (Cyto-RNAseq vs. Ribo-Seq), and translation (Ribo-Seq vs. Prot-MS). By performing Gene Ontology analyses, we could also assign different biological processes to these groups of “fast” and “slow” genes. For instance, immune-related processes, such as “leucocyte activation” and “neutrophil degranulation”, were enriched in the groups of “fast” genes. Finally, we are using a probabilistic framework, called PECAplus, capable of deconvoluting kinetic parameters, such as synthesis and degradation rates, for each gene at each step of the process. This framework is also allowing us to identify at which time points in a given cellular fraction the kinetic parameters change significantly. We expect this way we can unveil the dynamics of key layers of gene expression regulation in the context of transdifferentiation.

140

Raquel Garcia-Perez[1], Gloria Mas-Martin[2], Martin Kuhlwilm[1], Meritxell Riera[1], Antoine Blancher[3], Marc Marti-Renom[2], Luciano Di Croce[2], Jose Luis Gómez-Skarmeta[4], Tomas Marques-Bonet[1] and David Juan[1]

[1]*Institute of Evolutionary Biology, UPF-CSIC, Spain*
[2]*Centro Nacional de Análisis Genómico - Centro de Regulación Genómica, CNAG-CRG, Spain*
[3]*Laboratoire d’Immunogénétique moléculaire, Faculté de Médecine Purpan, Université Toulouse 3, France*
[4]*Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide, Spain*

Recent evolution of the epigenetic regulatory landscape in human and other primates

Abstract: Although comparative epigenomics has intensively studied switch on/off changes in regulatory regions associated to evolutionary differences, the detailed evolutionary changes in activity of enhancers and promoters during primate evolution remain mostly unexplored. Here we characterize the evolutionary dynamics of the epigenomic activity of these regulatory elements in the primate lineage. To that end we have comprehensively profiled lymphoblastoid cell lines (LCLs) from human, chimpanzee, gorilla, orangutan and macaque by performing ChIP-seq for five histone marks, ATAC-seq, WGBS and RNA-seq experiments.

Integration of genome-wide epigenomic and gene expression data has allowed us to identify very characteristic genomic and epigenomic conservation patterns associated to strong, weak and poised activities in promoters and enhancers. In particular, strong promoters are highly epigenomically conserved while poised and weak promoters are mostly species specific. Enhancers show an activity-related bimodal behavior with most of them been either highly conserved or species specific. We observed a clear activity-dependent association between genomic and epigenomic conservation.

Moreover, building of gene-specific epigenomic architectures based on integration of genomic annotations and available 3D contact maps showed that gene regulation concentrates a high proportion of epigenomically conserved strong promoters in primates and most of the poised promoters while only a very small proportion of strong, poised or weak enhancers could be associated to gene regulation with a weaker connection with epigenomic conservation. In order to study the dynamics of changes of epigenomic gene regulation in the primate lineage we analyzed the differences in calibrated RNA-seq and ChIP-seq and ATAC-seq signals. By exploring the interplay between these evolutionary patterns, we were able to disentangle the connection of gene expression and epigenomic changes at different levels of genomic resolution. Multivariate regulatory models based on the global epigenomic signals in regulatory architectures explain over 60% of intra and inter-species expression variability in differentially expressed genes in primates, showing that H3K27ac and H3K4me3 in promoters and H3K36me3 in genic enhancers are the most informative histone marks.

Interestingly, we could distinguish that different patterns of expression changes implicate different changes in activity in promoters, enhancers or the whole architecture associated to changes in the binding of different histone marks. For

instance, human-specific up-regulated genes are mostly associated to changes in H3K36me3 in genic enhancers, while down-regulated ones involves a stronger regulatory changes associated to removal of H3K27ac, H3K4me1 and/or H3K36me3 or of increase of H3K27me3 in the whole architecture. We have established an experimental and computational framework revealing activity-dependent evolutionary constraints and what changes in the epigenomic activity of the gene regulatory architectures are at the source of gene expression changes in the primate lineage.

141

Marina Esteban Medina[1], María Peña Chilet[1], Cankut Cubuk[1], Carlos Loucera[1] and Joaquín Dopazo[1]

[1]*Fundación Progreso y Salud, FPS, Spain*

Mechanistic models for drug repositioning in rare diseases

Abstract: Rare diseases' diagnosis and treatment development remain a challenge for the healthcare system. The majority of rare diseases present a lack of effective treatment, due, among other aspects, to the lack of research dedicated to the discovery and development of therapies. Drug repositioning emerges as a quick and effective way to obtain treatments that are already approved. The constant increase in knowledge about mechanisms of disease and action of drugs collected in various repositories, together with the arising of new methodologies under the framework of machine learning, are fostering the development and application of models capable to predict quite accurately different aspects of the cell phenotype in response to different conditions. In this project, we aimed to develop a novel tool for rare disease (RD) drug repositioning through the mechanistic modeling of the pathways implicated in the disease.

142

Adria Alcala[1] and Gabriel Riera Roca[1]

[1]University of Balearic Islands, Spain

AligNet: Alignment of Protein-Protein Interaction Networks

Abstract: A very difficult problem in systems biology is to discover protein- protein interactions as well as their associated functions. The analysis and alignment of protein-protein interaction networks (PPIN), has become a key ingredient to obtain functional orthologs as well as evolutionary conserved pathways and protein complexes. We propose AligNet, a new method and software tool for the pairwise global alignment of PPIN that produces biologically meaningful alignments and more efficient computations than state-of-the-art methods and tools.

143

Pablo Román-Naranjo Varela[1], Maria Del Carmen Moleón González[2], Álvaro Gallego Martínez[1], Dheeraj Bobbili[3], Patrick May[3] and José Antonio López Escámez[1]

[1]*Otology & Neurotology Group, GENYO - Centre for Genomics and Oncological Research, Granada, Spain*
[2]*Department of Otolaryngology, ibs.GRANADA, Hospital Universitario Virgen de las Nieves, Granada, Spain*
[3]*Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg*

Sensorineural hearing loss genes rare variant analysis and burden analysis in familial Meniere disease

Abstract: Introduction: Meniere disease (MD) is a rare inner ear disorder characterized by vertigo, sensorineural hearing loss (SNHL) and tinnitus with a prevalence of 75 cases in 100,000 people. Although most cases are considered sporadic, familial aggregation exists among 10% of MD cases, supporting a genetic background for this disease. Several genes have been described in familial MD such as COCH, DTNA, FAM136A, PRKCB, DPT and SEMA3D genes, showing genetic heterogeneity. Thus, the goal of this study is to identify variants associated with familial MD in SNHL genes.

Materials & Methods: In this study participated 62 familial MD cases, whose DNA was isolated from blood or saliva samples to perform whole-exome sequencing. Variant calling was performed with the Genome Analysis Tool Kit (GATK), following the GATK Best Practices Pipeline. Candidate genes were prioritized based on predicted variant pathogenicity and phenotypic similarity of diseases associated with the genes harboring these variants using PhenIX pipeline (http://compbio.charite.de/PhenIX/). Variants with a minor allele frequency (MAF) >0.001 were discarded for the single rare variant analysis and those with a MAF > 0.05 were excluded from gene burden analysis. Variants were checked in Deafness Variation Database (http://deafnessvariationdatabase.org/). Candidate variants were validated by Sanger sequencing.

Results: We have found and validated 15 novel or ultrarare variants in SNHL genes likely pathogenic in 17 familial MD cases using the single rare variant analysis. Among them, we highlight a rare homozygous variant in USH1C gene (chr11:g.17518327G>A), a rare heterozygous variant in TECTA gene (chr2:g.26684646C>T) in 3 affected cases within one family, a novel heterozygous variant in MYH14 gene (chr18:g.44121737A>G), and a novel heterozygous variant in DIABLO gene (chr12:g.122701379G>C) found in 3 cases in one family. According with the gene burden analysis, there is an accumulation of rare likely pathogenic variants in OTOF, OTOG and LOXHD1.

Conclusion: Forty percent of familial MD cases carry a novel or ultrarare variant in SNHL genes, suggesting that non-syndromic SNHL genes play an important role in familial MD. Besides, FMD patients showed an accumulation of rare variants in OTOF, OTOG and LOXHD1.

Funding: Supported by the Luxembourg National Research Fund INTER/Mobility/17/11772209

144

Rafael Domínguez Acemel[1], José María Santos Pereira[1], Silvia Naranjo[1] and José Luis Gómez Skarmeta[1]

[1]*Centro Andaluz de Biología del Desarrollo, Spain*

Exploring changes in the 3D genome involved in the invertebrate to vertebrate transition using HiChIP

Abstract: It has been shown that the 3D configuration of the genome plays a major role in the regulation of genes in animals. Briefly, many animal genomes are spatially partitioned in units called Topologically Associated Domains (TADs) and interactions between distant enhancers and promoters are only facilitated and productive when both of them belong to the same TAD. Thus, we wondered if changes in the 3D organization of the genome could lead to regulatory novelties that were fixed during the invertebrate to vertebrate transition. In addition, it has been proposed that mechanical stress occurring at TAD boundaries lead to an enhanced activity of the DNA repair machinery and turn these regions in hotspots for genomic rearrangements with potential evolutionary impact

In a previous study we used 4C-seq to compare the 3D topology of the Hox cluster in amphioxus with that of the Hox clusters in zebrafish. We found that the amphioxus Hox cluster is embedded in a single TAD. In vertebrates, however, it has been described that HoxA and HoxD clusters are placed in between two TADs and receive inputs from both of them. This has proven to be essential for the patterning of paired appendages, a vertebrate novelty. Therefore, changes in DNA topology were needed in order to establish the limb regulatory program in vertebrates.

To further investigate the evolution of DNA topology and its importance in evolutionary transitions we are performing HiChIP with the H3K4me3 antibody both in zebrafish and amphioxus embryos. This allowed us to easily identify the interactions of active promoters genome-wide at 5kb resolution. We plan to use this information to spot cases of topological reorganization potentially leading to regulatory novelties.

145

Pau Puigdevall Costa[1], Lucilla Piccari[2], Isabel Blanco[2,3], Joan Albert Barberà[2,3], Dan Geiger[4], Celia Badenas[5], Montserrat Milà[5], Robert Castelo[1] and Irene Madrigal[5]

[1]*Universitat Pompeu Fabra, UPF, Spain*
[2]*Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer, IDIBAPS, Spain*
[3]*CIBERES, Spain*
[4]*Technion Israel Institute of Technology, Israel*
[5]*Hospital Clínic, CIBERER, Spain*

Genetic linkage analysis of heritable pulmonary arterial hypertension in a large pedigree identifies candidate modulators of reduced penetrance

Abstract: Understanding the molecular basis of human inherited disease is a challenging task due to the complexity of mapping the genotype to the phenotype. The study of diseases running in families has shown that the inheritance of certain rare variants causes disease. However, sometimes not all carriers of such pathogenic variants express the clinical symptoms, a phenomenon known as reduced penetrance. We have investigated a potential mechanism of genetic modification for the penetrance of a missense BMPR2 mutation in heritable pulmonary arterial hypertension (HPAH) by applying genetic linkage analysis to a large multiplex family. We have identified a candidate region for a modifier in the distal promoter region of the FIGN gene supported by lung-specific regulatory activity and GWAS risk factors associated with blood pressure. Taken together, these results suggest that common regulatory variants may have an important role in determining the penetrance of pathogenic coding variants.

146

Pau Puigdevall Costa[1] and Robert Castelo[1]

[1]*Universitat Pompeu Fabra, UPF, Spain*

GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor.

Abstract: Genomewide position-specific scores assign a numeric value indicating different levels of conservation, constraint, fitness or mutation tolerance to each position. They are built on heterogenous information, including sequence homology, functional domains and the structural properties conferred by the amino acid residues. These scores are widely used in the context of variant filtering to assess the functionality of different genetic features. However, integrating such scores may be problematic given the lack of a standardization on the storage format and the large size of the files, especially in genomes of higher eukaryotes. To address these shortcomings we have developed

GenomicScores, an R package available at https://bioconductor.org/packages/GenomicScores that provides an efficient storage and access to genomic scores facilitating their integration into genome analysis workflows on top of R and Bioconductor. Its basic functionality is described in the accompanying vignette and in the article available at https://doi.org/10.1093/bioinformatics/bty311.

147

Alvaro Gallego-Martinez[1], Pablo Roman-Naranjo[2], Dheeraj Bobbili[1], Patrick May[1] and Jose Antonio Lopez-Escamez[2]

[1]*Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg*
[2]*Otology & Neurotology Group, GENYO - Centre for Genomics and Oncological Research, Granada, Spain*

Excess of predicted pathogenic rare variants in axonal guiding signaling pathway in sporadic Meniere disease

Abstract: Introduction: Meniere disease (MD) is a rare inner ear disorder characterized by vertigo, sensorineural hearing loss (SNHL) and tinnitus. Its prevalence of 75 cases in 100,000 people, although this prevalence fluctuates from 50/100.000 to 110/100.000 attending to different populations. MD has been described mostly in sporadic cases, being familial cases around 10% of total observed cases. The main suspected causes attributed to Meniere's disease are related to the development of the otic nerve, cell-cell adhesion and cellular matrix development, and leukocyte activity. In this study, we focus on the search for biomarkers in genes that interact in the main signaling pathways of those biological processes. Our objective is to define genes with a higher load of variants in sporadic cases of MD compared with the frequencies of the European and Spanish control population, and to find specific pathway nodes with involved interacting genes.

Materials & Methods: We performed a targeted-gene sequencing panel of 263 genes related with axonal guidance and neuronal development pathways (196), and cell adhesion and leukocyte activity (67). We pooled 870 Spanish sporadic MD cases for a total of 87 pooled samples. Target gene sequencing was performed following Agilent SureSelect protocol. We performed variant calling using CRISP pool-seq tool to manage genotypes for pooled samples and was compared with the default output of Agilent SureCall variant call tool. Gene burden analysis was performed using a selection of variants included in the common-to-rare threshold, were the minimum allele frequency considered was the minimum detected in the sporadic cases (0.001) whilst the maximum value was considered the threshold for common population variants (0.1). We managed gene burden analysis comparing gnomAD global population, gnomAD non-finnish European population and CSVS Spanish population with our sporadic MD cases. We scored genes that are more significant and calculated which pathways nodes are more affected through network analysis.

Results: We divided the output into LoF variant analysis, missense variant analysis, synonymous variant analysis and pathogenic analysis with variants annotated with CADD>20. A significant accumulation of pathogenic predicted variants in genes such as SLIT2, GNA14 and NTN4 was found in sporadic MD cases, suggesting a role for axonal guidance pathway. Synonymous analysis derived in a focus on cell adhesion genes instead of neuronal development genes. However, we did not found genes shared in missense and synonymous analysis. We could not find any significant accumulation of LoF variants in any gene.

Conclusion: We have observed rare missense variant accumulation in key genes on neuronal development and axon guidance pathways in sporadic MD patients.

148

Francesc Montardit-Tarda[1], Najla Ksouri[1], Yolanda Gogorcena[1] and Bruno Contreras-Moreira[1]

[1]*Consejo Superior de Investigaciones Científicas-Aula Dei (CSIC-EEAD), Spain*

Genomic delimitation of proximal promoter regions: three approaches in Prunus persica

Abstract: The increase of genomic data in plants has become a revolution for plant biologists. Nowadays, transcript expression can be investigated thanks to new sequencing technologies such as RNA-Seq. One of the objectives of these approaches is the de novo motif discovery in upstream sequences of co-expressed genes. However, a bottleneck of these methods is to determine the appropriate length of the upstream region to be sampled, something that has only been addressed so far in two plant species, rice and Arabidopsis thaliana. Here, three computational approaches are presented with the aim of delimiting proximal promoter regions at a genomic level and tested in three species: Prunus persica, A. thaliana and Brachypodium distachyon. Two of the approaches relied on nucleotide conservation, one aligning sequences of related species (Prunus sp.), and the other by computing nucleotide polymorphism density. Finally, the third approach was based on the frequency of detected putative regulatory DNA motifs. The comparative genomics approach suggested a limit of -576 nt or -434 nt upstream of transcription start sites for Prunus persica, depending on the source of the genome. The polymorphism density approach yielded proximal limits of -510 nt, -520 nt and -450 nt for P. persica, A. thaliana and B. distachyon, respectively. Finally, the frequency of predicted DNA motifs provided alternative proximal promoter for the three species, and the differences were discussed. A validation of these limits was carried out by calculating the percentage of experimentally determined regulatory motifs in plants inside the proposed ranges. We report that a -500 nt upstream window contains approximately 80% of the motifs curated in TRANSFAC. Therefore, plant proximal promoter regions were delimited with a length of -500 nt, with differences of -50/-100 nt between species and/or methodologies. The availability of the developed code will enable the scientific community to calculate the optimal length of proximal promoter regions for any plant species.

149

Najla Ksouri[1], Francesc Montardit-Tarda[1], Bruno Contreras-Moreira[1] and Yolanda Gogorcena[1]

[1]*Consejo Superior de Investigaciones Científicas-Aula Dei (CSIC-EEAD), Spain*

De Novo regulatory motif discovery in Prunus persica co-regulated genes

Abstract: Peach [Prunus persica (L.) Batsch], a model plant for Rosaceae family, is considered as an interesting species for genomics and computational researches. Considering the advances of experimental technologies, including genome sequencing and gene expression profiling, attention has been shifted towards deciphering the molecular mechanisms underlying gene regulation under various environmental stimuli. Indeed, the modulation of gene regulation is an intricate process occurring at various levels of which the transcriptional signature is the core control code. The transcription machinery is based on a combinatorial cooperation between transcription factors (TFs) and their cognate transcription factor binding sites (TFBSs) known as cis-regulatory elements (CREs) or motifs. By binding these specific non coding motifs upstream the promoters, TFs may act either as activators or repressors of gene expression leading to dynamic changes of the cellular responses. In peach, the transcriptional regulation is mediated by 2780 TFs grouped into 58 families. Whereas much is known about TFs, CREs involved in gene expression regulation remains a challenging task as the number of possible combinations between TFs and their targets is enormous. A widely used approach for motif discovery is co-expressed genes clustering as they are more likely to have their promoters bound by a common TF. In this study, we conducted a genome wide scale identification of regulatory motifs under various environmental stresses (drought, cold, hyper-hydricity,…) and in various tissue types (root, leaf, stigma and fruit). Eight peach RNA-sequencing data sets were used as raw input and gene expression profiling has revealed 11335 stress-related genes. Different clustering algorithms were tested for an accurate definition of the co-regulated genes according to their expression patterns. Promoter sequences for each gene within each cluster were retrieved and randomly generated clusters were used as a negative control to highlight the reliability of the putative (CREs). Two de novo motif finding

algorithms offered by RSAT::Plant tool (oligo and dyads analysis), were assessed to find out the over-represented sites upstream the genes. Out of 45 clusters, 12 were predicted to contain significant regulatory motifs. This work represents thus a great step toward scoping out the fundamental bases of transcriptional regulation in P. persica and offers a great potential for novel motif discovery for any plant.

150

Luis M Escudero[1,2]

[1]*Instituto de Biomedicina de Sevilla, Spain*
[2]*Universidad de Sevilla, Spain*

Scutoids, a geometrical solution to three-dimensional packing of epithelia

Abstract: As animals develop, the initial simple planar epithelia of embryos must be sculpted into complex three-dimensional tissues. However, the architecture and packing of curved epithelia remains largely unknown. Here, by means of computational modelling, we show that cells in bent epithelia are compelled to adopt a novel shape that we name “scutoids”. The detailed image analysis of diverse tissues and organs confirm the generation of apico-basal transitions among cell during morphogenesis. Using biophysics arguments we infer that scutoids allow the minimization of the tissue energy and stabilize the tree-dimensional packing of the tissue. Altogether, we argue that scutoids are nature´s solution to achieve epithelial bending and the missing piece for developing a unifying and realistic model of epithelial architecture. Our results pave the way to understand the biomechanics of morphogenesis in developing organisms and sheds light on the underlying logic of 3D cellular self-organization.

I will present new experiments and the computational tools that we are developing to characterize the molecules that are important for the appearance and maintaining of scutoids.