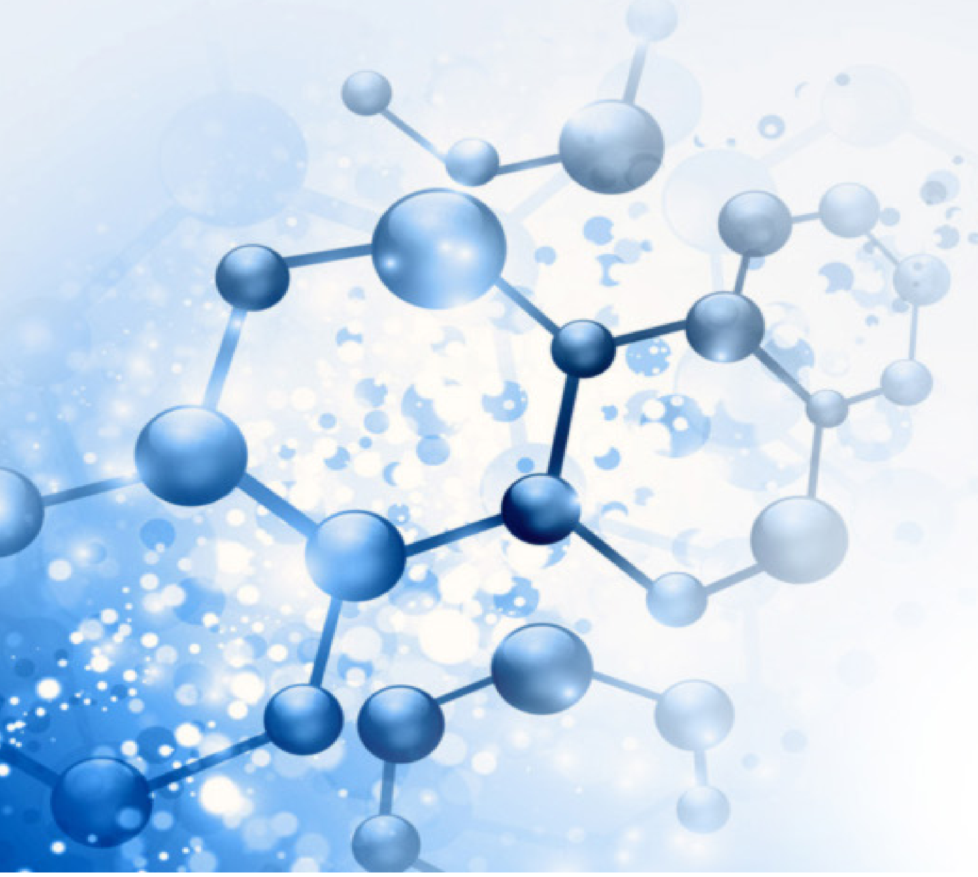
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

JBI2016

XIII Symposium on Bioinformatics

May, 10th -13th 2016



Valencia - Spain

INB Spanish National Bioinformatics Institute

Main sponsor: FUJITSU intel

Sponsors:

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

MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Instituto de Salud Carlos III

PRB²

etsinf Escola Tècnica Superior d'Enginyeria Informàtica

ESCUELA TÉCNICA SUPERIOR INGENIEROS INDUSTRIALES VALENCIA

ITACA

IUMPA Instituto Universitario de Matemática Pura y Aplicada

biobam BIOINFORMATICS SOLUTIONS

SEQUENTIA

elixir

Affiliated Conference:

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Collaborators:

acm UPV Chapter

BIB BIOINFORMATICS

Keynotes.

The Evolution of Enzyme Function.

Janet M. Thornton[1]

[1]*European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK.*

With sequenced genomes, the full complement of enzymes in an organism can be defined. To help assign accurate functions to these new sequences, we need to understand how enzyme families evolve functional diversity. Our objective is to map how enzymes evolve new functions. To do this we developed two software tools to explore the evolution of enzyme function. **EC-BLAST** performs quantitative similarity searches between enzyme reactions. **FunTree** integrates a wide range of data to unravel the evolution of novel enzyme functions. We have applied these methods to analyse a large number of families comprising very diverse relatives.

In order to get an overview of the universe of biochemical reactions we derived a network of the 5,073 representative reactions, connected according to the similarity of their reactions. As a test case, we have explored the ability of EC-BLAST to capture the overall chemistry of the isomerases (EC 5) and to reproduce their EC classification. Our results revealed that isomerase reactions are chemically diverse and difficult to classify automatically from first principles.

An analysis of >300 enzyme superfamilies revealed that almost all enzyme families perform multiple reactions. Most changes of function relate to substrate changes, whilst maintaining the enzyme chemistry. The molecular mechanisms involved in changing function include sequence variations, indels and domain composition changes. Although the catalytic residues are the most conserved of all residues, they do change both their type and position along the sequence in most families.

Our study highlights the complexity of enzymatic catalysis and the need for well-structured, accurate databases of enzyme reactions. There is no simple relationship between changes in enzyme function and sequence during evolution.

JBI2016 Keynote #2

Function and Evolution of Post-Translational Modifications.

Vera van Noort[1]

[1]*Centre of Microbial and Plant Genetics. KU Leuven, Belgium.*

Cells have to continuously adapt their internal systems to the environment. Are sufficient nutrients available for growth; is there another cell that I need to communicate with; am I under attack; questions that often need to be answered quickly in order for an organism to be successful. These fast answers are provided by signaling mechanisms, modifications of proteins that can be reversed and that other proteins can read and respond to or that directly change the protein activity or function. We apply computational methods to study all aspects of these post-translational modifications. These methods can reveal when the modifying enzymes appear in evolution, but also the direct effect of these post-translational modifications on the molecular functions of proteins. The explosion of genomic sequencing has enabled the large-scale computational analysis of protein sequence data that provides the raw material for our evolutionary studies. The technological advances in proteomics have provided a rich source of information on protein post-translational modifications that we exploit to further understand how organisms sense the environment.

JBI2016 Keynote #3

Towards a pan-eukaryotic kinome.

Berend Snel[1]

[1]*Utrecht Bioinformatics Center. Utrecht University, The Netherlands.*

Kinase are hallmark signaling proteins in life. Comparative genomics is needed for many open questions concerning kinases, such as how to assign kinases in newly sequenced eukaryotic genomes to an orthologous group or how to place the ~60 unclassified human kinases into one of the seven main kinase families. However many steps in conventional phylogenetics are pivotally hampered by the massive size of the kinase family across all eukaryotes.

We here present a preliminary analysis of the pan-eukaryotic kinome that relies on a recently proposed computational approach that was developed to solve similar questions for the RAB small GTPases [1]. Although our first results suggest that the resolution for kinases is less clear than for RABs, most human kinases can now be placed next to or inside one of the kinase main families. Subsequently we project the
resulting kinase orthologous groups onto a reference set of eukaryotic species. This projection reveals the fate of the ancestral eukaryotic kinase diversity across extant lineages. Some lineages such as the important model organisms *Saccharomyces cerevisiae* and *Arabidopsis thaliana* harbor relatively few ancestral kinase orthologous groups. In contrast, animals but also ciliates and especially the excavate
Naegleria gruberi are most similar to LECA with respect to their kinases.

[1] Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. Marek Elias, Andrew Brighouse, Carme Gabernet-Castello, Mark C. Field, Joel B. Dacks. J Cell Sci 2012 125: 2500-2508. doi: 10.1242/jcs.101378

Special Session:

Highligths (H).

#018

A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation

Rafael D Acemel[1], Juan J Tena[1], Ibai Irastorza-Azcarate[1], Ferdinand Marlétaz[2], Carlos Gómez-Marín[1], Elisa de la Calle-Mustienes[1], Stéphanie Bertrand[3], Sergio G Diaz[1], Daniel Aldea[3],Jean-Marc Aury[4], Sophie Mangenot[4], Peter W Holland[2], Damien P Devos[1], Ignacio Maeso[1], Hector Escrivá[3], and José Luis Gómez-Skarmeta[1]

[1]*CABD, Universidad Pablo de Olavide. Spain.*
[2]*Department of Zoology, University of Oxford. UK.*
[3]*Université Pierre et Marie Curie Université Paris 6, CNRS, UMR 7232, Biologie Integrative des Organismes Marins (BIOM), Observatoire Océanologique de Banyuls-sur-Mer. France.*
[4]*Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, France.*

The HoxA and HoxD gene clusters of jawed vertebrates are organized into bipartite three-dimensional chromatin structures that separate long-range regulatory inputs coming from the anterior and posterior Hox-neighboring regions. This architecture is instrumental in allowing vertebrate Hox genes to pattern disparate parts of the body, including limbs. Almost nothing is known about how these three-dimensional topologies originated. Here we perform extensive 4C-seq profiling of the Hox cluster in embryos of amphioxus, an invertebrate chordate. We find that, in contrast to the architecture in vertebrates, the amphioxus Hox cluster is organized into a single chromatin interaction domain that includes long-range contacts mostly from the anterior side, bringing distant cis-regulatory elements into contact with Hox genes. We infer that the vertebrate Hox bipartite regulatory system is an evolutionary novelty generated by combining ancient long-range regulatory contacts from DNA in the anterior Hox neighborhood with new regulatory inputs from the posterior side.

Acemel, RD. *et al*. Nature Genetics, 48 (2016), 336-341.

#092

Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs

David Juan[1], Juliane Perner[2], Enrique Carrillo-De Santa Pau[1], Simone Marsili[1], David Ochoa[3], Ho-Ryun Chung[4], Martin Vingron[4], Daniel Rico[1], and Alfonso Valencia[1]

[1]*Spanish National Cancer Research Centre. Spain*
[2]*Cancer Research UK/University of Cambridge. United Kingdom.*
[3]*European Bioinformatics Institute (EMBL-EBI). United Kingdom*
[4]*Max Planck Institut Fuer Molekulare Genetik. Germany*

Epigenetic communication through histone and cytosine modifications is essential for gene regulation and cell identity. Here, we propose a framework that is based on a chromatin communication model to get insight on the function of epigenetic modifications in ESCs [1]. The epigenetic communication network was inferred from genome-wide location data plus extensive manual annotation. Notably, we found that 5-hydroxymethylcytosine (5hmC) is the most-influential hub of this network, connecting DNA demethylation to nucleosome remodeling complexes and to key transcription factors of pluripotency. Moreover, an evolutionary analysis revealed a central role of 5hmC in the co-evolution of chromatin-related proteins. Further analysis of regions where 5hmC co-localizes with specific interactors shows that each interaction points to chromatin remodeling, stemness, differentiation, or metabolism. Our results highlight the importance of cytosine modifications in the epigenetic communication of ESCs.

[1] Juan, D. *et al*. Cell Reports, 14 (2016), 1246–1257.

#098

Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage

Marina Marcet-Houben[1,2], and Toni Gabaldón[1,2,3]

[1]*Bioinformatics and Genomics Programme. Centre for Genomic Regulation (CRG). Dr. Aiguader, 88. 08003 Barcelona, Spain.*
[2]*Universitat Pompeu Fabra (UPF). 08003 Barcelona, Spain.*
[3]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.*

Whole-genome duplications have shaped the genomes of several vertebrate, plant, and fungal lineages. Earlier studies have focused on establishing when these events occurred and on elucidating their functional and evolutionary consequences, but we still lack sufficient understanding of how genome duplications first originated. We used phylogenomics to study the ancient genome duplication occurred in the yeast Saccharomyces cerevisiae lineage and found compelling evidence for the existence of a contemporaneous interspecies hybridization. We propose that the genome doubling was a direct consequence of this hybridization and that it served to provide stability to the recently formed allopolyploid. This scenario provides a mechanism for the origin of this ancient duplication and the lineage that originated from it and brings a new perspective to the interpretation of the origin and consequences of whole-genome duplications.

Marcet-Houben, M. & Gabaldón, T. PLoS Biology, 13 (2015), e1002220.

#090

Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry.

Alexandros Pittis[1,2], and Toni Gabaldón[1,2,3]

[1]*Bioinformatics and Genomics Programme. Centre for Genomic Regulation (CRG). Dr. Aiguader, 88. 08003 Barcelona, Spain.*
[2]*Universitat Pompeu Fabra (UPF). 08003 Barcelona, Spain.*
[3]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.*

The origin of eukaryotes stands as a major conundrum in biology. Current evidence indicates that the last eukaryotic common ancestor already possessed many eukaryotic hallmarks, including a complex subcellular organization. In addition, the lack of evolutionary intermediates challenges the elucidation of the relative order of emergence of eukaryotic traits. Mitochondria are ubiquitous organelles derived from an alphaproteobacterial endosymbiont. Different hypotheses disagree on whether mitochondria were acquired early or late during eukaryogenesis. Similarly, the nature and complexity of the receiving host are debated, with models ranging from a simple prokaryotic host to an already complex proto-eukaryote. Most competing scenarios can be roughly grouped into either mito-early, which consider the driving force of eukaryogenesis to be mitochondrial endosymbiosis into a simple host, or mito-late, which postulate that a significant complexity predated mitochondrial endosymbiosis. Here we provide evidence for late mitochondrial endosymbiosis. We use phylogenomics to directly test whether proto-mitochondrial proteins were acquired earlier or later than other proteins of the last eukaryotic common ancestor. We find that last eukaryotic common ancestor protein families of alphaproteobacterial ancestry and of mitochondrial localization show the shortest phylogenetic distances to their closest prokaryotic relatives, compared with proteins of different prokaryotic origin or cellular localization. Altogether, our results shed new light on a long-standing question and provide compelling support for the late acquisition of mitochondria into a host that already had a proteome of chimaeric phylogenetic origin. We argue that mitochondrial endosymbiosis was one of the ultimate steps in eukaryogenesis and that it provided the definitive selective advantage to mitochondria-bearing eukaryotes over less complex forms.

Pittis, AA. & Gabaldón, T. Nature, 531 (2016), 101-104.

#097

Bacterial antisense RNAs are mainly the product of transcriptional noise

Verónica Lloréns-Rico[1], Jaime Cano[1], Tjerko Kamminga[2], Rosario Gil[3], Amparo Latorre[3], Wei-Hua Chen[4], Peer Bork[4], John I. Glass[5], Luis Serrano[1], and Maria Lluch-Senar[1]

[1]*Centre for Genomic Regulation (CRG). Spain*
[2]*Merck. Netherlands*
[3]*Institut Cavanilles de Biodiversitat i Biologia Evolutiva. Spain*
[4]*EMBL. Germany*
[5]*JCVI. United States*

cis-Encoded antisense RNAs (asRNAs) are widespread along bacterial transcriptomes. However, the role of most of these RNAs remains unknown, and there is an ongoing discussion as to what extent these transcripts are the result of transcriptional noise. We show, by comparative transcriptomics of 20 bacterial species and one chloroplast, that the number of asRNAs is exponentially dependent on the genomic AT content and that expression of asRNA at low levels exerts little impact in terms of energy consumption. A transcription model simulating mRNA and asRNA production indicates that the asRNA regulatory effect is only observed above certain expression thresholds, substantially higher than physiological transcript levels. These predictions were verified experimentally by overexpressing nine different asRNAs in Mycoplasma pneumoniae. Our results suggest that most of the antisense transcripts found in bacteria are the consequence of transcriptional noise, arising at spurious promoters throughout the genome.

Lloréns-Rico, V. *et al*. Science Advances, 2 (2016), e1501363.

#094

The potential clinical impact of the release of two drafts of the human proteome

Iakes Ezkurdia[1], Enrique Calvo[1], Angela Del Pozo[2], Jesús Vázquez[3], Alfonso Valencia[4], and Michael Tress[4]

[1]*Centro Nacional de Investigaciones Cardiovasculares. Spain*
[2]*IGEMM - Hospital La Paz Madrid, Spain*
[3]*CNIC. Spain*
[4]*Spanish National Cancer Research Centre. Spain*

The authors have carried out an investigation of the two “draft maps of the human proteome” published in 2014 in Nature. The findings include an abundance of poor spectra, low-scoring peptide-spectrum matches and incorrectly identified proteins in both these studies, highlighting clear issues with the application of false discovery rates. This noise means that the claims made by the two papers – the identification of high numbers of protein coding genes, the detection of novel coding regions and the draft tissue maps themselves – should be treated with considerable caution. The authors recommend that clinicians and researchers do not use the unfiltered data from these studies. Despite this these studies will inspire further investigation into tissue-based proteomics. As long as this future work has proper quality controls, it could help produce a consensus map of the human proteome and improve our understanding of the processes that underlie health and disease.

Ezkurdia, I. *et al*. Expert Rev. Proteomics 12 (2015), 579–593.

Special Session:

Student Symposium (I).

#017

cICB: a modular high-throughput computational pipeline for the annotation of proteins of unknown function by Integrative Cell Biology

Nicola Bordin[1], Juan Carlos González Sánchez[2], and Damien Devos[1]

[1]*Centro Andaluz de Biología del Desarollo. Spain*
[2]*University of Heidelberg. Germany*

During last decade, the gap between sequence determination and functional annotation has increased dramatically, representing an incomplete understanding of the data we have generated. Automatic annotation pipelines ease the burden of manual annotation, but are limited in scope and coverage. Computational tools for proteome annotation are intrinsically conservative in assigning a definitive function (76% of proteins in UniProt/TrEMBL are annotated as “unknown” or “uncharacterized”) and tend to focus on specific aspects of the protein such as functional domains, signal peptide prediction, or the presence of transmembrane helices, among others. Compartmentalizing the annotation gives a very specific characterization of an aspect of the protein, at the cost of losing the general overview of the protein's function and role in its environment. Integrating results from several databases and tools allows us to simultaneously question several related aspects of a protein's function. We have created a computational pipeline that combines the advantages of manual curation with the speed and power of bioinformatics. The pipeline is modular, open, can be installed at your location or run on our server. The pipeline allows the characterization of whole proteomes as well as single proteins. We applied the Integrative Cell Biology (ICB) pipeline to 40 bacterial proteomes belonging to the PVC superphylum and were able to increase the average number of annotated proteins from 54% to 78%. We illustrate some of the advantages of ICB with some global and detailed results. The system and results will be available at www.pvcbacteria.org.

#031

Integration of multiomics data to describe link between developmental exposure to pesticides and impaired neurodevelopment

Elena Bernabeu[1], Marta Llansola[1], Vicente Felipo[1], Sonia Tarazona[1], and Ana Conesa[1]

[1]*Centro de Investigación Príncipe Felipe. Spain*

In recent years, the omics disciplines have made their way across a wider spectrum of research groups, thus leading to the generation of multiomics data sets in a great number of studies. While traditional omics studies only focused on a single biological level, the multiomics approach has the potential of studying systems in further detail. However, along with this great potential comes the challenge of integrating and analyzing data far more complex in nature than that of a single omic discipline.

One of such multiomics data integration challenges is exemplified by the EU-funded DENAMIC project, which investigated neurotoxic effects of low-concentration mixtures of pesticides and a number of common environmental pollutants in children using a Rattus norvegicus animal model. The goal of this project was to assess risk of environmental pollutants for human developmental neurotoxicity, as well as to develop biomarkers associated with exposure to several low-concentration mixtures of pesticides. To meet this goal, several molecular levels were measured in rats treated with 4 different pesticides, using different omics platforms: proteomics, metabolomics and transcriptomics. In this work we aim to develop a strategy for the integration of multiomics and clinical data (in this case, learning tests) for the DENAMIC experimental set up. Our objective is to obtain multiomic models regarding the neural response to toxic compounds as well as further information about the global effect of pesticide developmental exposure at both molecular and physiological levels.

The strategy presented here aims to tackle different challenges of integrative analysis. First, the pre-processing of the multiomic data and the treatment of missing values, and second, the establishment of potential associations between mRNAs, miRNAs, proteins and metabolites, while trying to filter out spurious associations to increase the mapping specificity. These first steps are essential to model the interactions among omics and test the significance of such interactions with statistical methods. Multivariate approaches and regression models will be applied for such purpose as well as to relate omic data with the variables from learning tests. Finally, a proper graphical representation of these results will allow for an easy identification of relevant pathways and components that could potentially act as markers of neurotoxicity and explain the molecular basis of impaired neurodevelopment.

#055

New applications for computational docking: modeling of protein-peptide complexes and oligomeric structures

Chiara Pallara[1], and Juan Fernández-Recio[1]

[1]*Life Sciences Department, Barcelona Supercomputing Center, Spain*

After sequencing the complete genomes of several organisms, one of the current biological challenges consists in unravelling their intricate protein-protein interaction networks. To this aim, one of the fundamental steps is to provide structural details at atomic level for such interactomes, which is essential to understand biological processes at molecular level and contribute to improve therapeutic intervention. Nevertheless, given the high costs and the intrinsic limitations of available experimental methods, in-silico modeling based on computational docking represents a valuable tool to complement such data and thus improve our understanding of biological processes involving protein interactions.

In this context, CAPRI experiment (Critical Assessment of PRediction of Interactions) [1] provides a common ground for testing the predictive capability of currently available docking methods, like our docking scoring algorithm, pyDock [2], but also for identifying their limitations and finally guiding new developments in the field. In the last CAPRI edition, new thought-provoking challenges had to be faced, such as the modeling of protein-peptide complexes and the prediction of homo-multimers and domain-domain interactions as part of the first joint CASP-CAPRI experiment [3].

Given the transient and weak nature of protein-peptide complexes, as well as the high conformational flexibility of peptides, new docking strategies had to be devised to model such interactions. interestingly, the use of a new template-based approach proved to be more successful than the integration of peptide conformational sampling (based on molecular dynamics and energy minimization protocols) into our standard ab initio docking strategy.

Moreover, specific new protocols were also developed for the structural modeling of homo- and hetero-dimers, and homo-tetramers. Overall, they showed general good success in the prediction of homo and hetero-complexes but failed in modeling any tetrameric structures.

All in all, this CAPRI edition showed that protein-protein docking protocols can be usefully used for protein-peptide and homo-oligomer structural prediction.

[1] J. Janin et al. Proteins 52 (2003) 2-9.
[2] TM. Cheng et al. Proteins 68 (2007) 503-515.
[3] MF. Lensink MF et al. Proteins (accepted)

#069

When in development? Estimating natural selection through Drosophila life cycle using population genomics data

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

[1]*Genomics, Bioinformatics and Evolution. Departament de Genètica i Microbiologia, Universitat Autònoma de Barcelona, Spain*
[2]*EvoDevo Helsinki community, Centre of Excellence in Experimental and Computational Developmental Biology, Institute of Biotechnology, University of Helsinki, Finland*

The hourglass model of embryonic evolution claims that early and late development are more divergent between species than mid-development. At the molecular level this model has been tested calculating the ratio of non-synonymous to synonymous substitution per site (dN/dS) among species for genes expressed in different developmental stages. Although a low ratio indicates constraint, the interpretation of high values as adaptive selection is equivocal. Here we use a novel approach, the DFE-alpha method [1], a robust derivative of the McDonald and Kreitman test. This method combines genome data on polymorphism within a species (Drosophila melanogaster data from DGRP [2]), and divergence between two species (D. melanogaster and Drosophila yakuba), to estimate the fraction of non-synonymous fixations occurred from the separation of both species that are actually adaptive (α). We calculate α for the genes expressed in each developmental stage and through the life cycle using gene expression data are from RNA-seq experiments (modENCODE database [3]). for each stage, genomic determinants such as codon usage bias, messenger complexity, intron length, expression bias or number of transcripts and exons have been estimated.

We find that all pupa and adult male stages exhibit the highest levels of adaptive change while mid and late embryonic stages show high conservation. The earliest stages of embryonic development, especially the first two hours, are the least conserved but, in contrast to the pupa and adult stages, this relative lack of conservation is partially explained by relaxation of natural selection. The stages exhibiting the highest conservation, mid and late embryonic development, are the ones showing the most complex gene structure: they express, on average, larger genes, with more exons, more transcripts and longer introns. The contrary holds for the stages with high adaptation rates. We conclude that: (i) adaptation along the D. melanogaster life-cycle occurs in two discrete and widely separated periods; (ii) the pattern of adaptive substitutions mirrors those of genomic determinants; and (iii) gene structure complexity associated to each developmental stage seems the proximate explanation for the observed hourglass pattern.

[1] Eyre-Walker et al. Mol. Biol. Evol. 26 (2009) 2097–2108.
[2] Mackay et al. Nature 482 (2012) 173–178.
[3] Graveley et al. Nature 471 (2011) 473–479.

Scientific Session:

Structural Bioinformatics & Function (F).

#011

Structural diseasomes and hot-spot prediction enable detection of network-attacking mutations

Didier Barradas-Bautista[1], and Juan Fernandez-Recio[1]

[1]*Barcelona Supercomputing Center. Spain*

Next generation sequencing projects have demonstrated that mutations like the non-synonymous single nucleotide polymorphisms(nsSNPs) are responsible for population diversity and how individual are affected. Protein-protein interactions(PPIs) are involved in almost all essential cellular processes, and may be affected by nsSNPs causing a pathology[1],[2]. However, in spite of their importance, there is no available 3D structure for the vast majority of known PPIs[3]. Computational methods, such as protein-protein docking, can complement existing experimental efforts and help to build the human structural interactome[4]. The correct prediction of protein complexes by docking is still very challenging for many cases. However, the identification of hot spot interface residues, based on sequence conservation or physicochemical properties, is more accurate and can be applied at more large scale. When characterizing PPI interfaces, it would be significant to identify hot-spot residues, which are those that contribute significantly to the binding energy[5]. A method developed in our group, called pyDockNIP[6] can identify interface hot-spots with high precision, and has the distinct advantage of not needing prior information of the structure complex. We have developed and validated a variation of this method called ""pyDockNIP extended” that can be applied to identify pathological mutations on the binding surface thus altering the PPIs. Our method has 40% recall with 75% precision to detect such mutations.We constructed PPI individual networks with all the 3D protein structures available for the main elements of RASopathies network and six monogenic inheritable diseases. We merged a simple pathway analysis and our predictive method on phenotypes such as Colorectal cancer or Myocardial infarction. We found 292 nsSNPS associated with diseases that are located at protein interfaces and probably drive the network to pathological states. An example, we found 20 nsSNPs at the interface of proteins, that alter a subnetwork of key proteins in the Ras signaling cascade with no apparent effect on the binding energy, but affecting metabolic pathways similar to cancer and contribute to the development of the pathology.

[1] X. Wang, X. Wei, B. Thijssen, J. Das, S. M. Lipkin, and H. Yu, “Three-dimensional reconstruction of protein networks provides insight into human genetic disease,” Nat Biotechnol, 2012, vol. 30, no. 2, pp. 159 – 164.

[2] R. Mosca, J. Tenorio-Laranga, R. Olivella, V. Alcalde, A. Céol, M. Soler-López, and P. Aloy, “dSysMap: exploring the edgetic role of disease mutations,” Nat. Methods, 2015, vol. 12, no. 3, pp. 167–168, Feb.

[3] R. Mosca, A. Céol, and P. Aloy, “Interactome3D: adding structural details to protein networks,” Nat. Methods, vol. 10, no. 1, pp. 47–53, Jan. 2013.
[4] R. Mosca, C. Pons, J. Fernández-Recio, and P. Aloy, “Pushing Structural Information into the Yeast Interactome by High-Throughput Protein Docking Experiments,” PLoS Comput. Biol., 2009, vol. 5, no. 8, p. e1000490, Aug.
[5] O. Keskin, B. Ma, and R. Nussinov, “Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot-spot residues,” J Mol Biol, 2005, vol. 345, pp. 1281 – 1294.
[6] S. Grosdidier and J. Fernández-Recio, “Identification of hot-spot residues in protein-protein interactions by computational docking,” BMC Bioinformatics, 2008, vol. 9, no. 1, p. 447.

#028

Rationally designed drug blending as a mechanism to overcome drug resistance in cancer

Francisco Martinez-Jimenez[1], John Overington[2], Bissan Al-Lazikani[3], and Marc A. Marti-Renom[1]

[1] *CNAG-CRG. Spain*
[2] *Stratified Medical. United Kingdom*
[3] *Institute of Cancer Research, Sutton. United Kingdom*

Drug resistance is one of the major problems in cancer treatment. Rapid mutation and selective pressure can efficiently select drug-resistant mutants. Although there are many mechanisms for drug resistance, a classic mechanism is due to coding mutations in the drug-target binding site. Numerous efforts have been made to individually understand and overcome resistance to imatinib in chronic myelogenous leukemia (CML) treatment caused by the T315I mutation of ABL1 or resistance to gefitinib in non-small-cell lung cancers (NSCLC) due to point mutations in epidermal growth factor receptor (EGFR). However, there is a lack systematic analysis of the mutational landscape that can potentially cause resistance to targeted therapies.

To address this issue, we have developed a computational model that predicts mutations in a drug target with the highest likelihood-resistance ratio in a specific cancer class. Subsequently, our model finds the least sensitive compounds for these resistance mutations. Finally, it defines a blend of molecules potentially overcoming resistance caused by generation of spontaneous mutations in the target in a particular cancer type. We exemplified the applicability of the framework using the ERK1, ERK2, MEK1 and EGFR kinases. Our results show overlapping with previously reported resistant mutations in these proteins. We also provide a set of candidate molecules potentially insensitive to these mutations. In summary, our work aims to reduce the difficulties in the choice of the optimal treatment and thus, it is a step further in the development of the personalized medicine for the treatment of cancer.

#034

Machine learning-based method for residue-residue contact prediction.

Ruben Sanchez-Garcia[1], Joan Segura[1], Jesus Cuenca-Alba[1], Carlos Oscar S. Sorzano[1], and Jose Maria Carazo[1]

[1]*CNB/CSIC. Spain*

Protein-Protein Interactions (PPI) are critical for almost all cellular processes. Although much computational research have been done in this topic, including tools to predict whether or not two proteins may interact or to model the three-dimensional structure of a PPI, just a few works have focused on contact prediction at the residue-residue level. Residue-residue contact predictions, which may be regarded as an in silico alternative to Mass Spectrometry Cross-Linking experiments, present several applications. Thus, for instance, they can be employed as distance constrains for PPI docking or Electron Microscopy, simplifying both model construction and validation. State-of-the-art techniques aimed to predict residue-residue contacts includes correlated mutations-based methods [3] and machine learning-based approaches [1,2].

In this work we present a new machine learning-based method developed to predict residue-residue contacts of two proteins using as input either their sequences or their PDB files. Our approach is based on a two-step-random-forest classifier that employs sequence based features such as position specific score matrices and structural features such as accessible surface area, protrusion index or atomic depth. In the first step we compute a residue-residue contact score using a classifier that has been trained over a data set where each residue pair has been codified employing sequence features of the residues and their sequence neighbors. In the second step, each residue pair is codified using the first step scores combined with structural features of the residues and their structure neighbors. When PDB input files are available, their structural features and neighbor residues are computed directly. If there are not available protein structures, structural features are predicted from sequence.

Our method has been trained over the complexes compiled in Docking Benchmark 4.0 [4]. Evaluation, carried out by leave one complex out cross-validation, shows promising results since we have achieved state-of-the-art global performance levels using a limited number of features, which suggests that there is still room for improvement.

[1] S. Ahmad & K. Mizuguchi PLoS ONE 6(12) (2011) e29104.
[2] Fu. Minhas et al. Proteins 82(7) (2014) 1142-55.
[3] J. Iserte et al. Nucleic Acids Research 43(Web Server issue) (2015) W320-W325.
[4] H. Hwang et al. Proteins 15;78(15) (2010) 3111-14.

#056

Characterization and quantification of the global transcriptome in the Oligodendrocyte differentiation by means of PacBio and Illumina sequencing

Ana Conesa[1,2], Lorena de La Fuente Lorente[1], Hector Del Risco[2], Cristina Martí[1], Victoria Moreno[3], Susana Rodríguez[4], and Manuel Tardaguila[2]

[1]Centro de Investigación Príncipe Felipe, Genomics of Gene Expression, Valencia, Spain
[2]Institute for Food and Agricultural Sciences, Department of Microbiology and Cell Science, University of Florida, Gainesville, USA
[3]Centro de Investigación Príncipe Felipe, Gene Expression and RNA Metabolism, Valencia, Spain
[4]Centro de Investigación Príncipe Felipe, Neuronal and Tissue Regeneration, Valencia, Spain

Alternative splicing, a widespread means of creating functional diversity in higher eukaryotes, entails substantial challenges for its bioinformatic analysis. Paramount among these is the elaboration of the transcriptome to analyse, specially given the high similarity rate between isoforms and the incompleteness and/or variation in the annotation of the 5' and 3' ends of the mRNA. Here we have applied both PacBio (long reads) and Illumina (short reads) sequencing in a murine model of neural stem cell differentiation. PacBio sequencing detects whole transcripts (mean length resolved is 3000 bp) and is ideal to elaborate precise transcriptomes and perform isoform discovery. The tradeoff is the high error rate (around 5%) and the loss of quantification power. Complementarily, Illumina allows for quantification of expression and for the correction of error-prone long reads. Classification of our PacBio transcriptome based on the splice pattern of isoforms reveals 60% of transcripts match annotated references in Refseq and ENSEMBL, 35% show novel splice junctions (18% with alternative donors or acceptors) and 5% map to regions thought to be deprived of coding potential (genic introns and intergenic regions). Further characterization of the novel splice junctions involved the evaluation of their coverage using both short reads from the experiment and from external repositories as well as an in silico proteomic validation using large databases of mass spectrometry profiles. The results show the enrichment of novel splice junctions in UTRs relative to the distribution of known splice sites as well as their short read coverage validation. Furthermore, the use of a restricted transcriptome instead of a global reference, diminishes the amount of quantification artifacts. Lastly, important expression associations can be made from this data: we found that most of multi-isoform genes expressed at least one additional annotated isoform at greater levels, in the majority of the cases it being the so called Principal Isoform, while a reduced subset of genes only expressed the novel isoform. Altogether these results shed light into the complex dynamics of alternative splicing and points to the necessity of using restricted transcriptomes to adequately analyze gene expression at the isoform level.

#048

Nucleosome Dynamics portal

Ricard Illa[1], Laia Codó[2], Romina Royo[2], Adam Hospital[3], Isabelle Heath[4], Josep Lluís Gelpí[5], and Modesto Orozco[5]

[1]*Institute for Research in Biomedicine (IRB) - Molecular Modeling and Bioinformatics Group. Spain*
[2]*National Institute of Bioinformatics (INB), Barcelona Supercomputing Center (BSC). Spain*
[3]*National Institute of Bioinformatics (INB), Institute for Research in Biomedicine (IRB) - Molecular Modeling and Bioinformatics Group. Spain*
[4]*Institute for Reseach in Biomedicine (IRB) - Structural and Computational Biology Unit. Spain*
[5]*National Institute of Bioinformatics (INB), Institute for Research in Biomedicine (IRB) - Molecular Modeling and Bioinformatics Group, Barcelona Supercomputing Center (BSC), Universitat de Barcelona (UB). Spain*

Nucleosome positioning plays a major role in transcriptional regulation and most DNA-related processes. Here we present Nucleosome Dynamics, a new online tool for analyzing nucleosome positioning from MNase-seq experimental data. The tool works on cell/time averaged studies, as well as on comparative experiments to account for the transient and dynamics nature of nucleosome positioning under different cellular states.

Two R libraries, nucleR and NucleosomeDynamics, were specifically developed to perform such studies. nucleR performs a Fourier transform filtering and a peak calling, in order to efficiently and accurately define and classify nucleosome's location. NucleosomeDynamics compares different MNase-seq experiments, and analyze their variations, thus, identifying variations in the nucleosome location. It identifies upstream and downstream shifts, evictions, inclusions, and differences on chromatin digestion. Additionally, a list of other nucleosome related features, like the location of nucleosome-free regions, the theoretical prediction of nucleosomes periodicity at gene level, the classification of transcription start sites based on the nucleosomes' properties surrounding them, and the elastic properties of the nucleosomes derived from a fitting of nucleosome profiles into a Gaussian function, are also computed.

The calculations are made accessible in a web portal [1]. The interface allows to upload data on the server, select studies to be executed, launch calculations, and finally store and maintain the results in a private user workspace. Results can be downloaded, as GFFs files or BIGWIG, or visualized. The executions run asynchronously in parallel on the server, while the workspace allows to monitor and keep track of their state. for the visualization of results, we use JBrowse, a fast and embeddable genome browser built completely with JavaScript and HTML5. In the present version, the JBrowse is adapted to yeast data, and incorporates relevant annotations from the Saccharomyces Genome Database (SGD), data from several recent publications in the field, and can also incorporate users's own annotation tracks.

The Nucleosome Dynamics portal provides a single access point to a complete series of nucleosome positioning oriented tools, and contributes to a multiscale view of chromatin structure.

[1] http://mmb.irbbarcelona.org/NucleosomeDynamics

#093

Transcriptional atlas of Drosophila melanogaster imaginal discs

Cecilia Coimbra Klein[1], Alessandra Breschi[1], Silvia Perez-Lluch[1], Marina Ruiz-Romero[1], Amaya Abad[1], Emilio Palumbo[1], and Roderic Guigó[1]

[1]*Centre for Genomic Regulation (CRG) and Pompeu Fabra University. Spain*

The fruitfly is among the most studied model organisms, however the exploration of specific tissues and compartments along different developmental stages is restricted to few transcriptomics and epigenetics studies. Here, we aim to trace the spatial and temporal transcriptional profile from the eye, leg, genitalia and the wing imaginal discs of Drosophila melanogaster in three developmental stages: third instar larva, white pupa and late pupa. We further aim to explore the signatures of specific compartments of the wing imaginal disc.

We identified sets of commonly expressed genes, tissue and compartment-specific genes as well as time-specific genes. GO analysis of tissue and compartment-specific genes reveals an enrichment for development and pattern formation-related terms while time-specific genes are enriched in cell cycle and broader developmental terms. A comparison of the splicing patterns shows that there are fewer differences in splicing when compared to gene expression. Nevertheless, such differences in isoform usage are mainly found in late pupa stage, suggesting that splicing may play a role during differentiation.

Posters

#012

Archer: Predicting protein function using local structural features. A helpful tool for protein redesign.

Jaume Bonet[1], Javier Garcia-Garcia[1], Joan Planas-Iglesias[1], Narcis Fernandez-Fuentes[1,2], and Baldo Oliva[1]

[1]Structural Bioinformatics Lab (GRIB-IMIM). Department of Experimental and Life Sciences. Universitat Pompeu Fabra, Barcelona, Catalonia (Spain).
[2]Institute of Biological, Environmental and Rural Sciences (IBERS) Aberystwyth University Aberystwyth, Ceredigion, UK.

The advance of high-throughput sequencing methodologies has led to an exponential increase of new protein sequences, a large proportion of which remain unannotated. The gap between the number of known proteins and those with assigned function is increasing. In light of this situation, computational methods to predict the function of proteins have become a valid and necessary strategy. Here we present Archer, a server that exploits ArchDB's [1] hierarchy of super-secondary structures to map GO [2] and Enzyme [3] functions upon protein regions and, thus, infer the function of a protein. The server relies on either the sequence or structure of the protein of interest and returns the mapping of functional subclasses extracted from ArchDB. Moreover, it computes the functional enrichment and significance of each subclass, combines the functional descriptors and predicts the function of the query-protein. We compared our results with sequence-based annotation methods, such as Best-BLAST [4], BLAST2GO [5] and BLAST2GO+InterPro [6], showing a high accuracy (around 80%) even for the comparison of sequences with low similarity. Furthermore, users can select variants of the target sequence that swap the region of a super-secondary structure by another that putatively fits in the same scaffold. Only variants that modify the predicted function are offered for selection, thus providing a rational, knowledge-based, approach for protein design and functionalization. The Archer server is accessible at http://sbi.imim.es/archer.

1. Bonet J et al. Nucleic Acids Res 42(Database issue) (2014) D315-319.
2. Harris MA et al Nucleic Acids Res 32(Database issue) (2004) D258-261.
3. Bairoch A: Nucleic Acids Res28(1) (2000) 304-305.
4. Jones CE et al. BMC Bioinformatics 6 (2005) 272.
5. Gotz et al. Nucleic Acids Res 36(10) (2008) 3420-3435.
6. Hunter S et al Nucleic Acids Res 40(Database issue) (2012) D306-312.

#013

iFraG: a protein-protein interface prediction server based on sequence fragments

Javier Garcia-Garcia[1], Victoria Valls-Comamala[2], David Andreu[3], Francisco J. Muñoz[2], Narcis Fernandez-Fuentes[4], and Baldo Oliva[1]

[1]Structural Bioinformatics Laboratory, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, C/Dr. Aiguader 88, 08003 Barcelona, Spain
[2]Laboratory of Molecular Physiology and Channelopathies, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[3]Laboratory of Proteomics and Protein Chemistry,Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[4]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University. Gogerddan Campus, SY23 3EB, Aberystwyth, United Kingdom

Protein-protein interactions (PPIs) are crucial in many biological processes. The first step toward the molecular characterization of PPIs implies charting their interfaces, i.e. surface(s) mediating the interaction. To this end, here we present iFrag, a sequence-based web server that infers possible interacting regions between two proteins by searching minimal common sequence-fragments of the interacting protein pairs. By utilizing the sequences of two interacting proteins (queries), iFrag derives a two dimensional residue matrix computing a score for each pair of residues that relates to the presence of similar regions in interolog protein pairs. The scoring matrix is represented as a heat-map reflecting potential interface regions in both query proteins. We have studied the performance of iFrag predictions on a non-redundant dataset (less than 40% of sequence identity) of proteins extracted from Uniprot database[1] that are found in a complex structure in the PDB[2] filtered to Human. We have compared iFrag with: PIPE-Sites [3], PPIPP[4], SLIDER [5], and DOMINE [6] using high confidence predictions between domains of PFAM[7]. iFrag is competitive in terms of values of area under the ROC curve, precision and MCC to those of current approaches (70%, 0.1%, and 0,01 respectively). The predicted regions involved in the interaction can range from short fragments composed by few residues to complete domains or proteins, depending on available information on PPIs in each specific case; some examples are presented in the supplementary material. Moreover, as a proof of concept we have tested with iFrag the potential interaction of the amyloid beta peptide (Aβ with length 42 Aas) with clusterin (APOJ or CLUS_HUMAN) and serum albumin, in order to predict potential regions of the interface. Among the potential fragments, we have selected the C-tail peptide of albumin (Alb C-tail) for experimental validation. The results of iFrag suggests the potential of the Alb C-tail fragment to bind the amyloid beta peptide and hinder its aggregation. We have confirmed this prediction experimentally with solubilized Aβ40 peptide [8]. iFrag is freely accessible at http://sbi.upf.edu/iFrag

1. UniProt Consortium Nucleic Acids Res 43(Database issue) (2015) D204-212.
2. Velankar et al Nucleic Acids Res 44(D1) (2016) D385-395.
3. Amos-Binks et al. BMC Bioinformatics 12 (2011) 225.
4. Ahmad et al. PLoS One 6(12) (2011) e29104.
5. Boyen et al. IEEE/ACM Trans Comput Biol Bioinform 8(5) (2011) 1344-1357.
6. Yellaboina et al. Nucleic Acids Res 39(Database issue) (2011) D730-735.
7. Finn et al Nucleic Acids Res 44(D1) (2016) D279-285.
8. Bitan & Teplow Methods Mol Biol 299 (2005) 3-9.

Applying docking tools to estimate residue contribution to binding energy

Miguel Romero-Durana[1], and Juan Fernández-Recio[1]

[1]*Barcelona Supercomputing Center. Spain*

Protein-protein docking is one of the multiple computational techniques that has been developed over the last decades. It aims to predict the 3D structure of a complex, given the 3D structures of its individual components. Interestingly, several studies indicate that, besides its main goal, protein-protein docking tools can be valuable in other areas of structural proteomics. for instance, Grosdidier and Fernandez-Recio showed how the normalized interface propensity (NIP) parameter, obtained from rigid-body docking simulations, can be used to identify hot-spots, crucial residues for complex stability, that confer most of its binding energy [1].

In this work we follow the idea of applying protein-protein docking tools to study the energetic implications involved in complex formation by employing pyDock [2] scoring function to qualitatively estimate the residue contribution to the binding energy of complexes. This is remarkable since pyDock scoring function has been specifically designed to address the docking problem. Nevertheless, we show how pyDock docking energy, at the residue level, is able to identify those residues that contribute the most to the binding energy. We have exploited this capability to perform a detailed study of the MEK1-BRAF complex (PDB id 4MNE) in order to predict key residues whose mutation may have pathological consequences. Our study suggests that pyDock outcome is similar to those obtained with other methods like in silico Alanine scanning based on molecular dynamics simulations with AMBER14 package [3] or FoldX [4]), at a much lower computational cost. Given the promising results, we plan to extend the study in order to identify key residue-interacting-pairs that may result instrumental in complex formation.

[1] Solene Grosdidier and Juan Fernandez-Recio. Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinformatics, 9(1):447, 2008.

[2] Tammy Man-Kuang Cheng, Tom L Blundell, and Juan Fernandez-Recio. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins, 68(2):503–515, August 2007. PMID: 17444519.

[3] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. The Amber biomolecular simulation programs. Journal of Computational Chemistry, 26(16):1668–1688, December 2005.

[4] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. The FoldX web server: an online force field. Nucleic acids research, 33(Web Server issue):W382–388, July 2005. PMID: 15980494.

#041

PMut2: a web-based tool for predicting pathological mutations on proteins

Víctor López Ferrando[1], Xavier de La Cruz[2], Modesto Orozco[3], and Josep Lluís Gelpí[1]

[1]*Barcelona Supercomputing Center. Spain*
[2]*Vall D'Hebron Research Institute. Spain*
[3]*Institute for Research In Biomedicine. Spain*

Assessing the impact of amino acid mutations in human health is an important challenge in biomedical research. As sequencing technologies are more available, and more individual genomes become accessible, the number of identified variants has dramatically increased. PMut, released back in 2005 (1), has been one of the popular predictors in this field. PMut was a neural-network-based classifier using sequence data to provide a pathology score for point mutation in proteins.

PMut2 is a new, revised, and much more powerful version of the predictor. It introduces the use of state-of-the-art machine learning algorithms and an updated training set based on SwissVar. It achieves an accuracy of 80% with sensitivity and specificity of 0.79 and 0.75 respectively. PMut2 includes a fully featured training and validation engine that can be optimized to generate predictors adapted to user specific training sets. The engine is implemented in Python using MongoDB engine for data management. It has been adapted to run at the HPC level to cover large scale annotation projects.

#044

Long non-coding RNAs and genome structural organisation

Irene Farabella[1], and Marc Antoni Marti Renom[1,2]

[1]CNAG-CRG, Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain Gene Regulation, Stem Cells and Cancer Program, Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain.
[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

RNAs, and in particular nuclear-retained long noncoding RNAs (lncRNAs), are an organizing factor for shaping nuclear dynamic and architecture [1]. It is known that lncRNAs are an important structural component of DNA-free nuclear bodies as well as key regulators of important biological process mediating chromatin remodeling [1,2]. for example, it has been shown that certain type of lncRNAs, named repressive chromatin-associated lncRNAs, can act as scaffolds via RNA–DNA triplex formation regulating the activity and the localisation of chromatin-associated proteins and modulating gene expression [2]. Another class of lncRNAs, named 'chromatin-interlinking' RNAs, has been shown to act as architectural factors maintaining interphase chromatin conformation particular in actively transcribed chromatin compartment [3]. However, the exact mechanisms on how lncRNAs interact with chromatin and modulate the overall organisation of the chromatin structure are not fully understood and further investigations on the role of lncRNAs in the genomes spatial organisation is needed. To that end, we implemented a computational method combining detection of triplex forming oligonucleotides (via Triplexator [4]) with 3D co-localisation analysis using genome-wide chromosome conformation capture.

[1] J. Rinn and M. Guttman, Science 345 (2014) 1240–1241
[2] T. Mondal et al. Nat Commun. 6 (2015) 7743
[3] M. Caudron-Hergeret et al. Nucleus 2(5) (2011) 410-24
[4] F. A. Buske et al. Genome Res. 22(7) (2012) 1372-81

#086

GREENC: a database of plant lncRNAs

Andreu Paytuví[1], Antonio Hermoso[2,3], Irantzu Anzar[1], Walter Sanseverino[1], and Riccardo Aiese Cigliano[1]

[1]Sequentia Biotech SL, Calle Comte D'Urgell 240, 08036 Barcelona, Spain
[2]CRG Bioinformatics Facility, Centre for Genomic Regulation (CRG), Dr Aiguader 88, 08003 Barcelona, Spain
[3]Universitat Pompeu Fabra (UPF), Dr Aiguader 88, 08003 Barcelona, Spain

Long non-coding RNAs (lncRNAs) are functional non-translated molecules greater than 200 nt. Their roles are diverse and they are usually involved in transcriptional regulation. LncRNAs still remain largely uninvestigated in plants with few exceptions. Experimentally validated plant lncRNAs have been shown to regulate important agronomic traits such as phosphate starvation response, flowering time and interaction with symbiotic organisms, making them of great interest in plant biology and in breeding. There is still a lack of lncRNAs in most sequenced plant species, and in those where they have been annotated, different methods have been used, so making the lncRNAs less useful in comparisons within and between species. We developed a pipeline to annotate lncRNAs and applied it to 37 plant species and six algae, resulting in the annotation of more than 120 000 lncRNAs. To facilitate the study of lncRNAs for the plant research community, the information gathered is organised in the Green Non-Coding Database (GreeNC, http://greenc.sciencedesigners.com/).

#106

Using Formal Language Theory to characterize biosequences.

Jose M. Sempere[1]

[1]*Departamento de Sistemas Informáticos y Computación. Universidad Politecnica de Valencia. Spain*

Formal Language Theory has been a good framework to build bioinformatic tools, especially for motif prediction tasks. for example, D. Searls [1] proposed formal grammars to characterize DNA and RNA biomolecules. The purpose of this work is to describe some motif prediction tools that we have developed by taking the formal language approach.

In this work, we describe tools that characterize the following motifs: (1) RNA/single stranded DNA hairpin structures, (2) nucleotide mutations (pairwise alignments), (3) RNA pseudoknots, and (4) DNA recombination through splicing sites. for the first task, we use non-regular languages based on the iterated and bounded hairpin languages [2], for the second task, we use classical operations over strings [3], for the third task we use stochastic context-free grammars [4] and Watson-Crick finite automata [5] and, finally, for the fourth task we use finite automata [6].

We discuss the (dis)advantages of specific (stochastic) language parsers, and its corresponding computational complexity. We establish some guidelines for a general strategy to apply formal languages in the development of good motif prediction tools. We propose grammatical inference as a good approach for the training of this tools, and we conclude by introducing BIOLANG a formal language approach to characterize biosequences, which actually is work in progress.

[1] D. Searls. The language of genes. Nature vol 420 (2002) 211-217.
[2] D. Cheptea et al. A new operation on words suggested by DNA biochemistry: hairpin completion, In: Proc. Transgressive Computing (2006) 216-228.
[3] M. Crochemore et al. Algorithms on strings. (2007)
[4] E. Rivas et al. The language of RNA; a formal grammar that includes pseudoknots. Bioinformatics vol 16 No 4 (2000) 334-340.
[5] G. Paun et al. DNA Computing. New Computing Paradigms. (1998)
[6] F. Wang at al. Recognition of Simple Splicing Systems using SH-Automaton. Journal of Fundamental Sciences vol 4 No 2 (2008) 337-342

Scientific Session:

Metagenomics (B)

NGS Technologies: Genomics & Transcriptomics (D).

#019

Designing a sensitive software tool for methylation analysis on long-reads datasets

Joaquín Tárraga[1], Mariano Pérez[2], Juan M. Orduña[2], José Duato[3], Ignacio Medina[1], and Joaquín Dopazo[1]

[1]Department of Computational Genomics, Centro de Investigación Príncipe Felipe. Spain.
[2]Departamento de Informática, Universidad de Valencia. Spain.
[3]DISCA, Universidad Politécnica de Valencia. Spain.

DNA methylation is an important mechanism of epigenetic regulation in development and disease. It is a heritable modifiable chemical process that affects gene transcription, and it is associated with other molecular markers (e.g. gene expression) and phenotypes (e.g. cancer or other diseases) (Jones, 13). Although many methods for DNA methylation profiling have been developed, only bisulfite sequencing gives rise to comprehensive DNA methylation maps at single-base pair resolution (Laird, 2010). On other hand, current NGS sequencers can sequence short DNA or RNA fragments of lengths usually between 50 and 400 nucleotides (nts), though new sequencers with longer fragment sizes are being developed. Primary data produced by NGS sequencers consists of hundreds of millions or even billions of short DNA fragments which are called reads. This big data trend has shifted the pressure from the sequencers to the software analysis tools (Fonseca, 2012), which should be scalable enough to process increasing volumes of methylation data with acceptable sensitivity and reasonable execution times. In this paper, we present a software tool for mapping and determining the methylation state of bisulfite reads. It includes an innovative strategy that combines an efficient algorithm for those reads with a low rate of mutation errors, insertions or deletions (EIDs), and an algorithm with notably improved sensitivity that correctly aligns reads with a high rate of EIDs. This software shows excellent sensitivity and remarkable parallel performance for both short and long bisulfite reads, presenting runtimes that linearly depend on the number and the length of reads.

Fonseca, N. e. (2012). Tools for mapping high-throughput sequencing data. Bioinformatics, 3169–3177.
Jones, P. (2013). Functions of DNA methylation: islands, start sites, gene. Nat. Rev. Genet., 484–492.
Laird, P. (2010). Principles and challenges of genome-wide DNA methylation analysis. Nat. Rev. Genet., 191–203.

#027

Health and disease imprinted in the time variability of the human microbiome

Andrés Moya[1,2,3], José Manuel Martí[4], Daniel Martínez[1], Manuel Peña[4], César Gracia[4], Amparo Latorre[1,2,3], and Carlos P. Garay[4]

[1]*Área de Genómica y Salud, FISABIO, Salud Pública, Valencia.*
[2]*ICBiBE, Universitat de València, Paterna (Valencia).*
[3]*CIBER en Epidemiología y Salud Pública (CIBEResp), Madrid.*
[4]*Instituto de Física Corpuscular, CSIC-UVEG, Valencia.*

Human microbiota plays an important role in determining changes from health to disease. Increasing research activity is dedicated to understand its diversity and variability. We analyze 16S rRNA and whole genome sequencing (WGS) data from the gut microbiota of 97 individuals monitored in time. Temporal fluctuations in the microbiome reveal significant differences due to factors that affect the microbiota such as dietary changes, antibiotic intake, early gut development or disease. Here we show that a fluctuation scaling law describes the temporal variability of the system and that a noise-induced phase transition is central in the route to disease. The universal law distinguishes healthy from sick microbiota and quantitatively characterizes the path in the phase space, which opens up its potential clinical use and, more generally, other technological applications where microbiota plays an important role.

#076

Integrated solution to metagenomics, from taxonomy to functional analysis

David Pérez-Villarroya[1], Llúcia Martínez-Priego[1], Alejandro Artacho[2], and Giuseppe D'Auria[1,2,3]

[1]*Sequencing and Bioinformatics Service, FISABIO-Public Health, Valencia, Spain*
[2]*Department of Genomics and Health, FISABIO-Public Health, Valencia, Spain*
[3]*CIBER en Epidemiología y Salud Pública (CIBEResp), Madrid, Spain*

Next Generation Sequencing technologies have revolutionized biology not only in practical, but also in conceptual and formal terms. Moreover, the great amount of genomic data obtained has generated the need for new and more powerful computational, statistical and mathematical techniques. Metagenomics is one of the fields powered by the advent of NGS allowing the access to genomic diversity of natural population independently to the bacterial identity. These methods are based on massive parallel sequencing as alternatives to standard one-by-one cloning strategy.

With respect to first generation sequencing methodologies, NGS provided a deep change in the data throughput scale which is accompanied by new, fast and more robust data analysis approaches.

We developed an integrated framework for metagenomic studies going through a continuous stream from NGS data quality assessment to taxonomical and functional overviews. Whole datasets are split by kinds (ribosomal rRNA, ORFs, tRNAs, other) and each dataset is driven to proper pipeline offering finally a global picture of 16S taxonomy distributions trough descriptive statistics and functional annotation by several databases with following pathway analysis. Other kinds of genomic signatures are tabulated and stored for downstream uses. Moreover computation needs have been optimized trough massive parallel distribution of work.

#067

Searching for the chromatin determinants of hematopoiesis

Enrique Carrillo-De Santa Pau[1], David Juan[1], Felipe Were[1], Vera Pancaldi[1], Daniel Rico[1], and Alfonso Valencia[1]

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

As part of the BLUEPRINT Consortium, we are characterizing the epigenomes of blood cells to understand how changes in chromatin are connected with the different lineage differentiation options. In this work, we present our analyses using hematopoietic stem cells (HSCs), monocytes, macrophages, neutrophils, B-cells (naive from venous blood and tonsil-derived germinal center B-cells) and T-cells (CD4 and CD8), combining hematopoietic samples from BLUEPRINT, ENCODE and NIH Epigenomic Roadmap. We have developed a bioinformatics pipeline to generate a 'chromatin space' where the different cell types are clustered by epigenomic similarity.

Our analysis is based on Multiple Correspondence Analysis (MCA), the analog of Principal Component Analysis when working with categorical data. We used our previous approach to deal with protein multiple alignments (Rausell, Juan et al PNAS, 2010) with critical enhancements to deal with millions of regions in the same analysis.

The analysis of the orthogonal dimension of the space allows us to identify chromatin determinant regions (CDRs), genomic regions with different epigenomic characteristics between the different groups. Functional enrichment analysis of the neighbouring genes suggests that the chromatin state in this regions could be directly linked with the different cell identities. Our analytical approach allows to combine samples from different sources and identify the regions for which chromatin status associates with cell lineage determination or disease conditions.

FAIR, Functional Analysis At Isoform Resolution by using long reads technologies

Lorena de La Fuente Lorente[1], Ana Conesa[1,2], Manuel Tardaguila[2], Hector Del Risco[2], Cristina Marti[1], Victoria Moreno[3], Susana Rodriguez[4], Marissa Macchietto[5], and Ali Mortazavi[5]

[1]*Centro de Investigación Principe Felipe, Genomics of Gene Expression, Valencia, Spain*
[2]*Institute for Food and Agricultural Sciences, Department of Microbiology and Cell Science, University of Florida, Gainesville, USA*
[3]*Centro de Investigación Principe Felipe, Gene Expression and RNA Metabolism, Valencia, Spain*
[4]*Centro de Investigación Principe Felipe, Neuronal and Tissue Regeneration, Valencia, Spain*
[5]*University of California, Irvine, USA*

Transcriptomes of higher eukaryotes are characterized by the presence of multiple isoforms coded by the same gene. Although a dense structural catalog of isoforms and splice junctions has been created for many organisms, the study of the functional implications of alternative isoform expression (AIE) has not been yet addressed at a whole-genome level. However, individual studies have shown the importance of alternative splicing on the creation of isoforms that module the functionality of the cell and are even associated with disease states. Thus, based on the claimed role of alternative spliced isoforms in conferring functional meaning, we have developed a new methodology called FAIR to address functional profiling of transcript and protein isoforms at a genome-wide level by using long-reads technologies (third generation sequencing) such as PacBio. Moreover, we have implemented the FAIR methodology in an easy-to-use software application, Transcrip2GO.

Therefore, using as input PacBio and Illumina data, FAIR can be used to generate functional hypothesis about the role of alternative isoforms in our system. In the first part of the approach, FAIR allows the functional annotation of each PacBio-resolved isoform. It involves the prediction of ORFs and the annotation of a rich diversity of functional layers at isoform resolution: miRNA binding sites, PFAM domains, post-translational modifications, UTR motifs, binding elements, NMD prediction, Provean scores, GO terms, repetitive elements, etc. Finally, as FAIR aim is to figure out the relevance of AIE in the functionality of the system, it applies different statistical methods which combine both expression data and functional annotation over PacBio-resolved isoforms. Among the several included statistical methods, we can highlight the differential splicing analysis to test for genes with differential splicing pattern between conditions and the functional nested enrichment which point out functional elements affected by alternative isoform.

To test this new approach, we applied FAIR methodology in a mouse cell differentiation system from Neural Stem Cells to Oligodendrocytes. Using our rich functional annotation pipeline we found that nearly all genes expressing several isoforms have them annotated with at least one differential functional label, suggesting that functional profiling at isoform resolution is meaningful. We identified up to 100 enriched functions in genes regulated by differential splicing. Combining PAR-CLIP and isoform functional annotations, we also identified the cellular function specifically regulated by distinct RNA binding domains. Other functional insights of the relationship between function and differential splicing are easily revealed by the set of features implemented in the Transcript2GO software.

#108

A cheap and simple method for sample tracking for illumina TruSightOne clinical Exomes

Cristian Perez-Garcia[1], Merche Bermejo[1], Maria Garcia-Hoyos[1], Javier Garcia-Planells[1], Carlos Ruiz-Lafora[1], and Pablo Marin-Garcia[1]

[1]*Instituto de Medicina Genómica. Spain*

In high throughput clinical genetics environments it is crucial to keep track of samples and have methods to double check that results from the lab are correct and belongs to the original sample. There are such methods available commercially but are expensive and based in genotyping methodology so extra processing is needed. We have developed a new method for double checking sample integrity during the NGS pipeline. We have created a plasmid that contains the sequences for a region of a distal 3'UTR captured by the TSO probes that is not used in clinical reports and added synthetic unique barcodes in the middle of the sequence. Plasmids are added to the DNA before starting the NGS pipeline and the unique part of the sequences of the plasmids captured by the TSO sequencing allows us to obtain the ID of the sample and check the identity. We have created a software and test data for automatic deployment of this procedure in any exome capture sequencing facility.

Posters: Metagenomics (B).

#020

MDPbiome: Predicting temporal microbiome dynamics influenced by external perturbations

Beatriz García-Jiménez[1], and Mark Wilkinson[1]

[1] *Biological Informatics Group. Center for Plant Biotechnology and Genomics (CBGP) UPM-INIA. Spain*

Motivation: Most microbiome analyses consider the population to be static. We examine microbiomes as dynamic systems, and attempt to model, and predict, their response to interventions, with the goal of empowering microbiome engineering plans.

Methods: This study examines longitudinal metagenomics data modelled as a Markov Decision Process (MDP), with a new approach we have called MDPbiome. Given an external perturbation, the MDP predicts the next microbiome state in a temporal sequence, selected from a finite set of possible microbiome states, determined by sample clustering. MDPbiome innovates with the inclusion of actions, which has not been reported in any of the prior studies that discuss state transitions diagrams in microbiome analysis.

Results: We examined four distinct datasets to demonstrate this approach. An MDP created for a vaginal microbiome time series generates a variety of behaviour policies. for example, that moving from a state associated with bacterial vaginosis to a healthier one, requires avoiding perturbations such as lubricant, sex toys, tampons and anal sex. The flexibility of our proposal is verified after we applied MDPbiome to human (adult and premature baby) gut and chick gut microbiomes, taking nutritional intakes, breast milk and antibiotic uses, or salmonella and probiotic treatments, respectively, as perturbations. In the preterm infant guts, with a higher percentage of breast milk, the microbiome is more stable, with fewer transitions between different states. In the latter case, MDPs provided a quantitative explanation for why salmonella vaccine accelerates microbiome maturation in chicks. This novel analytical approach has applications in, for example, medicine where the MDP could suggest the sequence of perturbations (e.g. clinical interventions) to apply to follow the best path from any given starting state, to a desired (healthy) state, avoiding strongly negative states.

Further work: MDPbiome's predictive models should now be applied to de novo datasets, in collaboration with wet laboratories. Optimally, we would collaborate at the experimental design stage, where we would have the opportunity to assist in defining the nature and frequency of the interventions, and the meta-data collected at each time-point, to help further our joint discovery goals.

#024

Effect of rifampicin on the intestinal microbiota of Blattella germanica

Tania Rosas[1], Carlos García-Ferris[1], Rebeca Domínguez[1], Pablo Llop[2], Andrés Moya[1,2], and Amparo Latorre[1,2]

[1] *ICBiBE, Universitat de Valéncia, Paterna (Valéncia).*
[2] *Área de Genómica y Salud, FISABIO-Salud Pública, Valencia.*

Most insect species live in symbiosis with microorganisms that have great impact in the insect nutrition, reproduction and survival [1]. The cockroach *Blattella germanica* harbors an obligate endosymbiont, *Blattabacterium*, which plays an important role in nitrogen metabolism, and a complex gut microbiota that varies in composition depending of the insect diet and developmental stage [2]. To study the dynamics of the gut microbiota acquisition and colonization, we analyzed the bacterial composition and diversity in the gut of cockroaches fed with control diet versus rifampicin supplemented diet using 16S rRNA gene amplicons sequencing. We also studied the offspring of these cockroaches, this time fed with control diet, rifampicin supplemented diet or feces supplemented diet to analyze the patterns of bacterial prevalence and recolonization. Using the QIIME pipeline we identified drastic changes in the microbiota composition and a lowered diversity caused by the antibiotic treatment. Interestingly the normal microbiota recolonization was achieved just by removing the antibiotic from the diet as well as by the feces supplementation. Despite the changes in bacterial composition, the genera *Desulfovibrio* prevailed under all conditions, suggesting an important role in the cockroach gut.

[1] Engel & Moran, FEMS Microbiol Rev 37 (2013) 699-735
[2] Perez-Cobas et al. FEMS Microbiol Ecol 91:4 (2015)

#026

Evaluating fragmentation and coverage variation in viral metagenome assemblies using different assemblers

Rodrigo Garcia-Lopez[1,2,3,†], Jorge Francisco Vázquez-Castellanos[1,2,3,†], and Andres Moya[1,2,3]

[1]*Área de Genómica y Salud, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO)-Salud Pública, Valencia, Spain.*
[2]*Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València, Paterna, Spain.*
[3]*CIBER en Epidemiología y Salud Pública (CIBEResp), Madrid, Spain.*
[†]*These authors have contributed equally to this work*

Sequence assembling greatly influences downstream taxonomic classification and functional annotation of viral metagenomes and is mainly challenged by uneven species abundance, stochasticity in sample processing and sequencing errors, impact the success of accurate assemblies [1]. Diversity may be studied using Operational Taxonomic Units (OTUs), but their reliability in viral metagenomics is limited as de novo assembling clusters reads into larger contigs that are independent of reference genomes and must deal with chimeric structures and intra-species variability. Coverage calculations help determine relative completeness of viral genomes in a dataset but they overlook missing or overly separated fragments.

In this work we evaluated seven different assemblers using simulated Illumina datasets. We compared OTU assignment and alpha diversity calculation using a fragmentation score, the coverage and other assembly statistics. Fragmentation was dependent on genome coverage but not as heavily influenced by the assembler. Sequencing depth was the predominant factor in assembling success, whereas the number of contigs, coverage and fragmentation were the most influential alpha diversity estimations. RayMeta and CLC built the most accurate contigs with large datasets while Meta-IDBA dealt better with medium-sized ones. However, assembling highly fragmented genomes with high coverage may lead to the clustering of different OTUs belonging to the same genome.

[1] F. Vázquez-Castellanos et al. BMC Genomics 15:31 (2014) 13

#029

Temporal stability of the salivary microbiome and its association with salivary markers of oxidative stress

Mária Džunková[1,2], Daniel Martínez[1], Nuria Jimenez[1,2], Giuseppe D'auria[1,2], Katarína Janšáková[3], Michal Behuliak[4], Roman Gardlik[3], Peter Celec[4], Amparo Latorre[1,2], and Andrés Moya[1,2]

[1]*Department of Genomics and Health,FISABIO-Public Health, Valencia, Spain - Cavanilles Institute of Biodiversity and Evolutionary Biology, University of Valencia, Valencia, Spain*
[2]*CIBER en Epidemiología y Salud Pública (CIBEResp), Madrid, Spain*
[3]*Institute of Molecular Biomedicine, Faculty of Medicine, Comenius University, Bratislava, Slovakia*
[4]*Institute of Physiology, Academy of Sciences of the Czech Republic, Prague, Czech Republic*

Salivary markers of oxidative stress are associated with periodontal status and are altered in patients with periodontitis or caries. It has been hypothesized that the composition of oral microbiome and the presence of specific bacterial species might be associated with some salivary markers of oxidative stress. The day to day variability of salivary markers of oxidative stress has been described. More detailed studies are needed for the assessment of salivary microbiome stability. The objective of the present study is to study temporal stability of saliva microbiome and to analyze its association to variations of salivary markers of oxidative stress.

Saliva samples from 26 young healthy volunteers (13 women, 13 men) were collected every morning for 30 days. Advanced oxidation protein products as markers of protein oxidation and ferric reducing ability of saliva as well as total antioxidant capacity were measured. The 16S rDNA amplicons have been sequenced on Illumina MiSeq platform and clustered into operational taxonomic units (OTUs). Reference 16S rDNA sequences of all *Streptococcus* species have been incorporated into the clustering to distinguish between OTUs belonging to *Streptococcus* species previously associated with caries (such as *S. mutans*). The environmental factors (volunteers) have been tested for fitting on ordination of samples obtained by canonical correspondence analysis. Intra-individual temporal stability was calculated as variance of prevalence of each OTUs in the time and expressed as a final indicator for the whole microbiome stability. Spearman correlations of proportion of all individual OTUs with oxidative stress markers have been visualized in Bayesian networks.

The ordination of samples showed that microbiomes of 26 volunteers significantly differed between each other ($p<0.001$). The two most important OTUs *Streptococcus parasanguinis* and *Rothia* formed 37.51 ± 10.14% and 20.91 ± 11.31% of the microbiome, respectively, and have been found in all volunteers, but it different proportions. The major differences between individuals were caused by different proportions of less prevalent bacteria (1-5%), such as

Granulicatella, *Actinomyces*, *Atopobium*, *Saccharibacteria* and *Gemella*. The remaining 18,138 detected OTUs had an average proportion below 1%. The analysis on intra-individual temporal variation showed that the microbiome of the most individuals is relatively stable, but the variability detected in male volunteers was higher than in females. Direct correlation between *Streptococcus mutans* (formed less than 1% of the microbiome on average) and oxidative stress markers was not found, however, each volunteer had its own specific bacterial community which correlated with these markers. The results of this study indicate that shifts of the whole microbial community, rather than the presence of single species are associated with variations of oxidative stress markers in saliva.

#050

Distinct gut microbiota composition in gay men

Marc Noguera-Julian[1], Muntsa Rocafort[1], Yolanda Guillén[1], Mariona Parera[1], Piotr Nowak[2], Falk Hildebrand[3], Georg Zeller[3], Anders Sönnerborg[2], Peer Bork[3], Roger Paredes[1], and The Metahiv Study Group[1]

[1]*IrsiCaixa AIDS Research Institute, Catalonia, Spain*
[2]*Department of Medicine, Unit of Infectious Diseases, Karolinska Institutet, Sweden*
[3]*Structural and Computational Biology, European Molecular Biology Laboratory, Germany.*

Background. The precise effects of HIV-1 on the human microbiome are unclear. Initial cross-sectional studies provided contradictory associations between microbial richness and HIV status and suggested shifts from Bacteroides to Prevotella predominance following HIV-1 infection, which have not been found in animal models or in studies matched for HIV-1 risk groups.

Methods. This was a cross-sectional study[1] where we first tested 129 HIV-1-infected subjects and 27 HIV-negative controls in Barcelona (BCN0). Findings were internally validated in 110 subjects from BCN0 providing a second fecal sample 1 month later (BCN1). External validation was obtained in 77 HIV-1-infected and 7 non-infected subjects from Stockholm (STK). In all study participants, we produced MiSeqTM 16S rRNA sequence data on fecal microbiomes and collected comprehensive metadata. Alpha and beta diversity analyses of the gut microbiota were performed. LASSO regression was used to quantify the strength of the association between sexual practice, HIV-1 status and global fecal microbiota composition.

Results. Men who have sex with men (MSM) consistently had a significantly richer and more diverse fecal microbiota than non-MSM individuals. After stratifying for sexual practice, HIV-1 infection remained consistently associated with reduced bacterial richness. The lowest microbial richness was observed in HIV-1-infected individuals with immune-virological discordant phenotype. Fecal microbiomes strongly clustered by sexual practice rather than by HIV-1 serostatus, with high concordance between BCN0 and BCN1 (Procrustes m2=0.3475, PROTEST p=0.001) The fecal microbiota composition in BCN and STK significantly differed by sexual practice, with MSM and non-MSM subjects mostly belonging to the Prevotella and Bacteroides enterotypes, respectively. Cross-validation accuracy of the LASSO model was very high for sexual practice (mean AUC=95%), confirming a different fecal microbiota composition in MSM and non-MSM individuals after excluding multiple other potential confounders. In contrast, HIV-1 status was not associated with consistent changes in the global fecal microbiota composition at the genus level.

Conclusions. Gay men have a distinct gut microbiota composition, which is a potential confounder of all human fecal microbiome studies. Yet, HIV-1 infection remains independently

associated with reduced bacterial richness, which offers new avenues for interventions to improve HIV immune dysfunction.

[1]Noguera-Julian M, Rocafort M. et al Gut Microbiota Linked to Sexual Preference and HIV Infection. EBioMedicine 2016.

#102

Comparative analysis of stress response systems in the human gut microbiome

Elizabeth Hobbs[1], and Ivan Erill[1]

[1]*University of Maryland Baltimore County. United States*

Metagenomic projects provide a unique window into the genetic composition of microbial communities. Next-generation sequencing metagenomics enables the analysis of bacterial population composition and the study of emergent population features, such as shared metabolic pathways. Recently, we have shown that metagenomics datasets can be leveraged to characterize population-wide transcriptional regulatory networks, or meta-regulons, providing insights into how bacterial populations respond collectively to specific triggers. Here we formalize a Bayesian inference framework to analyze the composition of transcriptional regulatory networks in metagenomes and we apply it to the comparative analysis of different stress response systems in the extensive human gut microbiome metagenomic dataset compiled by the Integrated Reference Catalog of the Human Gut Microbiome. Our results reveal key differences across bacterial clades and human populations in the span and composition of bacterial stress responses involved in antibiotic and heavy-metal resistance. The ability to perform comparative analysis on metagenomics-inferred regulatory systems can be leveraged to customize antibiotic treatment and defines a general framework to study the evolution of bacterial regulatory networks using metagenomic data.

Computational workflow to RNA-seq differential expression analysis.

Llúcia Martínez-Priego[1], David Pérez-Villarroya[1], and Giuseppe D'Auria[1,2,3]

[1]*Sequencing and Bioinformatics Service, FISABIO-Public Health, Valencia, Spain*
[2]*Department of Genomics and Health, FISABIO-Public Health, Valencia, Spain*
[3]*CIBER en Epidemiología y Salud Pública (CIBEResp), Madrid, Spain*

With the advent of massive parallel sequencing and the increasing development of bioinformatics tools, RNA sequencing (RNA-seq) give us the opportunity to get a complete picture of the transcriptome of a given organism.

Read mapping, transcriptome assembly, transcript annotation and expression quantification are the key process for RNA-seq analysis. Several bioinformatic tools had been created to carry out all of these steps.

Here we describe a complete and integrative pipeline built using standard data formats and time-effectively computation needed to infer biological conclusions from raw sequences data. We developed a complete RNA-seq analysis pipeline from raw sequencing reads quality assessment to gene-set differential expression analysis and visualization. Sample datasets are aligned to reference genome(s) retrieving new alternative splicing events and detecting polymorphisms for each transcriptome. Alignment is used to infer transcriptome assembly providing a basis to calculate gene expression level and testing the statistical significance of differences between groups by provided metadata. Finally, transcript annotation, differential expressed functional groups (KEGG, GO) and transcriptional or post-transcriptional regulation levels between transcripts are summarized in a report for easy visualization and/or publication.

Posters: NGS Technologies: Genomics & Transcriptomics (D).

#057

Different approaches and automatic decision to obtain a high quality 'de novo' transcriptome for Castanea sativa

Marina Espigares[1], Pedro Seoane[1], Rocio Bautista[2], Isabel González Gayte[2], Luis Gomez[3], Julia Quintana Gonzalez[3], and M. Gonzalo Claros[1,2]

[1]*Departamento de Bioquímica y Biología Molecular, Universidad de Málaga, 29071 Málaga, Spain*
[2]*Plataforma Andaluza de Bioinformatica, Universidad de Málaga. Spain*
[4]*Departamento de Biotecnología. Universitat Politècnica de Madrid, Spain*

The European chestnut (*Castanea sativa*) is a tree that has been widely grown around the Mediterranean area. About 80% of grafted trees in El Bierzo (Spain) are affected by blight disease with important consequences in nut and timber production, and therefore a reduction of the economic value of this resource. The long term purpose of this study is to shed light on the control of blight disease, in particular by the fungus *Cryphonectria parasitica*, which causes an important impact on the chestnut industry. Since *C. sativa* is a non-model organism, our analysis must start with the construction of a high accurate de novo transcriptome for this tree, accurately annotated, that enables the discovery of genes with biotechnological potential to improve disease resistance. We have thus automatized a parallelized workflow where a primary assembly of short reads from Illumina Platform and long reads from Roche-454 technology, are performed individually using a set of k-mers from 25 to 35 and using different assemblers. The resulting contigs are then reconciled with the aim of obtaining the best transcriptome. All these combinations enable the generation of a large number of assemblies (29 plus 2 references) in a single run whose evaluation according to the sequencing technology, software and strategy, becomes very challenging. To automate the process, we implemented a metric system that is then data mined with a principal component analysis to identify patterns and understand how the assembly variability impacts the final result and which one is the best assembly. In our case scenario, the 31 different assemblies are gathered and segregated in five clusters in a very interesting way. We observed that the group of assemblies that reconcile 454 long reads using MIRA4 produces more complete and accurate reconstruction of genes than the assemblies reconstructed using MINIMUS from 454 and Illumina primary assemblies. This workflow can be tailored to the user's convenience in order to optimize it for a better transcriptome.

This work has been supported by co-funding from the ERDF (European Regional Development Fund) and PAI (grant P10-CVI-6075). The authors also thankfully acknowledge the computer resources and the technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga.

#063

The landscape of polymorphic inversions and their functional impact in the human genome

Jon Lerga-Jaso[1], Meritxell Oliva[1], Sergi Villatoro[1], David Izquierdo[1], Lorena Pantano[1], Sònia Casillas[1], and Mario Cáceres[1]

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

Recent advances in genomic techniques have generated an increasing interest in structural variants (SVs). However, there is still limited data on their functional impact. This is particularly true for inversions, in which their balanced nature and the complex repetitive regions where they appear make their study especially challenging. In this regard, we have designed a null model to elucidate the functional characteristics of these variants, taking as reference the most reliable set of human polymorphic inversions so far from the InvFEST database. Human inversions show the signature of natural selection, avoiding more often than expected not only genes, but also other functional elements such as enhancers, chromatin domain boundaries and highly conserved regions. Moreover, we measured the effect of 45 experimentally-genotyped polymorphic inversions on gene expression based on the transcriptome data from lymphoblastoid cell lines of 175 individuals from the Geuvadis project. Using a combination of well-established tools and further filtering criteria to minimize false positives, we found that although the majority of inversions do not appear to have any significant regulatory effect in this cell type, around half a dozen of them have a clear influence on specific genes, both in cis and in trans. Finally, we intersected SNPs reported as GWAS hits of different human phenotypes with the inversions in our dataset. All together, these results illustrate the potential impact of inversions on the human genome, and contribute to shed light on the molecular mechanisms responsible for phenotypic variability.

Support: European Research Council (ERC) Starting Grant (INVFEST).

#065

Toolkit to explore and analyze nanopore sequencing data on a Hadoop framework

Asunción Gallego[1], Joaquín Tarraga[1], Vicente Arnau[2], Ignacio Medina[3], and Joaquín Dopazo[1]

[1]*Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF). Spain*
[2]*Departamento de Informática, ETSE, Universidad de Valencia. Spain*
[3]*HPC Service, University Information Services, University of Cambridge. United Kingdom*

Background. The use of nanopore technologies is expected to spread in the future because they are portable and can sequence long fragments of DNA molecules without prior amplification. The first nanopore sequencer available, the MinION™ from Oxford Nanopore Technologies, is a USB-connected, portable device that allows real-time DNA analysis. In addition, other new instruments are expected to be released soon, which promise to outperform the current short-read technologies in terms of throughput. Despite the flood of data expected from this technology, the data analysis solutions currently available are only designed to manage small projects and are not scalable.

Results. Here we present HPG Pore [1], a toolkit for exploring and analysing nanopore sequencing data. HPG Pore can run on both individual computers and in the Hadoop distributed computing framework, which allows easy scale-up to manage the large amounts of data expected to result from extensive use of nanopore technologies in the future.

Conclusions. HPG Pore allows for virtually unlimited sequencing data scalability, thus guaranteeing its continued management in near future scenarios. HPG Pore is available in GitHub at http://github.com/opencb/hpg-pore.

[1] Tarraga J, Gallego A, Arnau V, Medina I, Dopazo J. 2016 HPG pore: an efficient and scalable framework for nanopore sequencing data. BMC Bioinformatics. 17(1):107.

#073

Pan-genomics pipeline for real-time comparative genomics analysis in public health

Giuseppe D'Auria[1]

[1]*FISABIO-Public Health, Valencia, Spain*

Whole microbial genome analysis based on massive sequencing of second and third generations represents a real framework in public health systems. Ad-hoc pipelines for genomic analysis provide in short laps of time (hours) information about taxonomy, comparative genomics (pan-genome) and single polymorphisms profiles. Thus, pathogenic organisms of interest can be tracked at the genomic level, allowing monitoring at one-time several variables including: epidemiology, pathogenicity, resistance to antibiotics, virulence, persistence factors, mobile elements and adaptation features. Such information can be obtained not only at worldwide level from public repositories, but also at the “local” level, by on-demand genome sequencing especially in the event of recurrent or emergent outbreaks.

We describe a streaming data analysis in microbial comparative genomics for a public health context. Such pipeline goes from massive sequencing data quality assessment through data mining for comparative analysis (pan-genomic). The pipeline searches for differential genetic features such as virulence, resistance/persistence factors and mutation profiles (SNPs and InDels) formatting comprehensible and summarized results. Such analytical protocols will enable a quick response to the needs of locally circumscribed outbreaks, providing information on the causes of resistance as well as genetic tracking elements for rapid detection and monitoring actuations for present and future occurrences.

#074

Differential functional activity and phenotype prediction from gene expression or mutations using models of signalling pathways

Marta Hidalgo[1], Cankut Çubuk[1], Alicia Amadoz[1], Francisco Salavert[1], Jose Carbonell-Caballero[1], and Joaquín Dopazo[1]

[1] *Centro de Investigación Príncipe Felipe. Spain*

Because of their multigenic nature, complex diseases are better understood as failures of functional modules caused by different combinations of perturbed gene activities or functionalities rather than by the malfunction of a unique gene. The increasing availability of detailed repositories of cell functionality (KEGG, Reactome, etc.) along with recent advancements in computational modelling of biological systems offer novel realistic alternatives to better understand complex biological systems (Fisher, 2015 Nat Biotech).

Here we present HiPathia, a web tool for the interpretation of the consequences that gene expression levels and/or genomic mutations occurring within signalling pathways can have over cell functionality. HiPathia transforms low-informative gene expression and/or genomic variation data into stimulus-response signalling circuit (sub-pathway) activities, a sophisticated type of biomarker which carries information on the different cell functionalities triggered by signals and that has recently proven to be superior to conventional biomarkers (Amadoz et al., 2015 Sci Rep 5:18494; Fey et al., 2015 Sci Signal 8:ra130). Such signalling circuit activities not only account for the underlying molecular mechanisms of diseases or the mode of action of drugs but they can also be used as mechanistic features for the prediction of complex phenotypes.

HiPathia implements an improved version of the Pathiways server (Sebastian-Leon et al., 2013 NAR 41:W213-W217) in which gene expression from any technology (microarray, RNA-seq) can be input. It also includes the estimation of deleteriousness of variants, like the PathiVar server (Hernansaiz-Ballesteros et al., 2015, NAR 43:W270-W275). HiPathia inputs files containing gene expression (in CSV format) and/or genomic variants (in VCF format) data, identifies all the signalling circuits within KEGG pathways, calculates a value of activation of each one, based on the individual gene expression values and/or the degree of deleteriousness of the variants carried by them, and identifies those with a significant differential activity between the two conditions compared. If gene expression and variants are simultaneously provided, both omic data are integrated. If only variation data are available, tissue-specific data from public databases (Expression Atlas or Human Protein Atlas) are used to define gene activity.

The graphical output represents the pathways analysed in which the possible ways by which the signal is transmitted from receptor proteins to the corresponding effector proteins are highlighted. In this way, disruptions or activations in the signal flux can be easily visualized and understood in terms of changes in gene expression and/or deleterious mutations.

In addition to estimate differential signalling activity, both omic data (alone or in combination) can be used to predict discrete classes (e.g., cancer type) or continuous variables (see an example of drug sensitivity prediction in Amadoz et al., 2015 Sci Rep 5:18494). A predictor can be built using a training dataset and can be further used to identify new, unknown samples.

HiPathia can be found at: http://hipatia.babelomics.org.

#075

Actionable pathways: interactive discovery of therapeutic targets using signalling pathway models

Francisco Salavert[1], Marta Hidago, Alicia Amadoz[1], Cankut Çubuk[1], Daniel Crespo[1], Jose Carbonell-Caballero[1], and Joaquín Dopazo[1]

[1] *Centro de Investigación Príncipe Felipe. Spain*

The discovery of actionable targets is crucial for targeted therapies and is also a constituent part of the drug discovery process. The success of an intervention over a target depends critically on its contribution, within the complex network of gene interactions, to the cellular processes responsible for disease progression or therapeutic response. Here we present PathAct, a web server that predicts the effect that interventions over genes (inhibitions or activations that simulate drug treatments or over-expressions) can have over signal transmission within signalling pathways and, ultimately, over the cell functionalities triggered by them. PathAct implements an advanced graphical interface that provides a unique interactive working environment in which the suitability of potentially actionable genes, that could eventually become drug targets for personalized or individualized therapies, can be easily tested.

The PathAct tool can be found at: http://pathact.babelomics.org.

#080

Comparison of reference-based and reference-free SNP matrix methods for outbreak detection of *Salmonella enterica* subsp. *enterica serovar* Kentucky

Jorge de La Barrera[1], Silvia Herrera[2], Isabel Cuesta[2], and Sara Monzón[2]

[1]*CNIC. Spain*
[2]*National Centre for Microbiology, Instituto de Salud Carlos III. Spain*

The advent of whole genome sequencing (WGS), has allowed a fundamental change in the way diagnostics and typing are performed for foodborne diseases. Replacing the myriad of classical bacterial typing techniques by one that is able to handle genome at single base level could reveal WGS as the sole standard method for microbiology. In this work, information obtained by some of the classical typing techniques (Pulse-field gel electrophoresis, Multilocus sequence typing, Serotyping and antibiogram) is compared to the information obtained from WGS for 10 isolates of Salmonella enterica subsp. enterica serovar. kentucky with the aims to elucidate the role of WGS in outbreak detection and pathogens surveillance in the near future. Two alternative typing approaches to study community-based outbreaks using SNPs matrix and phylogenetic analysis are evaluated: reference-based and reference-free. Based on results obtained in this work it is not able to conclude that one method is preferable than the other for outbreak detection. Suitable approach as well as reference selection (when pertinent) should be selected depending on bacterial strain and matter undertook (outbreak detection, traceback, emerging pathogen, etc.). for the samples included in this study, WGS have proved to be reliable detecting outbreaks identified using classical techniques and solved cases that such techniques cannot. However, it is still necessary to integrate epidemiological and molecular data (phylogeny and resistance profile) when deciding to declare an outbreak based on cluster detection.

#083

APPRIS selects the dominant cellular protein isoform

Jose Manuel Rodriguez[1], Angel Carro[2], Alfonso Valencia[2], and Michael Tress[2]

[1]*Spanish National Bioinformatics Institute (INB-CNIO). Spain*
[2]*Spanish National Cancer Research Centre (CNIO). Spain*

The APPRIS database (http://appris.bioinfo.cnio.es) uses protein structural and functional features and information from cross-species conservation to annotate splice isoforms from protein-coding genes [1]. APPRIS selects one of these isoforms to be the principal protein isoform for each gene. Generally this main isoform has the most conserved protein features and the most evidence of cross-species conservation. Those isoforms with unusual, missing or non-conserved protein features are flagged as alternative.

APPRIS principal isoforms have a wide range of uses and are applicable in all fields of research. Determining a principal isoform is important for research groups studying individual genes, and the designation of a single variant as the principal isoform is a critical first step for any genome analysis, for example studies of cancer mutations would be able to use APPRIS data to determine whether the mutations are in principal or alternative exons.

APPRIS principal isoforms have great practical use as experimental evidence clearly shows that principal isoforms are the main protein isoforms in the cell. APPRIS principal isoforms coincide overwhelmingly with the main protein isoform detected in proteomics experiments [2], with the variant with clearest cDNA evidence [3] and with the transcript with most reliable RNAseq evidence [in house results]. In addition they are under significantly greater selective pressure than alternative isoforms [4].

The APPRIS Database houses annotations for seven Ensembl [5] species (human, mouse, rat, pig, zebra-fish, fruit-fly and C. elegans), and for the RefSeq [6] human gene set. The recently developed APPRIS WebServer and WebServices [7] allow users to check Ensembl annotations for nine other species, dog, cat, cow, opossum, chicken, zebra-finch, lizard, xenopus and fugu, and to interrogate the APPRIS database in an automatic fashion.

APPRIS is stable and is implemented as part of the GENCODE/Ensembl human genome annotation [8], and can be visualized in the UCSC Genome Browser [9] as public track hub.

[1] Rodriguez JM et al. (2013) Nucleic Acids Res. 41:D110-7.
[2] Ezkurdia I et al. (2015) J Proteome Res. 14:1880-7.
[3] Harte RA et al. (2012) Database 2012:bas008.
[4] Liu T and Lin K. (2015) Mol. Biosyst. 11:1378-88.

[5] Yates A et al. (2016) Nucleic Acids Res. 44:D710-6.
[6] O'Leary NA et al. (2016) Nucleic Acids Res. 44:D733-45
[7] Rodriguez JM et al. (2015) Nucleic Acids Res. 43:W455-9.
[8] Harrow J et al. (2012) Genome Res. 22:1775-89.
[9] Kent WJ et al. (2002) Genome Res.12:996-1006.

#101

Super-Cap an integrated workflow to capture rare variants, structural variations and to build private gene sequences

Valentino Ruggieri[1], Irantzu Anzar[1], Andreu Paytuví[1], Roberta Calafiore[2], Amalia Barone[2], Riccardo Aiese Cigliano[1], and Walter Sanseverino[1]

[1]*Sequentia Biotech. Spain*
[2]*UNINA. Italy*

The recent development of Sequence Capture methodology represents a powerful strategy for enhancing data generation to assess genetic variation of target regions. However, the association of the genomic variation with certain traits requires a reliable detection and a systematic investigation of the entire spectrum of DNA variation, including single nucleotide polymorphisms, insertions/deletions as well as copy number variation (CNV), and presence/absence variation (PAV). In this study, an ad hoc pipeline, for reads mapping, variant calling and reconstruction of private sequences was developed. SUPER-CAP pipeline profiting by SUPER-W, a tool developed for SNPs and SVs calling (Sanseverino et al, 2015), has been specific written to handle Sequence Capture data, fine calculate the allele frequency of variations and to build private sequence of captured genes. In this study, a survey of 378 loci and related regulative region in a collection of 44 tomato landraces was carried out. By using Roche-Nimblegen technology, about 14000 high-quality variants were identified. A validation of a subset, by exploiting shared data present in tomato public SNP repository, showed >95% of similarity. The high depth (> 40X) and the proper filtering criteria adopted allowed to identify about 4000 rare variants and 10 genes with a putative CNV or PAV. In addition, in order to reconstruct private sequences for each genotype the variants detected were linked into single haplotypes. This allowed for example to evaluate the combined effect of multiple variants and to assess perturbations of the promoter cis-acting elements.

#103

Identification of Cancer Key Metabolic Patterns Using RNA-Seq Data

Cankut Çubuk[1], Marta Hidalgo[1], Alicia Amadoz[1], Jose Carbonell-Caballero[1], and Joaquín Dopazo[1]

[1] *Centro de Investigación Príncipe Felipe. Spain*

Metabolic abnormalities are the main cofounders of the cancer cells. Characterization of the metabolic biomarkers and targets are of paramount importance for efficient cancer diagnosis and treatment.

In our study, we used K-shortest Elementary Flux Modes (EFMs) [1], metabolic network topology of Recon1 [2] and the shortest path approach. Since, EFMs guarantee the steady-state conditions, the reaction stoichiometry were used only to generate the reaction-reaction interaction network. In this network, the nodes represent the reactions and the edges represent the connections between the metabolic reactions. Moreover, each of the EFMs was thought as a particular cellular system and exchange reactions were considered as entrance points of the metabolic flow. The shortest paths between exchange reactions were found and gene expression data were mapped to their metabolic reactions using gene-protein-reaction (GPR) associations.

Two different behaviours for metabolic reactions were observed between cancer and normal cells; all reaction values inside a shortest path were altered in a similar portion or some particular reactions were altered independently from the others. Based on these observations, we classified the alteration status of the reactions as moderate and disturbed. Later on, we compared the recurrence percentage of these reactions among all shortest paths. This method has been applied to 3 different RNA-Seq datasets [kidney renal clear cell carcinoma (KIRC), breast invasive carcinoma (BRCA), bladder urothelial carcinoma (BLCA)] from The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/).

Unlike the moderately altered reactions, the recurrency percentage of the disturbed reaction series were non-uniform between different cancer types. Using all these disturbed reactions we were able to predict the cancer type with high accuracy (>%90).

These findings let us to think that the disturbed reactions, profile of their metabolic enzymes and metabolites can be cancer cell specific. The moderately altered reactions were related with lipid metabolism, cell growth and proliferation which are essential cellular functions to differentiate tumor cells.

So far, our results were highly correlated with the literature. Moreover, we obtained metabolic patterns to distinguish the different cancer types which can be used for cancer type specific treatment and drug development. However, the missing GPRs are the main limitation of our study. We believe that the novel models such as ReconX series will provide more accurate results.

[1] Rezola, A., Pey, J., de Figueiredo, L., Podhorski, A., Schuster, S., Rubio, A. and Planes, F. (2013) Selection of human tissue-specific elementary flux modes using gene expression data.Bioinformatics, 29, 2009-2016.
[2] Duarte, N., Becker, S., Jamshidi, N., Thiele, I., Mo, M., Vo, T., Srivas, R. and Palsson, B. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences, 104, 1777-1782.

#112

MGvizCE: Clinical Exome QC and Analytics

Pablo Marin-Garcia[1], Daniel Perez-Gil[2], Cristian Perez-Garcia[1], Alba Sanchis-Juan[2], Azahara Fuentes[2], Jose M. Juanes[3], Alberto Labarga[4], Antonio Fabregat[5], Vicente Arnau[6], Javier Chaves-Martinez[2], Javier Garcia-Planells[1], and Ana Barbara Garcia-Garcia[2]

[1]*Instituto de Medicina Genómica. Spain*
[2]*INCLIVA. Spain*
[3]*Seqplexing. Spain*
[4]*NAVARRABIOMED. Spain*
[5]*EBI. United Kingdom*
[6]*Escuela Técnica Superior de Ingenierías (ETSE, UVEG). Spain*

NGS facilities dedicated for clinical genomics need high QC standards and they need continuously keep track of their experiments and their metrics. As part of the Medical Genomics Visualization toolset (MGviz) we have developed an interactive software suit with R-Shiny and Python (Bokeh, crossfilter, flask and ReportLab) for automatic reports of QC for the whole NGS experiments in clinical diagnostics labs. The tool allows the comparison of the current experiment with historic data to see the performance of the sequencher, check different metrics for coverage and variations, warns for large copy number regions, remember decisions over annotations, helps in variant prioritization, segregation and finding compound heterozygotes and make automatic historical reports of pathogenic variants informed by the lab.

#114

Analysis of methylated DNA from psoriasis patients treated with anti-TNF drugs

Teresa Cabaleiro[1], Rocío Prieto-Pérez[1], María Talegón[1], Miriam Saiz Rodriguez[1], Esteban Daudén[2], Manuel Román[1], Dolores Ochoa[1], Francisco Abad-Santos[1], and María C Ovejero-Benito[1]

[1]*Clinical Pharmacology Service, Hospital Universitario de La Princesa, Instituto Teófilo Hernando, University Autónoma de Madrid (UAM), Instituto de Investigación Sanitaria La Princesa (LP), Madrid, Spain*
[2]*Dermatology Service, Hospital Universitario de La Princesa, Instituto de Investigación Sanitaria La Princesa (LP), Madrid, Spain*

Psoriasis is a chronic, autoimmune and inflammatory skin disorder related to a combination of genetic, environmental and immune factors that affects to 1.3-2.2% of the world population. This strongly disabling disease interferes with patients' daily life and presents a wide range of comorbidities such as cardiovascular diseases, cancer and depression that can decrease the life expectancy of psoriasis patients. Anti-TNF drugs have been the main biologic drug to treat moderate-to-severe psoriasis so far and its effectiveness can reach 80%. However, the clinical response to the administration of these drugs varies depending on the genetic and the environment of the patient. To understand this phenomenon, analyses of DNA methylation of patients treated with anti-TNF drugs were performed. Blood samples were collected from 72 patients who suffered from moderate-to-severe psoriasis. DNA extracted from these samples was treated with sodium bisulfite, amplified, labeled, hybridized to methylated and unmethylated probes and microarray scanning platform in HiScanSQ Illumina Inc. Human Methylation450 BeadChips technology was used as it allows the simultaneous analysis of 485,000 individual CpGs sites. After analyzing the results with bioinformatic tools such as Genome Studio, Circos and R version 3.1.2, significative differences in the degree of methylation of several CpG islands were found. These islands regulate the expression of genes that have not been involved so far in the pathology of psoriasis. These results are very promising because they help to seed light to the mechanisms involved in this disease and path the way to find new drugs to treat Psoriasis.

DEANN: Developing an European American NGS Network

Eugenia Flores de La Luna[1] and Ana Conesa[1]

[1]*Centro de Investigación Príncipe Felipe (CIPF). Spain*

The DEANN project is a network formation initiative that involves seven research institutions and universities from four EU countries (United Kingdom, Sweden, Spain and Italy) and eight research entities from four Latin American countries (Argentina, Mexico, Brazil and Chile). The project is being performed under the Marie Sklodowska Curie action from the European Commission, IRSES. The overall goal of this initiative is to strengthen research partnership among project participants by developing a shared scientific know how in the field of NGS data analysis. This will lead to increase the scientific competence of consortium members at the international level. More specifically, the objectives of the project are to reinforce collaboration, nurse scientific excellence, elevate the critical mass, improve education and achieve translational opportunities.

In the last two years a total of 59 exchange visits have been performed involving 47 researchers, 29 ESRs and 18 ERs (21 visits have been performed from EU to LA and 38 from LA to EU). In the remaining two years a total of 218 visits will be accomplished. During these collaborative exchanges the consortium has work on the following topics:

1. Genome analysis of the Southamerican admix populations to identify candidate genes for distinctive phenotypic traits.
2. (Functional) annotation of novel genomes for species biomedical and biotechnological relevance.
3. Annotation of ancient DNA samples.
4. Methods to integrate complex omics datasets such as genomics, RNA-seq, microRNAs, proteomics, etc.
5. Genomics and transcriptomics of the biodiversity, including forest and plant species, marine and sponge organisms and Artic species.
6. Moreover, the project has organized a large number of training activities.

By presenting this work we want to highlight the importance that NGS technologies have within the European Research Programme and how they are being tackled between continents in a common and shared aim.

Scientific Session:

Systems and Synthetic Biology (G).

#009

Learning the genome

Davide Bau[1,2]

[1]CNAG-CRG, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.
[2]Gene Regulation, Stem Cells and Cancer Program, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain

The three-dimensional organization of chromatin drives gene expression by bringing together genes and their interacting partners.

Different genes in different cell types are expressed differently, despite sharing the same DNA sequence. Epigenetic modifications defines cell specificity; specific patterns of covalent modifications of the histone tails impact gene expression by altering chromatin structure or recruiting histone modifiers in a cell specific way. Learning how the epigenomic landscape contributes to the cellular outcome could help understanding how the genome is regulated.

Using machine learning techniques, I have integrated histone modifications, chromatin accessibility and gene expression data to study the impact of post-translational modifications on genome architecture. My results show that learning from a subset of all the chromosomes is sufficient to accurately predict a set of features on the remaining chromosomes. My predictions provide a useful tool to help understanding how the genome is organized and to complement incomplete genomic data.

#045

Computational RNA models guide the engineering of novel gene regulatory systems from sequence to function

Guillermo Rodrigo[1]

[1]*IBMCP (CSIC-UPV). Spain*

A grand challenge in synthetic biology is to use our current knowledge of RNA science to perform the automatic engineering of completely synthetic sequences encoding functional RNAs in living cells. We developed a computational algorithm based on a physicochemical model to produce novel RNA sequences by exploring the space of possible sequences compatible with predefined structures [1, 2]. for that, we followed an evolutionary design strategy (mutation then selection). We tested our methodology by designing several riboregulators with diverse structures and interaction models, suggesting that only the energy of formation and the activation energy are sufficient criteria to engineer RNA interaction and regulation in vivo. Extending these seminal ideas, we implemented a new algorithm in the form of webserver, called RiboMaker, with relaxed structural requirements [3]. It optimizes the sequences of a small regulatory RNA and a 5' untranslated region for an efficient intermolecular interaction. In addition, we applied this bioinformatic framework to design novel gene regulatory systems. In particular, we designed and implemented in vivo RNA-mediated signal transduction cascades able to sense small molecules and small RNAs [4, 5]. The engineered systems integrate RNA-RNA interaction with available ribozyme or riboswitch elements, providing new ways to engineer arbitrary complex gene circuits. We also engineered a riboregulator as the negative-sense strand of another riboregulator [6], illustrating the versatility of our approach. Finally, we also studied the dynamic behavior of these engineered systems. for that, we developed theoretical framework by following a themodynamic equilibrium description to predict the response with dose and time. The theoretical calculations were confirmed experimentally.

[1] G. Rodrigo, et al. Proc. Natl. Acad. Sci. USA 109 (2012) 15271-15276.
[2] G. Rodrigo G, et al. PLoS Comput. Biol. 9 (2013) e1003172.
[3] G. Rodrigo et al. Bioinformatics 30 (2014) 2508-2510.
[4] S. Shen S, et al. Nucleic Acids Res. 43 (2015) 5158-5170.
[5] G. Rodrigo et al. Nucleic Acids Res. – (2016) to appear.
[6] G. Rodrigo G, et al. J. Mol. Biol. – (2016) to appear.
[7] G. Rodrigo G, et al. Biophys. J. 109 (2015) 1070-1076.

#047

Functional Meta-Analysis for Genomic Studies

Francisco García-García[1], Iván Ansari Toledano[2], Cristina Escribano[1], Joaquin Dopazo[3], and David Montaner[4]

[1]*Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain*
[2]*Departamento de Informática, ETSE, Universidad de Valencia, Valencia, Spain*
[3]*Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF). Bioinformatics of Rare Diseases (BIER), Ciber de Enfermedades Raras (CIBERER). Functional Genomics Node, (INB) at CIPF, Valencia, Spain.*
[4]*Genomics England, London, United Kingdom.*

Introduction. Computational methods play a key role in the resolution of clinical and biological problems. Generation of large amounts of data from high-throughput technologies and increasing information accessible in biological databases has boosted the demand for new methodologies able to link both. Functional enrichment analysis of genomic data provides outcomes that are nowadays an integral part of the results of the experiment. However, the small sample size of most of the experiments and their connection to a specific scenario, represent limiting factors when evaluating such studies. Therefore, to improve the integration of various experiments in the functional context and provide clarity in the interpretation, we present a meta-analysis method to detect functional results of global interest, reducing experiment-specific context effects.

Data. We used two different genomic datasets to evaluate this methodology. From Gene Expression Omnibus (GEO) we selected 26 microarrays studies related to psoriasis and dermatitis. In the initial screening, a case-control experimental design for human was required, where cases showed skin lesion and controls were free of injury. On other hand, we downloaded a second dataset with 20 microRNA data from The Cancer Genome Atlas project, which contain both tumoral and healthy samples.

Methods. After preprocessing these data, differential expression analysis from Babelomics [1] and enrichment analysis using logistic models [2][3] were carried out for each study. Biological information from Reactome, KEGG pathways and Gene Ontology databases was used. Odds ratios for each functional term were combined to detect a global association between experimental groups and studies. Variability for each study was estimated from several methods for fixed and random effects (DerSimonian-Laird, Hedges, Hunter-Schmidt, Sidik-Jonkman, ...)

Results. We detected common significant functions in psoriasis and dermatitis along all studies. This point confirms the relationship between both diseases. There is a clear impact in immunological processes and highlights the involvement of macrophages, lymphocytes, and mast cells.

For the second dataset, the functional evaluation of tumors generally shows some affected functions for most of individuals and detects groups of tumors with specific functionalities.

Conclusions. This methodology is useful to detect relevant genetic alterations in disease, able to confirm functionality already described and show new functional relationships that may be of interest to initiate new studies. The method is flexible when using functional information: Gene Ontology terms, signaling pathways or any other function could be incorporate in the Gene Set Analysis for each study. The proposed methodology is also flexible in the selection of different methods to estimate the variability between studies (fixed and random effects), adjusting the method to specific characteristics for studies.

[1] Babelomics 5.0: functional interpretation for new generations of genomic data. Alonso R et al. Nucleic Acids Res. 2015 Jul 1;43(W1):W117-21.
[2] Multidimensional gene set analysis of genomic data. Montaner D et al. PloS One. 2010 Apr 27;5(4).
[3] LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Sartor MA et al. Bioinformatics. 2009, 25(2), 211–217.

#082

Prediction of bacterial optimal growth rates in insect-endosymbiont systems at exponential growth.

Jorge Calle Espinosa[1], Francisco Montero[1], and Juli Peretó[2]

[1] *Universidad Complutense de Madrid. Spain*
[2] *Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València. Spain*

Nutritionally driven symbioses between insects and one or more intracellular, maternally inherited bacteria (endosymbionts) are widespread. These mutualistic associations allow insects to colonize novel ecological niches with unbalanced nutritional sources. In these systems, the bacterial endosymbiont takes the nutrients which are in excess in the insect diet and produce others which are essential for its host. for these relation to function in optimal conditions, the endosymbiont's growth rate along with the fraction of it production derived to the insect may be regulated properly. Here we show that for the time to adulthood to be minimized, a wide range of growth rates for the endosymbiont are acceptable, even if the production of the compounds needed for the insect competes directly with the endosymbiont's growth. These results suggest that the control over the endosymbiont population does not need to be strict until adulthood is reached and the needs for nutrients decreases.

#085

Integration of proteomic and metabolomic data in genome-scale metabolic models and its application to the cyanobacteria *Synechocystis sp. PCC 6803*.

Marina Pérez-Naveira[1], Maria Siurana[2], and Javier F. Urchueguía[1]

[1]*Institute for the Applications of Advanced Information and Communication Technologies (ITACA). Spain*
[2]*Institute for Applied and Pure Mathematics (IUMPA). Spain*

A genome-scale metabolic model is an informatic assembly of the whole set of reactions that are present in the metabolism of an organism that, used in combination with an analytic method, enables to partially simulate its metabolic response. The result obtained from the simulations is a flux distribution throughout the reaction landscape of the organism.

Cyanobacteria are important due to its ability to perform photosynthesis and can be used as biofactories. A detailed knowledge of their behavior allows the optimization of their performance on the production of high-valuable compounds.

From our cyanobacterial Genome Scale Metabolic Model (GSMM) of *Synechocystis sp. PCC 6803* [1,2] is possible to analyze the general behavior of cyanobacteria and to optimize the yield of production of compounds of interest.

The main objective of this work is to apply an algorithm created by Yizhak et al. [3] – termed IOMA - that allows the utilization of metabolomic and proteomic data, in connection with our GSMM allowing, for the first time, the use of this type of 'omic' data to constrain simulated flux landscapes in a way that is consistent with proteomic information. To this purpose we have incorporated a large set of kinetic and 'omic' data into the model and adapted the IOMA algorithm to perform a whole set of metabolic simulations.

This work has shown that this method gives better qualitative results when compared to standard Flux Balance Analysis in terms of glycogen production and pathways behavior when *Synechocystis* is under stress. The metabolomics and proteomics data used allow the simulation to be representative even when the stress is not present in the model, which represents a substantial step forward in comparison with other state-of-the-art analytical methods like FBA or MOMA.

[1] Montagud, A., Navarro, E., Fernandez de Cordoba, P., Urchueguia, J. F., & Patil, K. R. (2010). Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. BMC Systems Biology, 4, 156. doi:10.1186/1752-0509-4-156

[2] Montagud, A., Zelezniak, A., Navarro, E., de Córdoba, P. F., Urchueguía, J. F., & Patil, K. R. (2011). Flux coupling and transcriptional regulation within the metabolic network of the photosynthetic bacterium Synechocystis sp. PCC6803. Biotechnology Journal, 6(3), 330–342. doi:10.1002/biot.201000109

[3] Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E., & Shlomi, T. (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Bioinformatics, 26(12), 255–260. doi:10.1093/bioinformatics/btq183

#096

Modelling non-steady state metabolic fluxes using dynamic elementary modes

Abel Folch-Fortuny[1], Bas Teusink[2], Henk A.L. Kiers[3], Huub C.J. Hoefsloot[4], Age K. Smilde[4], and Alberto Ferrer[1]

[1]*Universitat Politècnica de València. Spain*
[2]*Free University of Amsterdam. Netherlands*
[3]*University of Groningen. Netherlands*
[4]*University of Amsterdam. Netherlands*

Principal component analysis (PCA) and multivariate curve resolution (MCR) models have been proposed to obtain a set of key pathways in metabolic networks, assuming steady state conditions [1,2]. These pathways or modules in the network are identified using the existing relationships between metabolic fluxes, measured experimentally. Recently, a new method called principal elementary mode analysis (PEMA) [3] has been proposed to model this kind of data. The methodology is based on the flux projection to a reduced set of elementary modes (EMs) of the metabolic network. The EMs are the simplest representations of pathways crossing the metabolic network. Basically, each EM connects substrates with end-products concatenating reactions in a thermodynamically feasible way. for non-steady state cases, e.g. when measuring the concentrations of the metabolites at early stages after perturbation, 13C-metabolic flux analysis (MFA) [4], dynamic flux balance analysis (DFBA) [5,6], and the Goeman's Global test [7] have been proposed, among other methods.

Here we define a new framework to model non-steady state metabolic fluxes. This methodology is based on the novel concept of dynamic EMs (dynEMs), i.e. EMs that are used partially at each time point of the experiment. In this way, we propose dynamic elementary mode regression discriminant analysis (dynEMR-DA) to identify the set of dynEMs whose activation pattern allows discriminating between different biological conditions. Actual [7] and simulated [8] non-steady state flux data sets from Saccharomyces cerevisiae are analysed using this methodology, identifying the most discriminant dynEMs when changing the initial concentrations of glucose, and when changing from aerobic to anaerobic conditions.

[1] J.M. González-Martínez et al. Chemometr. Intell. Lab. 134 (2014) 89-99.
[2] A. Folch-Fortuny et al. Chemometr. Intell. Lab. 142 (2015) 293-303.
[3] A. Folch-Fortuny et al. Mol. BioSyst. 12 (2016) 737-746.
[4] W. Wiechert, Metab. Eng. 3 (2001) 195-206.
[5] R. Mahadevan et al. Biophys. J. 83 (2002) 1331-1340.
[6] A.M. Willemsen et al. Mol. BioSyst. 11 (2015) 137-145.
[7] D.M. Hendrickx et al. Anal. Chim. Acta 719 (2012) 8-15.
[8] J.H. van Heerden et al. Science 343 (2014) 1245114.

Posters

#030

MEAL: Methylation and Expression AnaLyzer

Carlos Ruiz Arenas[1], Carles Hernández Ferrer[1], and Juan R Gonzalez[2]

[1]*CREAL. Spain*
[2]*Center for Research in Environmetal Epidemiology. Spain*

Summary: There is an increasing interest in the integrative analysis of the transcriptome and the methylome to understand their joint role in complex phenotypes. There have been some efforts in providing tools for integrating transcriptomics and methylomics using multivariate methods. However, there is not a common tool where integrative analyses can be performed with the most commonly used models.

Results: We introduce MEAL, an R package that enables the analysis by pairs of methylation and expression probes as well as multivariate approaches. MEAL can also analyze the global relationship between methylation and expression in a specific genomic region. The package can also analyze methylation or expression data separately, incorporating the state-of-the-art outputs and plots. It allows controlling for the effect of SNPs to avoid finding probes that can be eQTL or meQTL. To this end, new classes have been created to handle the inclusion of different omic data.

Availability and implementation: MEAL is freely avail-able as a Bioconductor package. The development version, other documentation and data used in illustrating examples can be found in BRGE web page (Bioinformatics Research Group in Epidemiology, http://www.creal.cat/brge.htm)

#037

INTEGRATIVE VISUALIZATION of MULTI-OMICS DATA: THE PAINTOMICS 3 PLATFORM

Rafael Hernández-De-Diego[1], Pedro Furió-Tarí[1], Sonia Tarazona[1], and Ana Conesa[1,2]

[1]*Genomics of Gene Expression Lab, Centro de lnvestigación Principe Felipe, Valencia, Spain*
[2]*Microbiology and Cell Science Department, University of Florida, Gainesville, USA*

A logical consequence of the advances of high-throughput technologies and the arising of new omics approaches is the consideration of combining those complementing measurements at the same experiment. Nevertheless, the high dimensionality and heterogeneity of multi-omics data add several layers of complexity to the integrative analysis, becoming a challenge to extract meaningful information and use them to answer fundamental biological questions. Within this scenario, joint visualization of multi-omics data arises as a powerful tool which can lead toward a comprehensive understanding of biological systems.

Our purpose is to present Paintomics 3, a web-based platform for integrated visualization of multiple omics datatypes onto KEGG pathway diagrams. Paintomics 3 combines the server-side capabilities for data analysis with the potential of modern web resources for data visualization, providing to researchers a visual and interactive exploration of their multi-omics data.

In opposition to other visualization tools, the system covers a complete pathway analysis workflow, including automatic feature name/identifier conversion, pathway enrichment analysis and network analysis. Paintomics 3 displays the user data onto interactive KEGG pathway diagrams, where concentrations levels are represented with intuitive fashions for visualization of changes in biological data, such as interactive heatmaps and trend charts, which are complemented with other useful resources such as clustering methods and links to external databases.

Paintomics 3 accepts a wide variety of omic types, including popular omics types such as transcriptomics, Proteomics or Metabolomics, as well as region-based approaches such as DNase-seq or ChiP-seq data; for numerous KEGG organisms. The tool is free to use and open source and its available at
[http://bioinfo.cipf.es/paintomics].

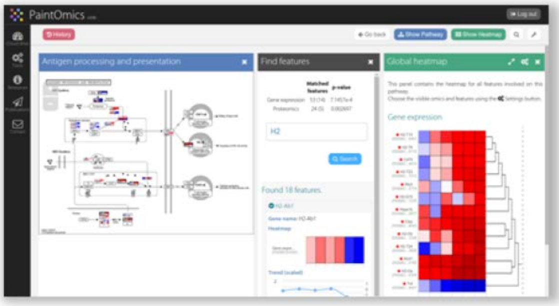


Figure 1. Integrative visualization of multiple omics data types onto the KEGG pathway "Antigen processing and presentation pathway" for *Mus Musculus*.

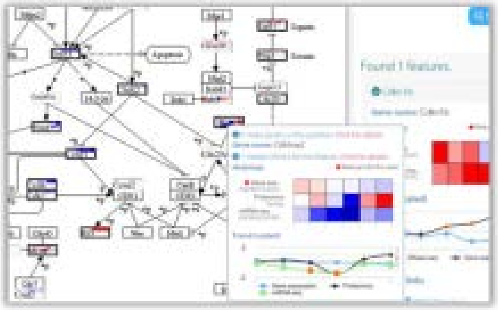


Figure 2. Detail for a joint visualization of transcriptomics, proteomics, DNAse-seq and miRNA-seq data using Paintomics 3.

#046

PanelMaps: a web tool for detection and visualization of altered regions for targeted sequencing

José M. Juanes[1], Francisco García-García[2], Joaquin Dopazo[3], and Vicente Arnau[1]

[1]*Departamento de Informática, ETSE, Universidad de Valencia, Valencia, Spain*
[2]*Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia Spain*
[3]*Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF). Bioinformatics of Rare Diseases (BIER), Ciber de Enfermedades Raras (CIBERER). Functional Genomics Node, (INB) at CIPF, Valencia, Spain*

Introduction. Gene panel sequencing allow us to detect variants associated with different diseases. Sometimes, the possible cause of the disease is not due to variations in SNPs (Single Nucleotide Polymorphism) or INDELs (small insertions-deletions) and can be motivated by the presence of a larger variation: deletion or insertion.
The aim of this work is the design and development of a web tool for detection and visualization of altered regions from targeted sequencing data. How does PanelMaps work?

Inputs. One or several BAM files (one file for each individual) and a BED file including all regions for gene panel. After loading data, the coverage of each sample is calculated and these values are normalized between all samples considering the total number of reads of each sample.

Methods. The tool includes two modules: visualization of genes or regions from coverage data and a module analysis to detect regions of interest. for detection of altered regions, PanelMaps uses a control sample selected by the user or combines all samples to get a common reference. Users can also specify comparisons between subgroups of samples of interest. The analysis incorporates a sliding window algorithm with various parameters adjustable by the user and related to the precision and characteristics of the region to be detected.

Outputs. PanelMaps shows a graphical description of coverage levels for genes and samples to confirm that all regions are covered. This web tool visualizes all regions included in the panel and shows a selection of altered regions between samples.

Conclusions. Panelmaps is a useful tool for detection and visualization regions of genes altered in panels that improves the knowledge of the genetic basis of diseases and produces useful information for diagnosis in clinical contexts.
This web tool is an alternative to the use of molecular biology techniques such as MLPA (Multiplex Ligation-dependent Probe Amplification), which are very costly and sometimes have some technical problems such as failure to detect variants or micro-deletions in positions where primers are incorporated.

#049

MBROLE 2.0 - Functional enrichment of chemical compounds

Javier Lopez-Ibáñez[1], Florencio Pazos[1], and Mónica Chagoyen[1]

[1]*Centro Nacional de Biotecnología. Spain*

Here we present the ongoing work on MBROLE 2.0, the new version of a web-based tool for performing enrichment analysis of metabolomic data.

Metabolomics is one the more recent 'omic' approaches, and it is growing in importance not only as a complement to other 'omics' but also due to its importance in biomedical and applied research [1]. Metabolomic experiments may generate huge amounts of data, which makes it necessary to use bioinformatic approaches to retrieve relevant biological information.

MBROLE was the first web-tool for performing enrichment analysis of chemical and biological annotations of metabolomic data on a wide range of organisms [2].

Since its release, MBROLE have been used in many published metabolomic studies on different organisms: human [3], e. coli [4], thermobifida fusca [5], synechococcus elongates [6]. This new version of MBROLE will include data from ten new databases, comprising metabolite annotations, such as interactions with enzymes, proteins or drugs, more organism-dependent pathways, uses and applications etc.

Other new features are an automatic conversion of compound identifiers (IDs) so users could mix IDs from different databases as input, a new and more intuitive interface and an improved reporting of results with sortable fields, coloring of most significant results and more export capabilities.

[1] Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omic trilogy. Nat. Rev. Mol. Cell Biol. 13, 263–269 (2012).
[2] Chagoyen, M. & Pazos, F. MBRole: enrichment analysis of metabolomic data. Bioinformatics 27, 730–731 (2011).
[3] García-Cañaveras, J. C., Donato, M. T., Castell, J. V. & Lahoz, A. A comprehensive untargeted metabonomic analysis of human steatotic liver tissue by RP and HILIC chromatography coupled to mass spectrometry reveals important metabolic alterations. J. Proteome Res. 10, 4825–4834 (2011).
[4] Carneiro, S., Villas-Bôas, S. G., Ferreira, E. C. & Rocha, I. Influence of the RelA Activity on E. coli Metabolism by Metabolite Profiling of Glucose-Limited Chemostat Cultures. Metabolites 2, 717–732 (2012).

[5] Vanee, N. et al. Proteomics-based metabolic modeling and characterization of the cellulolytic bacterium Thermobifida fusca. BMC Syst. Biol. 8, 86 (2014).

[6] Diamond, S., Jun, D., Rubin, B. E. & Golden, S. S. The circadian oscillator in Synechococcus elongatus controls metabolite partitioning during diurnal growth. Proc. Natl. Acad. Sci. U. S. A. 112, E1916–1925 (2015).

#052

sRNAtoolboxVM: small RNA analysis in a box

Michael Hackenberg[1], Antonio Rueda[2], Ricardo Lebrón[1], José L. Oliver[1], and Cristina Gómez-Martín[1]

[1]*University of Granada. Spain*
[2]*Queen Mary University of London. Spain*

In recent years, small RNA molecules have become one of the most interesting topics in genomics. This is due to its role in many vital processes such as cell differentiation [1]. More recently, microRNAs were shown to be useful as prognostic and diagnostic biomarkers, therapeutic agent in medicine [2], or as transgenes in plant science [3], among others applications. This, together with the decreasing cost of High-Throughput Sequencing (HTS), leads to a tremendous spread of small RNA research.

At the downside, the analysis of such enormous amount of data has now become the bottleneck in HTS-based research. One of the reasons is that a part of the software is only available for Linux operating systems which need to be installed and properly maintained, a task usually difficult to handle for non-expert users that often requires skilled personal. Additionally, the corresponding costs usually cannot be afforded by small to mid-sized laboratories. An alternative is the implementation in web-servers, like miRanalyzer [4], but those have problems with privacy issues, which is very important for medical data.

We present here a virtual machine, sRNAtoolboxVM, which tries to overcome these problems and limitations. It is platform independent and can be run by non-expert users. It implemented a Linux OS and counts with a complete set of preinstalled software for small RNA analysis: 1) expression profiling from NGSdata with sRNAbench [5]; 2) prediction of novel microRNAs; 3) characterize unmapped reads (sRNAblast [6]); 3) detection of other smallRNAs (sRNAbench [5]); 4) differential expression of different RNA types (sRNAde [6]); 5) downstream analysis: target gene prediction (miRNAconsTargets[6]) and functional analysis of target genes (pathways, GO-terms). It also includes helper tools that automatically populate the local database and keep the machine up to date.

[1] Ivey KN et al. Cell Stem 7 (2010) 36.41.
[2] Iorio M V. et al. EMBO Mol Med 4 (2012) 143-59
[3] Zhou M et al. Plant Mol Biol 83 (2013) 59-75
[4] Hackenberg M et al. Nucleic Acids Res 39 (Web Server Issue) (2011) 132-8
[5] Barturen G et al Methods in Next Generation Sequencing 1 (Web Server Issue) (2014) 21–31
[6] Rueda A et al. Nucleic Acids Res 43 (Web Server Issue) 467-473

#053

MethFlow: a pipeline for high-quality methylation calling

Guillermo Barturen[1], Cristina Gómez-Martin[2], José L. Oliver[2], Michael Hackenberg[2], and Ricardo Lebrón[2]

[1]*Centro de Genómica e Investigaciones Oncológicas, Pfizer-Universidad de Granada - Junta de Andalucía. Spain*
[2]*Universidad de Granada. Spain*

DNA methylation is an essential epigenetic mark involved in several cellular process such as transcription, chromatin structure or X chromosome inactivation among others. Moreover, it has been demonstrated to be mandatory for cell fate and differentiation. As is well known, most expressed genes show hypomethylation at promoter region and hypermethylation of the gene body. In pathological conditions, such as cancer, a global reduction of DNA methylation levels and the aberrant hypermethylation of CpG islands have been observed. However, methylation is not static also in physiological conditions and it varies among individuals, tissues and even cells from the same tissue. There are even some evidences that methylation depends on sequence variation in plants, but this relationship is still unclear in mammals.

We present MethFlow, a ready-to-use pipeline for simultaneously calling methylation levels and SNVs. This simultaneous calling is a keystone for exploring the impact of variation in methylation patterns. Our pipeline performs 1) preproccesing of methylation data, including format converting and adaptor trimming; 2) aligment against bisulfite-converted genomes; 3) methylation and SNVs calling, using MethylExtract; 4) several quality controls, such as fix bisulfite bias, phred score and coverage thresholds. Data input from third-party software is allowed (SRA, FASTQ, SAM and BAM format). The software is also able to use alternative genome assemblies hierarchically, thus accounting for epigenome changes in local populations.

Due to these unique features and its high-quality controls and flexibility, MethFlow is a convenient tool to study differential methylation among tissues, individuals, populations and physiopathological conditions.

[1] Karyn L. Sheaffer et al. Genes & Dev 28 (2014) 652-664.
[2] Duncan Sprout et al. Briefings in fuctional genomics 12 (2012) 174-192.
[3] Roadmaps Epigenetics Consortium et al. Nature 518 (2015) 317-330.
[4] Manu J Dubin eLife 4 (2015)
[5] Guillermo Barturen et al. F1000research 2 (2014) 217.

#059

Interactive web tool to manage sequencing data for the detection of viral insertion sites in gene therapy experiments

José M. Juanes[1], Joaquín Tárraga[2], Asunción Gallego[2], Ignacio Medina[2], Vicente Arnau[1], and Joaquín Dopazo[3]

[1]*Departamento de Informática, Escola Tècnica Superior d'Enginyeria, Universitat de València. Spain*
[2]*Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF). Spain*
[3]*Computational Genomics Department, Centro de Investigación Príncipe Felipe. Bioinformatics of Rare Diseases (BIER), Ciber de Enfermedades Raras (CIBERER). Functional Genomics Node, (INB, PRB2, ISCIII) at CIPF, Valencia, Spain.*

The possibility of integrating viral vectors to become a persistent part of the host genome makes them a crucial element of clinical gene therapy. However, viral integration has associated risks, such as the unintentional activation of oncogenes that can derive in cancers. Therefore, the analysis of integration sites of retroviral vectors is a crucial step in developing safer vectors for therapeutic use. Here we present ISMapper, a vector integration site analysis web server to analyse next-generation sequencing data for retroviral vector integration sites. Because it uses novel mapping algorithms, ISMapper is remarkably faster than previous available options and provides a useful interactive graphical interface to analyse the integration sites found in the genomic context.

ISMapper reads standard FASTQ or FASTA files containing reads corresponding to the insertion sites of the virus. These reads are mapped onto the reference human genome using BWA [1] or HPG-Align [2]. The results are presented using a graphical environment with a kariotype viewer (that provides a general perspective of the insertion sites) and a genome viewer implemented with GenomeMaps [3] (that provides a more detailed information about the insertion sites).

ISMapper can be found at: http://ismapper.babelomics.org.

[1] Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760.
[2] Tarraga, J., Arnau, V., Martinez, H., Moreno, R., Cazorla, D., Salavert-Torres, J., Blanquer-Espert, I., Dopazo, J. and Medina, I. (2014) Acceleration of short and long DNA read mapping without loss of accuracy using suffix array. Bioinformatics, 30,3396-3398.
[3] Medina, I., Salavert, F., Sanchez, R., de Maria, A., Alonso, R., Escobar, P., Bleda, M. and Dopazo, J. (2013) Genome Maps, a new generation genome browser. Nucleic Acids Res, 41, W41-46.

#060

TilingScan: an application for the identification of differentially expressed DNA regions in Tiling Microarray data.

José M. Juanes[1], Ana Miguel[2], José E. Pérez-Ortín[3], and Vicente Arnau[1]

[1]*Departamento de Informática, ETSE, Universidad de Valencia, Valencia, Spain. Spain*
[2]*Departamento de Bioquímica y Biología Molecular, Facultad de Biología, Universidad de Valencia, Valencia, Spain.*
[3]*Departamento de Bioquímica y Biología Molecular, Facultad de Biología, Universidad de Valencia, Valencia, Spain.*

Genomic technologies allow laboratories to produce large-scale data sets, either through the use of next-generation sequencing or microarray platforms. To explore these data sets and obtain maximum value from the data, researchers view their results alongside all the known features of a given reference genome. To study transcriptional changes that occur under a given condition, researchers search for regions of the genome that are differentially expressed between different experimental conditions. In order to identify these regions several algorithms have been developed over the years, along with some bioinformatic platforms that enable their use. However, currently available applications for comparative microarray analysis exclusively focus on changes in gene expression within known transcribed regions of predicted protein-coding genes, the changes that occur in non-predictable genetic elements, such as non-coding RNAs.

Here, we present a web application for the visualization of strand-specific tiling microarray or next-generation sequencing data that allows customized detection of differentially expressed regions all along the genome in an unspecific manner, that allows identification of all RNA sequences, predictable or not.

We have developed a version of the geometric moving average algorithm [1]. This algorithm scans the signal based in scalable and sliding windows and searches for over (or under) expressed regions of a minimum size of nucleotides flanked by neutral regions.
In addition, provided that microarray data signals often display noisy profiles, TilingScan enables the application of a smoothing algorithm, allowing signal noise removal prior to the search and clearer visualization of the data.

The application can be used to analyze NGS data provided that BAM files are converted into compatible files using freely available software.

[1] Basseville,M. and Nikiforof,I.V. (1993) Detection of abrupt changes: theory and application. Prentice Hall, Englewood Clifss, NJ.

#087

iGOP: an iterative algorithm to resolve inconsistencies in genome-scale metabolic models

Miguel Ponce de Leon[1], Juli Peretó[2], and Francisco Montero[3]

[1]*Departamento de Bioquímica y Biología Molecular. Facultad de Químicas. Universidad Complutense de Madrid. Spain*
[2]*Departament de Bioquímica i Biologia Molecular and Institute for Integrative Systems Biology I2SysBio (Universitat de València-CSIC), Valencia, Spain*
[3]*Departamento de Bioquímica y Biología Molecular. Facultad de Químicas. Universidad Complutense de Madrid. Spain*

In order to study the metabolic capabilities of organisms, the reconstruction of genome-scale metabolic models (GSM) has probed to be a powerful tool in systems biology. A genome-scale metabolic reconstruction draft is usually created from the genome annotation of an organism, which it is used to infer the enzymatic complement. Since, in general, genome annotations are incomplete, with missing genes, lots of hypothetical proteins with unknown function, as well as wrong annotated sequences, the networks inferred from such sources of information are inherently incomplete. Then, the missing enzymatic activities, as well the wrong inferred ones, will lead to inconsistencies in the network formulation. From a model point of view, the inconsistencies will manifest as deadend metabolites and blocked reactions. for these reason, reconstructing a metabolic model requires refinement steps, where the gaps are filled. Gaps can be detected and filled automatically using network-based methods that relies on the use of optimization techniques, to find candidate reactions to fill the gaps, from a universal model.

Most of the developed algorithms focus on finding and filling those gaps that prevent a model to predict biomass formation. Moreover, commonly used algorithms such gapfind/gapfill or fastcore were conceived to perform a global gap-filling, i.e. to solve all the gaps at once. However, in many cases some gaps are the consequence of wrong annotated genes, and thus the inconsistent elements should be pruned instead of filling the gaps. Thus, global gap-filling algorithms have the drawback that do not allow a proper classification of inconsistencies in such a way that an expert can decide whether a gap should be filled, or it is the consequence of a wrong annotation. Here we propose a method, named iGOP (iterative gapfill or prune), which combines the previously developed concept of unconnected modules (UM) with the fastcore algorithm.

The approach is an iterative procedure in which the resolution of each UM is conducted independently and thus, it allows to evaluate how many orphan reactions should be included in the model to restore the UM connectivity. for example, if the candidate reactions predicted to fill the gaps in a UM, outnumber the restored reactions, it may indicate that the UM is an artifact. The method was tested using a data set of 130 GSM reconstructed using The SEED Model. As a proof of concept, we applied iGOP to the histidine biosynthetic pathway, which was found inconsistently reconstructed in 20 models. The results allowed to classify and restore those cases were the pathway was likely to be coded by the organism, from those the inconsistent pathways most certainly represent an artifact, product of annotation errors.

#104

Synbiocraft: A Minecraft mod for simulating Synthetic Biology constructions

Alex Barberá-Mourelle[1], Daniel Pellicer-Roig[2], J. Alberto Conejero[3], and Diego Orzáez[4]

[1]*ETS Ingeniería Informática. Universitat Politècnica de València. Spain*
[2]*ETS Ingeniería Agronómica y del Medio Natural. Universitat Politècnica de València. Spain*
[3]*Instituto Universitario de Matemática Pura y Aplicada. Universitat Politècnica de València. Spain*
[4]*Instituto de Biología Molecular y Celular de Plantas, CSIC. Spain*

The idea of Minecraft is simple. Create your own constructions using blocks by moving them from one site to another in a 3D virtual world. This idea has catch more than 34 million players that are moving around, manipulating their space in order to build everything that comes out from their imagination. From an engineering point of view, Minecraft is just a huge 3D matrix where every cell can contain a lot of information. With a little knowledge of java language, the blocks inside the game can be modified; more kinds of blocks can be created, and the most important, more different traits can be implemented to those blocks. This allowed us to implement Synthetic Biology inside the game.

DNA pieces, machines, DNA ligation, Petri dishes… this powerful tool provides all possibilities to recreate the practices that can be done in a lab without spending material and without having to wait days in order to carry the experiment out. In fact, ligations always occur, no mistakes, no one opens your termocycler while the ligation is being carried out and plasmid insertions are always done correctly, there are always white colonies. By using a DNA database and BioBricks assembly method, we have reached the next level, we have recreated a represilator that was created by Elowitz & Liebler [1] and also the SEXY Plant project in iGEM [2]. But this can be expanded to all BioBricks database. We expect that, in the future, you will be able to replicate all the projects that have already been done in iGEM competition [3]. Further information concerning Synbiocraft in [4].

[1] M.B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. Nature 403 (2000) 335-338.
[2] Team Valencia_UPV. Sexy Plant. iGEM (2014)
http://2014.igem.org/Team:Valencia_UPV Last access on March 24th, 2016.
[3] International Genetically Engineered Machine (iGEM) Competition. The iGEM Competition is the premiere student competition http://igem.org/ Last access on March 24th, 2016.
[4] Team Valencia_UPV. AladDNA iGEM (2015)
http://2015.igem.org/Team:Valencia_UPV/Practices/Minecraft Last access on March 24th, 2016.

#109

Use of neuronal networks based on epigenomic maps for the prediction of transcriptional regulation via CRISPR/Cas9

Alex Barberá-Mourelle[1], J. Alberto Conejero[2], and Diego Orzáez[3]

[1]*ETS Ingeniería Informática. Universitat Politècnica de València. Spain*
[2]*Instituto Universitario de Matemática Pura y Aplicada. Universitat Politècnica de València. Spain*
[3]*Instituto de Biología Molecular y Celular de Plantas. CSIC. Spain*

Apart from its use as a tool for genomic edition, the CRISPR/Cas9 system can be used to turn on-off genes freely. This can be done by using mutated versions of Cas9 that have lost its nuclease activity, but not its power to join themselves to DNA, guided by specific gRNA's. Therefore, CRISPR/Cas9 system is enabled to design transcriptional regulators tailored to each gene. Their regulatory properties depend on the exact location of each Cas9 in the genome. In general, they are placed close to the codifying sequence of the gene under consideration.

On the one hand, the nucleosomes, in which the DNA is packed, determine the accessibility of the Cas9, and therefore its capacity to perform its regulatory activity, see for instance the recent work [1]. On the other hand, epigenomic maps of different organisms have appeared in the last years. They supply information of the DNA packaging degree at every point of the genome. This information can help to predict the result of locating a particular gRNA in the target gene, that will be of interest to determine the most successful gRNA's for every gene. Examples of epigenomic maps can already be retrieved from [2].

We are working in a model of neuronal network that predicts the genic regulation activity via CRISP/Cas9 at particular locations of the genome. This model will be based on epigenomic maps of Arabidopsis. The training dataset for the network will be given by experimental data of genic repression produced at the lab.

[1] M.A Horlbeck et al. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. Accepted to eLife (2016);10.7554/eLife.12677.
[2] Plant DNase I hypersenitive Sites (DHSs) Database http://plantdhs.org/ Last access, March 24th, 2016.

#110

Description and application of a pipeline for generating versions of genome-scale metabolic models from sequence information

Erik Zuchantke[1], Maria Siurana[2], David Fuente[3], Lenin Lemus[3], Röbbe Wünschiers[1], and Javier F. Urchueguía[3]

[1]*Experimental and Computational Biology Group, University of Mittweida. Germany*
[2]*Institute for Applied and Pure Mathematics (IUMPA). Spain*
[3]*Institute for The Applications of Advanced Information and Communication Technologies (ITACA), Universitat Politècnica de València, Valencia. Spain*

Genome-scale metabolic models (GSMM) are valuable tools for design and discovery of metabolic functions in organisms of interest. The process to obtain such a model starts with an automated reconstruction from the annotated genome sequence. However, it always requires a massive amount of manual work to curate and debug the resulting network. Thus, it takes time and effort to obtain a reliable accurate GSMM of a given organism based on its reference genome sequence. Often there appear other strains of interest related to the reference, from which only a non-annotated genome sequence is available. Therefore, we developed a pipeline that aims at the elucidation of how differences in sequence might affect the metabolic function of these strains, allowing the construction of models for the new strains based on the original one. This pipeline starts with sequence alignments, from which changes in genes involved in the metabolic model are identified. These genes are then translated to check changes in the functionality of the encoded proteins. From the results of this analysis, new versions of the metabolic model can be generated manually. We have applied this pipeline to obtain GSMM of different strains of Synechocystis sp. PCC 6803 based on their sequences and a detailed model previously reconstructed in our group [1,2]. This sequence-to-model pipeline used progressive mauve [3] for the alignment. To check if the mutations were already known, a web search is integrated to check information from CyanoBase [4] and UniProt [5] to get additional information about well-known proteins. To check if an exchange of amino acids is known for a specific protein, the entries of these proteins families from PFAM [6] were compared.

[1] A. Montagud et al. BMC Syst. Biol., 4 (2010) p. 156.
[2] A. Montagud et al. Biotechnol. J., 6 (2011) pp. 330–42.
[3] A. E. Darling et al. PLoS One, 5 (2010) p. e11147.
[4] Y. Nakamura et al. Nucleic Acids Res., 26 (1998) pp. 63–67.
[5] A. Bateman et al. Nucleic Acids Res., 43 (2015) pp. D204–212.
[6] R. D. Finn et al. Nucleic Acids Res., 42 (2014) pp. D222–230.

Scientific Session:

Biomedical Informatics: from Biomedical Bioinformatics to Mining Electronic Health Records (A).

#033

Inferring significant pathway regulators from the integration of multi-omic NGS data in Generalized Linear Models

Mónica Clemente-Císcar[1], Patricia Sebastián-León[1], Ana Conesa[1], and Sonia Tarazona[1]
[1]*Centro de Investigación Príncipe Felipe. Spain*

Understanding gene regulatory networks is an important goal in transcriptomics studies because it allows for the characterization of biomarkers and relevant pathways in diseases. Gene networks have been traditionally derived from gene expression data. However, the cost decreasing for NGS technologies has made it easier to obtain genome-wide multi-omic data for the same biological system, and therefore the generation of complex models for gene regulation that consider both transcriptional (e.g. transcription factors or epigenomic elements) and post-transcriptional regulation (e.g. microRNA or other non-coding RNA regulation) is now possible. Defining these models is still a challenge because issues such as apping among omics features, different noise levels in omics technologies or underpowered datasets must be faced. Hence, strategies and statistical methods to integrate multi-omic information in a common regulatory network are required.

We present MORE (MultiOmicsREgulation), a methodology to create complex gene regulatory models based on generalized linear regression (GLMs) to explaine gene expression as a function of its regulators and the experimental conditions. Including regulators in the models can result in unsolvable equations. To increase the model power while maintaining flexibility, we have implemented filters to exclude regulators with low variation, aggregation of regulators in case of multicollinearity or special stepwise variable selection procedures. MORE generates a different regulatory model per gene and provides both a gene-wise and global summary of the results, as well as a graphical representation. As far as we know, it is the first tool for integrating multi-omic data and experimental factors in order to investigate gene expression regulation.

We applied MORE to the huge collection of NGS data generated within the European STATegra project: time course data from a B-cell differentiation process in mouse under control and Ikaros-induction conditions. RNA-seq, miRNA-seq, DNase-seq and RRBS-seq data were used in this analysis to obtain regulatory programs for the 5,865 differentially expressed genes in this system. The GLM significant regulators were added to KEGG pathways to obtain multi-omic regulatory networks, which also allowed us to understand the regulatory cross-talk among the different pathways. We found that, in our system, the transcriptional regulation is more prevalent than post-transcriptional regulation.

#007

RD-Connect and identification of rare disease patients with similar genotype and phenotype combinations in other platforms through GA4GH Matchmaker Exchange

J. Protasio[1], O.J. Buske[2], Ma. Gonzalez[3], Rf. Acosta[3], D. Piscia[1], S. Laurie[1], A. Papakonstandinou[1], S. Zuchner[3], I. Gut[1], and S. Beltran[1]

[1]*Centro Nacional de Análisis Genómico (CNAG-CRG), Center for Genomic Regulation, Barcelona. Spain*
[2]*Department of Computer Science, University of Toronto, Toronto, Canada.*
[3]*The Genesis Project Inc., Miami, Florida. United States*

The increase in high-throughput genome sequencing and analysis has enabled important advances in rare disease research. These efforts, often scattered, have become a valuable source of information for many specific studies, as well as the rest of researchers in this field. RD-Connect (rd-connect.eu), an EU FP7 funded project under the auspices of the International Rare Diseases Research Consortium (IRDiRC), is building a platform to integrate clinical, biosample and –omics data from a huge number of patients. With the aim of sharing knowledge beyond the project, RD-Connect is part of the IRDiRC/GA4GH Matchmaker Exchange (MME, matchmakerexchange.org) project. The main goal of the MME is to identify similar patients (in terms of genotype and phenotype) in other rare disease platforms harbouring pre-filtered or full genomic information. The MME has a double challenge: defining a standard API, de facto, for genotype and phenotype data sharing, and the creation of a federated network that interconnects other platforms (e.g., Phenome-Central, Gene-Matcher, Café Variome, The Genesis Project, RD-Connect etc.). The first version of the MME API [1] allows researchers to look for patients with a similar phenotypic profile and/or overlap of manually selected candidate genes. The prototype of the second version of the API (led by RD-Connect and The Genesis Project with the support of the GA4GH Data Working Group) is extending the functionality in order to allow queries on unfiltered genomic data (e.g. full panels, whole exomes or genomes) to enable matching even when candidate genes have not been identified (termed 1-sided and 0-sided hypothesis matching). New components and filters enable more powerful and specific questions such as “Do you have any patients similar to one with phenotypes X, Y, Z and with one rare (allele frequency < 0.01), harmful (missense or stopgain) variant in NGLY1 or TTN?”. The new Genome component introduces filtering options such as pathogeneicity/deleteriousness score (e.g., CADD, SIFT), allele frequency (e.g., ExAC, 1000GP, ESP6500), gene annotation (e.g. one or more involved) or consequence (e.g. VEP, Jannovar). Matched results are securely returned with an internal identifier of patient and sorted by similarity (float number between zero -no similarities- and one - perfect match-) allowing the researcher to perform the query and to contact the contributor of the matched patient record.

#064

Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia

Simone Ecker[1], Vera Pancaldi[2], Daniel Rico[2], and Alfonso Valencia[2]

[1]*UCL Cancer Institue, University College London. United Kingdom*
[2]*Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO). Spain*

Background: Chronic lymphocytic leukemia (CLL) presents two subtypes which have drastically different clinical outcomes, IgVH mutated (M-CLL) and IgVH unmutated (U-CLL). So far, these two subtypes are not associated to clear differences in gene expression profiles. Interestingly, recent results have highlighted important roles for heterogeneity, both at the genetic and at the epigenetic level in CLL progression.

Methods: We analyzed gene expression data of two large cohorts of CLL patients and quantified expression variability across individuals to investigate differences between the two subtypes using different measures and statistical tests. Functional significance was explored by pathway enrichment and network analyses. Furthermore, we implemented a random forest approach based on expression variability to classify patients into disease subtypes.

Results: We found that U-CLL, the more aggressive type of the disease, shows significantly increased variability of gene expression across patients and that, overall, genes that show higher variability in the aggressive subtype are related to cell cycle, development and inter-cellular communication. These functions indicate a potential relation between gene expression variability and the faster progression of this CLL subtype. Finally, a classifier based on gene expression variability was able to correctly predict the disease subtype of CLL patients.

Conclusions: There are strong relations between gene expression variability and disease subtype linking significantly increased expression variability to phenotypes such as aggressiveness and resistance to therapy in CLL.

#066

Spanish Population-Specific Differences in Disease-Related Genetic Variation

Alicia Amadoz[1], Marta Bleda[1], Luz Garcia-Alonso[1], Alejandro Alemán[1], Francisco García-García[1], Juan A. Rodríguez[2], Josephine T. Daub[2], Gerard Muntané[2], Antonio Rueda[3], Alicia Vela-Boza[3], Francisco J. López-Domingo[3], Javier P. Florido[3], Pablo Arce[3], Arcadi Navarro[2], Salud Borrego[4], Javier Santoyo-López[5], Guillermo Antiñolo[3], and Joaquín Dopazo[1]

[1]*Centro de Investigación Príncipe Felipe (CIPF). Spain*
[2]*Universitat Pompeu Fabra. Spain*
[3]*Genomics and Bioinformatics Platform of Andalusia (GBPA). Spain*
[4]*Universidad de Sevilla. Spain*
[5]*University of Edinburgh. United Kingdom*

Recent results from large-scale genomic projects suggest that allele frequencies, which are highly relevant for medical purposes, differ considerably across different populations. The need for a detailed catalog of local variability motivated the whole-exome sequencing of 267 unrelated individuals [1], representative of the healthy Spanish population. Like in other studies, a considerable number of rare variants were found (almost one-third of the described variants). There were also relevant differences in allelic frequencies in polymorphic variants, including ~10,000 polymorphisms private to the Spanish population. The allelic frequencies of variants conferring susceptibility to complex diseases (including cancer, schizophrenia, Alzheimer disease, type 2 diabetes, and other pathologies) were overall similar to those of other populations. However, the trend is the opposite for variants linked to Mendelian and rare diseases (including several retinal degenerative dystrophies and cardiomyopathies) that show marked frequency differences between populations. This observation agrees with the fact that, while high-frequency variants and variants underlying complex diseases tend to be shared across populations [2], low-frequency alleles tend to be private [3]. Moreover, we observed a total of 121 variants affecting the binding sites of different drugs, suggesting an important role for local variability in population-specific drug resistances or adverse effects. In addition, our findings highlight the relevance of local variability to distinguish real disease associations from population-specific polymorphisms [1]. We have made available the Spanish population variant server that contains population frequency information for the complete list of 199,669 variant positions we found in the 267 healthy individuals (http://csvs.babelomics.org/).

[1] Dopazo et al. Molecular Biology and Evolution In press (2016)
doi:10.1093/molbev/msw005
[2] Marigorta et al. PLoS Genetics 9(6) (2013) e1003566.
doi:10.1371/journal.pgen.1003566
[3] Casals et al. PLoS Genetics 9(9) (2013) e1003815.
doi:10.1371/journal.pgen.1003815

#072

Leveraging text mining, expert curation and data integration to develop a database on psychiatric diseases and their genes

Alba Gutiérrez Sacristán[1], Àlex Bravo[1], Olga Valverde[2], Marta Torrens[3], Ferran Sanz[1], and Laura I. Furlong[1]

[1]*Research Group on Integrative Biomedical Informatics (GRIB)(IMIM-UPF). Spain*
[2]*Neurobiology of Behaviour Research Group (GRENEC), IMIM, DCEXS, UPF. Spain*
[3]*Institute of Neuropsychiatry and Addiction, Parc de Salut Mar, Universitat Autónoma de Barcelona, Barcelona, 08003, Spain*

During the last years there has been a growing research in psychiatric disorders' genetics, supporting the notion that most psychiatric disorders display a strong genetic component. However, there is still a limited understanding of the cellular and molecular mechanisms leading to psychiatric diseases, which has hampered the application of this wealth of knowledge into the clinical practice to improve diagnosis and treatment of psychiatric patients. This situation also applies to psychiatric comorbidities, which are a frequent problem in these patients. Some of the factors that explain the lack of understanding of psychiatric diseases etiology are the heterogeneity of the information about psychiatric disorders and its fragmentation into knowledge silos, and the lack of resources that collect this wealth of data, integrate them, and supply the information in an intuitive, open access manner to the community along with analysis tools. PsyGeNET (http://www.psygenet.org/) has been developed to fill this gap, by facilitating the access to the vast amount of information on the genetics of psychiatric diseases in a structured manner, providing a set of analysis and visualization tools. PsyGeNET is focused on mood disorders (e.g. depression and bipolar disorder), addiction to substances of abuse and schizophrenia.

In this communication we describe the process to update the PsyGeNET database, which involves i) extraction of gene-disease associations (GDAs) from the literature with state-of-the-art text mining approaches, ii) curation of the text-mined information by a team of experts in psychiatry and neuroscience, iii) integration with data gathered from other publicly available resources. BeFree, a text mining tool to extract gene-disease relationships, is used to identify genes associated to the psychiatric diseases of interest from a corpus of more than 1M publications. BeFree has a performance of 85% F-score for the identification of genes associated to diseases by exploiting morpho-syntactic features of the text. In addition, it normalizes the entities to standard biomedical ontologies and vocabularies. The text-mined data is then reviewed by a team of experts to validate the GDAs, following specific curation guidelines. Each expert is assigned a disease area according to her/his area of expertise. A web-based annotation tool was developed to assist the curation process. The tool supports a multi-user environment by user and password assignment. It displays the evidence that supports the association for each GDA to the curator. More specifically, it shows the sentences that support the association between the gene and the disease, highlighted (both the sentence and the entities involved in the association) in the context of the MEDLINE abstract. The curator has to validate the particular association based on the evidence of each publication, and select an exemplary sentence that states the association. We also describe the protocol designed to assign the curation tasks to the different experts and the method to assess the inter-annotator agreement. Finally, we describe the approach to integrate the expert-curated data with GDAs identified from other publicly available resources.

#095

The pan-cancer pathological regulatory landscape

Matias M. Falco[1], José Carbonell-Caballero[1], and Joaquín Dopazo[1]

[1] *Centro Investigación Principe Felipe. Spain*

In the last decades, cancer has increased its relevance and impact on society. This disease produces an uncontrolled cell proliferation, which is triggered by a dysfunction of signal transduction networks that regulate molecular communications and cellular processes. Studying these regulatory mechanisms is an essential task for understanding and developing effective therapies against cancer. Recently, a new approach for studying cancer and its molecular mechanisms has emerged, pan-cancer analysis. This analysis aims to examine the similarities and differences among the genomic alterations found across diverse tumor types, with the purpose of designing new therapies against cancer and being able to apply them to other similar tumor profiles.

Here we produced for the first time a comprehensive catalogue of transcription factor activities altered in several cancer types, a number of these significantly associated to patient's survival. In order to achieve it, genomic data was retrieved from different databases (Ensembl, ICGC y TCGA) and a tumor stage analysis and a survival analysis were performed. After a series of statistical analysis (survival analysis, differential expression analysis and GSEA), the results pointed that any of the studied transcription factors were not related with a certain tumor stage. In fact, the alteration of a transcription factor generally occurs, if so, in all stages. Also, a group of transcription factors, for each cancer, were also selected as predictors of a patient survival, and some of them were capable of influencing the survival function just by themselves.

Posters

#005

rexposome: A bioinformatic tool for characterizing multiple environmental factors and its association with different omics biomarkers and disease

Carles Hernandez-Ferrer[1], Martine Vrijheid[1], and Juan R. Gonzalez[1]

[1]*Centre for Research in Environmental Epidemiology. Spain*

Exposome encompasses all environmental factors from conception until old age. Due to the ever changing environment and habits, exposure to environmental contaminants is growing increasingly complex. The HELIX 'early-life exposome' approach involves combining all environmental hazards that mothers and children are exposed to, and linking this to the health, growth and development of the children.

The main objectives of HELIX project include the measurement of a wide range of chemicals and physical environmental hazards in food, consumer products, water, air, noise, and the build environment. Also to define a multi-pattern and individual exposure variability while determining a molecular profile and biological pathways associated with those multiple exposures. To this end, a Bioconductor package has been developed. The package incorporates functions for exploring exposome and its interaction with outcomes, its integration with different omic data as well as downstream analyses to facilitate biological insights though pathway analyses and retrieving data from public databases such as DisGeNET and Comparative Toxicogenomics Database.

The usefulness of the package will be illustrated by analysing data belonging to the HELIX's Spanish cohort where exposome, genome, transcriptome, methylome and proteome data is available joint with information about respiratory and neurocognitive outcomes.

#008

Meta-analysis in rare diseases: getting the most from scarce microarray experiments

Carlos Óscar S.-Sorzano[1], José María Carazo[1], Alberto Pascual-Montano[2], Mònica Franch[1], and Marta Martínez[1]

[1]*CNB-CSIC. Spain*
[2]*Perkin-Elmer España S.L.. Spain*

Rare diseases, also known as orphan diseases, present low prevalence among the general population. To increase the quality of their lives, patients depend on special efforts from the research community and Pharma Corporations, usually focused on wide spectrum diseases ensuring either high scientific impact or good returns on investment. Not surprisingly, public databases contain sparse and heterogeneous gene expression data from rare diseases, turning a difficult task to analyze them through a meta-analysis. Like most of orphan diseases, Neurofibromatosis is caused by a genetic deficiency. Alterations in the tumor suppressor gene neurofibromin drive to the most common phenotypic manifestation of the disease, Neurofibromatosis type 1. Patients may develop different anomalies in skin, eyes, skeleton, and cardiovascular, endocrine and nervous systems. In the peripheral nervous system, disorders typically manifest as benign neurofibromas (NF) that eventually may degenerate to malignant peripheral nerve sheath tumors (MPNST). In order to characterize the genetic determinants involving the transition from benign to malignant tissue, individual gene expression studies based on microarrays were carried out in recent years. We inspected GEO and ArrayExpress databases finding five studies with high quality raw data, four data sets from human and one from mouse. Trying to define a unique gene signature combining the differentially expressed genes extracted from the comparison MPNST vs. NF in the five different microarray platforms, we have designed a new method of meta-analysis useful for rare diseases. This method is based on a similar but more robust stringent normalized score than the previously proposed [1]. The score evaluates both the size effect and the significance level for each gene in each experiment. We calculated the final score for each gene adding the individual scores for each experiment, ignoring mouse data differing from human data, resulting in 10,896 genes ranked according to the final score. Only genes with absolute logFC median value higher than 1 and genes with consistent signs between the final score and the logFC median were included in the gene signature characterizing the malignant transformation, yielding a total of 1,576 genes, 851 up- and 725 down-regulated. In order to assess the homogenous contribution of each study to the final score, we computed the Bhattacharya distance for each gene. Low Bhattacharya distance values indicated similar contribution of the individual experiments to the final score and correlated with high scores. Finally, and as a new and valuable contribution of the method, gene signature was validated through Jaccard and Euclidean distance calculations. The distances between the determined gene signature and any individual study gene pattern corroborated the

power of this meta-analysis method averaging little and heterogeneous data. We also applied the method to culture cells and to the comparisons NF vs. Control and MPNST vs. Control. Through a detailed functional analysis and a hierarchical clustering, we characterized all gene signatures obtained, linked to the phenotypic manifestations of the Neurofibromatosis, and potentially useful as prognostic and predictive tools.

[1] Rasche et al. BMC Genomics 9 (2008) 310.

#010

comoRbidity: An R package to analyze comorbidities from clinical data

Alba Gutiérrez Sacristán[1], and Laura Ines Furlong[1]

[1]*Research Group on Integrative Biomedical Informatics, Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Department of Experimental and Health Sciences (DCEXS), Universitat Pompeu Fabra (UPF). Spain*

Clinical databases contain large amount of information about patient history. Using a limited number of data types, such as the age, gender and the patient's diagnosis, it provides us an opportunity to improve patient outcomes through research and the development of clinical decision support tools in the personalized medicine.

A clinical data analysis system is fundamental to understand and predict outcomes from past patient data. Here, we present a software that process massive amounts of healthcare data to identify comorbidity (coexistence of diseases in one patient) patterns in patient level data. Understanding the etiology of comorbidities has a great impact on the health status evolution, selection of appropriate treatments and health system costs, being a key aspect to identify new preventive and therapeutic strategies.

In this communication we describe the comoRbidity R package, a new tool for for studying disease comorbidity from clinical data. It lets the user having a global overview about a disease comorbidity, and analyse it in a clear and easily way. comoRbidity R package extract the statistically significant comorbidities according to age interval and gender population. The input of this software is the user's clinical dataset that should follow a simple tabulated format. Apart from comorbidity analysis, other tasks can be performed with comoRbidity package, such as the analysis of the population suffering a disease of interest. Moreover, the sex ratio parameter, that allows to see if the disease co-occurrence is equally likely for both genders or not, as well as the temporal direction analysis is assessed for the comorbidities identified in the analysis. A special focus is made on the results visualization, providing a variety of representation formats, such as networks, heatmaps or bar plots.

#015

Reactome web services and widgets for third-party integration

Antonio Fabregat[1], Konstantinos Sidiropoulos[1], Guilherme Viteri[1], Florian Korninger[1], Peter D'Eustachio[2], Lincoln Stein[3,4,5], and Henning Hermjakob[1,6]

[1]*European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory,*
Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[2]*NYU School of Medicine, New York, NY 10016, USA,*
[3]*Ontario Institute for Cancer Research, Toronto, ON M5G0A3, Canada*
[4]*Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA*
[5]*Department of Molecular Genetics, University of Toronto, Toronto, Canada*
[6]*National Center for Protein Sciences, Beijing, China*

Reactome (http://www.reactome.org) is a free, open-source, curated and peer-reviewed knowledge base of biomolecular pathways. It aims to provide intuitive bioinformatics tools for visualisation, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education. Thus, the mainstays of its software development are usability and responsiveness from the user’s point of view, likewise modularity and reusability from the developer’s side.

Reactome has developed web services and widgets (http://goo.gl/koRvhp) to facilitate integration in third-party software. One service allows access to the database and another performs enrichment and expression analysis as well as species comparison. Widgets for the Pathways Overview (http://goo.gl/pmeut1) and Pathway Diagrams (http://goo.gl/ntGiYG) are provided for JavaScript and GWT. Both widgets overlay the results of the Analysis Service (http://www.reactome.org/AnalysisService/) to help users find pathways of interest.

Data are automatically integrated from and to different resources by (1) cross-referencing them during the release process or (2) retrieving and visualising content on demand. A use case for the second scenario is the Molecular Interaction (MI) overlay (http://goo.gl/xsohW0), which is integrated in the diagram viewer widget. This is used in the Pathway Portal (http://www.reactome.org/PathwayBrowser) and made available as a stand-alone widget.

In the MI overlay, interactors are retrieved on demand from several resources available via the PSICQUIC service [1] and displayed as an overlay on the pathway diagram. Molecules in the diagram that have hits for the selected PSICQUIC resource have a summary icon on the top-right corner displaying the number of interactors. When this is selected, a maximum of ten interactors appear in a radial layout surrounding the molecule.

In summary, Reactome has facilitated data integration by providing easy-to-use services and reusable widgets. Several resources such as CTTV (www.targetvalidation.org/), ChEBI (https://www.ebi.ac.uk/chebi/), BluePrint (http://www.blueprint-epigenome.eu/), PRIDE (http://www.ebi.ac.uk/pride/archive/) and SCRIPPS (https://www.scripps.edu/) have integrated these services and widgets, which we intend to further improve by overlaying different data from other resources.

[1] Aranda, B. et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat Meth 8, 528-529 (2011)

Bioqueries: a collaborative environment to create, explore and share SPARQL queries in Life Sciences

María Jesús García-Godoy[1], Esteban López-Camacho[1], Ismael Navas-Delgado[1], and Jose F Aldana Montes[1]

[1]*Departamento de Lenguaje y Ciencias de la Computación, Universidad de Málaga, Andalucía Tech, Ada Byron Research Building E-29071 Málaga, Spain.*

The amount of information published is constantly growing in any area of interest, and Life Sciences is not an exception. However, these data can be published in any format, producing heterogeneity challenges when integrating them. Linked Data technology has emerged as a set of good practices based on W3C standards (i.e. Resource Data Framework, RDF) to publish and connect information. In Life Sciences, there are some platforms, such as Bio2RDF [1] and EBI RDF [2], producing datasets using Linked Data technologies. Despite using these standards, the information provided is accessible through a technological solution: SPARQL queries. Bioqueries aims to disseminate the use of these data repositories and connect end users in a collaborative environment. Bioqueries contributes to the consumption of Linked Data by enabling end users creating, exploring and sharing parameterized queries. These queries will target single repositories or several ones (federated SPARQL queries).

The main contributions of Bioqueries are summarized as follows: 1) a community of users creating, testing and sharing SPARQL queries; 2) a set of software modules to run queries and visualize their results (Relfinder [3]); and 3) the support for federated SPARQL queries (complex queries requiring joining data from more than one data source).

Bioqueries was initially populated with a set of SPARQL queries to serve as seed for the Bioqueries' community to grow. Since its initial publication [4], Bioqueries has grown up to 373 queries (59 federated queries) classified into different groups according to manual annotations, near 300 registered users and 68 endpoints. In this period, we have also include additional functionalities such as manual curation (public queries are verified and a quality sign added to those manually curated), improved administration support, additional formats to download query results and support for registering new data repositories. Bioqueries is freely available at http://bioqueries.uma.es.

As ongoing work, we are finishing the integration on a number of additional functionalities such as semi-automatic curation, visual support for designing queries and usage statistics.

[1] F Belleau, et al. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J. Biomed. Inform. 41(2008) 706-716.

[2] S Jupp, et al. The EBI RDF platform: Linked Open Data for the Life Sciences. Bioinformatics 30(2014) 1338-1339.

[3] Heim P, et al. RelFinder: revealing relationships in RDF knowledge bases. In: Proceedings of the 4th International Conference on Semantic and Media Technologies (SAMT). Graz, Austria. Lecture Notes in Computer Science, Springer, 5887 (2009)182-187.

[4] María Jesús García-Godoy et al. Sharing and executing linked data queries in a collaborative environment. Bioinformatics 29(13) (2013)1663-1670.

#025

kpath: An example of metabolic pathway data integration using Linked Data

Ismael Navas-Delgado[1], María Jesús García Godoy[1], Esteban López-Camacho[1], Maciej Rybinski[1], Armando Reyes-Palomares[2,3], Miguel A. Medina[2,3], and Jose F Aldana Montes[1]

[1]*Departamento de Lenguaje y Ciencias de la Computación, Universidad de Málaga, Andalucía Tech, Ada Byron Research Building E-29071 Málaga, Spain.*
[2]*Bioquímica, Facultad de Ciencias, Universidad de Málaga, Andalucía Tech, and IBIMA (Biomedical Research Institute of Malaga) E-29071 Málaga, Spain.*
[3]*CIBER de Enfermedades Raras (CIBERER) Málaga, Spain.*

Over the last few decades, the biological database community has been witnessing the increase on the amount of information available, which are available and accessible for scientists around the world. In the metabolism field, databases such as Kegg [1], Brenda [2], Reactome [3] and Biocyc [4] store information on metabolic pathways. However, these databases are only partially interlinked. This lack of interlinked resources causes that the level of interoperability between them is limited to independent searches to retrieve cross-database information on metabolism. This limitation restricts the use of more complex searches to discover new knowledge or relationships. Some studies [5] have also stress the importance of metabolism integration in the construction of complete human metabolic networks with complementary information from different databases.

Here we present kpath, a database that integrates information on metabolic pathways from different sources (Bio2RDF's Kegg, SwissProt and NCBI Taxonomy) using Linked Data. Thus, we provide a RDF database accessible via SPARQL queries. Additionally, kpath provides a navigational interface to ease the use of the integrated data by end users. This user interface includes three different applications: the Pathway Graphical Viewer, the Pathway Graphical Editor and the Relationship Search. The Pathway Graphical Viewer provides graphs with the participating biochemical reactions (i.e. metabolites, enzymes and genes) in a given pathway. Components related to other pathways are also shown allowing comparative analyses between pathways. The Pathway Graphical Editor includes the edition functionality of pathways, enabling users to customize pathways and save a local copy of their version. The Relationship Search tool enables the graphical browsing of relationships between pathway components (independently of their source). The public Linked Data repository can be queried at http://sparql.kpath.khaos.uma.es using the graph URI "www.khaos.uma.es/metabolic-pathways-app". The GUI providing navigational access to kpath database is available at http://browser.kpath.khaos.uma.es.

[1] Kanehisa, M., Goto, S., Kawashima, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acid Res. 42 (2014) D-199-D205.

[2] Schomburg, I., Chang, A., and Schomburg, D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 30 (2002) 47-49.

[3] Croft, D., O'Kelly, G., Wu, G. et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 39 (2011) D691–D697

[4] Caspi, R., Altman, T., Dreher, K. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 40 (2012) D742–D753.

[5] Stobbe, M., Houten, S., Jansen, G. et al. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. BMC Systems Biology 5 (2011) 165.

#032

Tools for the design and analysis of multi-omic experiments

Sonia Tarazona[1], Carlos Martínez-Mira[1], David Gómez-Cabrero[2], and Ana Conesa[1]

[1]*Centro de Investigacion Principe Felipe. Spain*
[2]*Karolinska Institute. Sweden*

Advances in massive sequencing technologies are favoring the proliferation of experiments studying several omics on the same biological system. Consequently, there is an increasing need of bioinformatics tools to help the researchers at different stages of these analyses. On the one hand, an important aspect that is often neglected is the experimental design. The different characteristics of each omic may lead to a different optimal sample size for each one of them, and it should be properly estimated when designing the experiment. On the other hand, there is also a lack of tools to validate integration methodologies. A popular approach for method validation is the utilization of synthetic data, but no publicly available algorithms exist so far that simulate different omic data types as well as the interaction among omic features. In this work, we propose two tools that give solution to the aforementioned problems: MultiPower strategy and MultiOmicSimulator.

The MultiPower tool estimates the optimal sample size of each omic in order to minimize the cost of the experiment. It assumes that the aim is identifying the omic features with significant changes between two populations. The estimation of the omics variability, the effect sizes to be detected, the minimum statistical power to be achieved, and the cost of generating each sample are considered by an integer programming model to render the optimal sample size per omic. We applied MultiPower to assess the suitability of the experimental design in STATegra project, thus illustrating that it can be useful for either knowing the limitations of the actual design or to design a new experiment. We observed that each omic had very different levels of variability and therefore the sample size to reach a minimum power in all cases varied from 3 in RNA-seq to 6 in metabolomics.

The MultiOmicSimulator algorithm simulates multi-omic data that serve for tuning or validating multi-omics integration strategies. It generates count data from sequencing technologies such as RNA-seq, miRNA-seq, ChIP-seq, DNase-seq or methylation, with a flexible experimental design (different experimental conditions, time series, etc.). More importantly, it also simulates the regulatory programs that relate gene expression with the rest of omic regulators (CpG sites, transcription factors, miRNAs, etc.) by modulating the values of the gene regulators accordingly with the regulatory effect of activation or inhibition on gene expression.

#036

3DBIONOTES: Unifying structure and biochemical annotations

J. Segura[1], D. Tabas-Madrid[1], R. Sanchez-Garcia[1], J. Cuenca-Alba[1], C.O.S. Sorzano[1], and J.M. Carazo[1]

[1]*GN7 of the Spanish National Institute for Bioinformatics (INB) and Biocomputing Unit, National Center of Biotechnology (CSIC), Madrid, Spain.*

With the advent of next generation sequencing methods, the amount of proteomic and genomic information is growing faster than ever. Several projects have been undertaken to annotate the genomes of most important organisms, including human. for example, the GENECODE project (1) seeks to enhance all human genes including protein-coding loci with alternatively splices variants, non-coding loci and pseudogenes. Another example is the 1000 genomes (2), a repository of human genetic variation, including SNPs and structural variants, and their haplotype contexts. These projects feed most relevant biological databases as UNIPROT (3) and ENSEMBL (4), extending the amount of available annotation for genes and proteins.

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 platform -3DBIONOSTES- that integrates proteomic, genomic and functional annotations with structural data, providing a unified and interactive view of the different sources of information. The main interface comprises three panels: the 3D viewer, the protein sequence viewer and the annotations panel. The three views are interactively connected and the different annotations can be displayed at sequence level, highlighting the amino acids of a selected annotation and at structural level mapping the corresponding residues into the protein structure. Current development offers an integrated access to EMDB/PDB (5), UNIPROT and ENSEMBL data, and further versions will include new sources of genomic, proteomic and interactomic data as well as a network viewer, where annotations will be also displayed at protein interaction levels.

(1) Harrow J et al. Genome research. 2012; 22;9;1760-74.
(2) 1000 Genomes Project Consortium et al. Nature. 2010; 10;28;467(7319):1061-73.
(3) UniProt Consortium et al. Nucl. Acids Res. 2015; 28;1;43;(D1): D204-D212.
(4) Cunningham F et al. Nucl. Acids Res. 2015; 28;1;43;(D1): D662-D669.
(5) Gutmanas A. et al. Nucl. Acids Res. 2014; 1;1;42;(D1): D285-D291.

#038

Characterization of RNA processing alterations in small cell lung cancer

Juan Luis Trincado Alonso[1], Jun Yokota[2], and Eduardo Eyras[1,3]

[1]*Universitat Pompeu Fabra, E08003 Barcelona, Spain*
[2]*Institute of Predictive and Personalized Medicine of Cancer (IMPPC), Badalona, Barcelona E08916, Spain*
[3]*Catalan Institution for Research and Advanced Studies, E08010 Barcelona, Spain*

Small cell lung cancer (SCLC) accounts for 15% of all lung cancers. Previous studies have shown high frequency of mutations in TP53 and RB1 [1], and amplification of MYC [1,2]. However, no targeted therapies have been approved for use in treatment of SCLC, contrary to other lung cancer types like adenocarcinoma. Accordingly, chemotherapy remains the only treatment, which is initially effective but is inexorably followed by rapid relapse in the majority of the patients. Understanding the molecular mechanisms underneath this disease is thus necessary for improving treatment. We have analyzed RNA-seq from 73 RNA-seq SCLC patient samples from [1] and characterized the transcriptomic changes between tumor and normal tissues. We have validated these changes on other 2 cohorts of 31 and 19 RNA-seq SCLC patient samples [3,4]. In order to identify those changes specific of SCLC, and to account for the fact that SCLC tumors have different cell type of origin than other lung tumors, we performed comparisons against more than 1000 non-small cell lung samples from The Cancer Genome Atlas and against neuroendocrine lung carcinoid tumors [5]. Additionally, using 71 WGS SCLC samples [1], we looked for somatic mutations disrupting intronic and exonic splicing regulatory motifs that could be responsible for these changes in the transcriptome. This is the largest analysis performed to date of RNA processing alterations and associated mutations in SCLC, which could lead to the uncovering of novel targets of therapy.

[1] Peifer et al. Nature Genetics (2012)
[2] George et al. Nature (2015)
[3] Rudin et al. Nature Genetics (2012)
[4] Iwakawa et al. Genes Chromosomes Cancer (2013)

#039

The BLUEPRINT Data Analysis Portal

José María Fernández González[1,2], Víctor de La Torre[1,2], Enrique Carrillo de Santa Pau[1], David Richardson[3], Romina Royo[4], Montserrat Puiggròs[4], Valentí Moncunill[4], Stamatina Fragkogianni[4], Laura Clarke[3], Paul Flicek[3], Daniel Rico[1], David Torrents[4], and Alfonso Valencia[1,4] on behalf The BLUEPRINT consortium

[1]*Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain*
[2]*Spanish National Bioinformatics Institute (INB), Spain*
[3]*European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK*
[4]*Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain*

The BLUEPRINT Data Analysis Portal (BDAP) is a tool to explore and query the data obtained with the analysis pipelines developed by the BLUEPRINT Consortium. The BDAP offers upper level analysis capacities to users with no need to re-analyze the raw data. It shows the differences between the epigenomes of hematopoietic cell types, analysed in the context of cell differentiation, thereby enabling the genomic regions, genes and pathways acting in the different hematopoietic lineages to be explored. The current version of the BDAP contains data from 439 samples from over 50 different cell types, obtained from healthy donors and in relation to 11 diseases. It includes WGBS (whole-genome bisulfite sequencing), DNaseI-Seq and RNA-Seq data, as well as that regarding seven histone modifications (H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K27me3, H3K9me3 or H2A.Zac in 2015-08 release). Through their gene names, pathways or coordinates, users can retrieve all the epigenomic and transcriptomic data available for their genomic region of interest. The data retrieved and high quality plots can be easily downloaded for downstream analysis and publication. The BDAP is an example of a new generation bioinformatics infrastructure based on non-relational databases and fast query mechanisms, and it can be accessed at http://blueprint-data.bsc.es

#042

Network-based epigenomic dataset integration in 3D: zooming in on RNA polymerase II

Vera Pancaldi[1], Enrique Carrillo-De-Santa-Pau[1], Biola Maria Javierre[2], Peter Fraser[2], Mikhail Spivakov[2], Alfonso Valencia[1], and Daniel Rico[1]

[1]*Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain*
[2]*Nuclear Dynamics Programme, The Babraham Institute, Cambridge, United Kingdom*

Despite the increasing number of studies mapping 3D chromatin interaction networks, there are surprisingly few network analysis performed on these datasets. Assortativity is a property that is widely used in social networks to detect whether similar people tend to connect with each other.

We propose a new related measure, the Chromatin feature Assortativity Score (ChAS), to identify epigenetic features that are associated to Chromatin Interaction Networks (CINs). We demonstrate the power of our methodology using promoter-centred interaction networks in mouse embryonic stem cells, generated with promoter-capture HiC experiments. We characterise each interacting chromatin fragment with the presence of epigenetic marks (histone and cytosine modifications as well as chromatin related protein binding peaks) and evaluate the relationship between these features and the network topology.

We confirm the importance of Polycomb proteins and associated marks in the 3D interaction network, extending previous findings. Moreover, we observe differences in the importance of various forms of RNA Polymerase II (RNAPII), which suggests a fundamental role for actively elongating RNAPII in the contacts between genes and their active enhancers.

Our method can be applied to any CIN generated through sequence based methods, as shown by our validation on two independent promoter-capture and 2 ChIA-PET datasets. It can also be extended to investigate any feature that can be assigned to the chromatin interacting fragments. We believe this new approach will aid in unravelling the complexities of this important layer of epigenomic regulation.

#061

Multisource and temporal variability in the distributions of biomedical data repositories: definition, methods and case studies

Carlos Sáez[1], Montserrat Robles[1], and Juan M García-Gómez[1]

[1] *Grupo de Informática Biomédica, Instituto ITACA, Universitat Politècnica de València.*

Biomedical data repositories (BDR) increasingly integrate data from diverse sources, from multiple health services to massive multicentre repositories. Besides, those health information systems are constantly updated with new data over time. In this manner, when reusing such data in managerial or clinical decision making, monitoring of healthcare indicators, clinical trials, or research studies, it is of upmost importance being aware of the possible variability that may exist among the data sources or over time. Specifically, we refer to the variability in data probability distributions, which could still be found even when interoperability or integration issues are satisfied, and may be caused by differences in data acquisition methods, protocols or health care policies, systematic or random errors during data input and management, demographic differences in populations, or even falsified data. When this probabilistic variability is unexpected or undesired, it affects to the overall data quality, and can lead to a suboptimal or even inaccurate data reuse [1-5].

Data quality in BDRs is a challenge for biomedical science professionals and researchers [6], who require complete and reliable data as well as evaluation tools and metrics [7]. Multisource and temporal data variability are sometimes considered in DQ controlled BDRs [7,8], the second specially in clinical trials [9,10]. The commonly used methods consist on describing basic data statistics, or comparing samples or time batches using classical statistical tests or process control methods [8,9,12,13]. However, these methods may prove inadequate to multi-modal, multi-type and multi-variate data [14] and result insufficient in Big Data environments, particularly due to large sample sizes [15,16].

Aiming to overcome the aforementioned problems and to provide new DQ metrics and exploratory tools, we developed a set of methods for multisource and temporal variability assessment [17,18]. They are based on an information theory and geometry probabilistic framework based on the normalized distances of data distributions among sources or over time. We will review these methods and show the results of their application to several case studies including: the Public Health Mortality and Cancer Registries of the Region of Valencia, Spain; a Spanish Breast Cancer and In-Vitro Fertilization datasets for predictive analytics, and a Perinatal Repository for healthcare monitoring.

[1] SL Krein *et al*. Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. Health services research 37 (2002): 1159–1180.

[2] D Blumentha *et al*. Information technology comes to medicine. The New England journal of medicine 356 (2007): 2527–2534.

[3] RJ Cruz-Correia *et al*. Data quality and integration issues in electronic health records. Information Discovery Electronic Health Records (2009) 55–95.

[4] McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. Carter KW, editor. PLoS ONE 8(3) (2013): e55811.

[5] C Sáez *et al*. Applying probabilistic temporal and multi-site data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories. Journal of the American Medical Informatics Association. *In press*.

[6] SL MacKenzie *et al*. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. Journal of the American Medical Informatics Association. 19 (2012):119–24.

[7] BL Massoudi *et al*. An informatics agenda for public health: summarized recommendations from the 2011 AMIA PHI Conference. Journal of the American Medical Informatics Association. 19(5) (2012):688–95.

[8] NG Weiskopf *et al*. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association. 20(1) (2013):144–51.

[9] KL Walker *et al*. Using the CER Hub to ensure data quality in a multi-institution smoking cessation study. Journal of the American Medical Informatics Association. 21(6) (2014):1129–35.

[10] JJ Gassman *et al*. Data quality assurance, monitoring, and reporting. Controlled Clinical Trials.16(2) (1995):104–36

[11] G Svolba *et al*. Statistical quality control in clinical trials. Controlled Clinical Trials. 20(6) (1999):519–30.

[12] MG Kahn *et al*. A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. Medical Care. 50 (2012):S21–9.

[13] Shewhart WA, Deming WE. Statistical method from the viewpoint of quality control. New York: Dover; 1986.

[14] C Sáez *et al*. Comparative Study of Probability Distribution Distances to Define a Metric for the Stability of Multi-source Biomedical Research Data. IEEE EMBC 2013:3226–3229.

[15] M Lin *et al*. Too Big to Fail: large samples and the p-value problem. Information Systems Research. 24(4) (2013):906–917.

[16] LG Halsey *et al*. The fickle P value generates irreproducible results. Nature Methods. 12(3) (2015):179–185.

[17] C Sáez *et al*. Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Data Mining and Knowledge Discovery. 29(4) (2015): 950-75.

[18] C Sáez *et al*. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. Statistical Methods in Medical Research. Published On-Line First (In Press).

#070

BiER collaborative projects

Francisco García-García[1], Alejandro Alemán[1], Francisco Salavert Torres[1], Mercedes Medina García[1], Julen Mendieta Esteban[2], and Joaquin Dopazo[1]

[1]*Computational Genomics. Príncipe Felipe Research Center. Spain*
[2]*Departamento de Informática, ETSE, Universidad de Valencia, Valencia, Spain*

BIER (Bioinformatics Platform for Rare Diseases) is a transversal working group whose mission is to provide bioinformatics and technological support to experimental CIBERER units.

During the last years the BiER platform has maintained an intense collaborative relationship with more than 25 CIBERER groups, mainly within the context of intramural sequencing projects, addressing transcriptomic and genomic studies (exomes and panels of genes). BiER has provided advice as well as technological and bioinformatics support in 70 projects of groups belonging to different CIBERER programs: Medical Genetics, Hereditary Metabolic Medicine, Endocrine Medicine, Pathology Neurosensorial, Neuromuscular and Mitochondrial Medicine and Related Syndromes Hereditary Cancer. We have also worked on the development of new methods of transcriptome analysis in the context of signaling pathways, functional meta-analysis and functional enrichment analysis for microRNAs.

We actively participate in collaboration intra-groups receiving more than 50 researchers in our unit and organizing the training activity "NGS course: from reads to candidate genes" which has been held during the last 4 years with an average of 25 attendants from different groups CIBERER per edition.

BiER has provided support and development of new versions of web tools for processing and analysis of genomic data: Babelomics (http://www.babelomics.org/) BiERapp (http://bierapp.babelomics.org/), TEAM (http://team.babelomics.org/), CIBERER Spanish Variant Server (http://csvs.babelomics.org/), among others.

The results of these analysis and bioinformatics developments have contributed to the discovery of 13 new disease genes in which 27 new mutations were identified and the identification of 36 new causal mutations in known disease genes, and also generated 56 collaborative scientific publications in the last three years
(http://bioinfo.cipf.es/publications) [1] [2] [3].

[1] Global Transcriptome Analysis of Primary Cerebrocortical Cells: Ientification of Genes Regulated by Triiodothyronine in Specific Cell Types. Gil-Ibañez P et al. Cereb. Cortex. 2015.

[2] Deregulation of key signaling pathways involved in oocyte maturation in FMR1 premutation carriers with Fragile X-associated primary ovarian insufficiency. Alvare-Mora MI et al. Gene. 2015.

[3] Differential Features Between Chronic Skin Inflammatory Diseases Revealed in Skin-Humanized Psoriasis an Atopic Dermatitis Mouse Models. Carretero M et al. J. Invest. Dermatol. 2015.

#081

Platform for Combining Electronic Health Record Data with Gene Research

Diego Bosca[1], Alberto Maldonado[1], and Montserrat Robles[1]

[1] *Universitat Politecnica de Valencia. Spain*

One of the key challenges in biomedical informatics is the lack of methodologies and tools to facilitate the reuse and mining of Electronic Health Records (EHR). If we focus on clinical research, interoperable EHRs provide a computable collection of fine-grained patient profiles, facilitating cohort-wide investigations and knowledge discovery on an unprecedented scale.

Ideally clinical research systems should be smoothly integrated with the computer tools that are routinely used by clinicians, in particular with the EHR. An important problem is the heterogeneity of clinical data sources, which may differ in the data models, schemas, naming conventions, and degree of detail used to represent similar data. Furthermore, clinical research systems very often require data at a level of abstraction higher than raw clinical data, leading to impedance mismatch problem.

We approach this problem by providing a methodology based on the combined use of EHR standards and web technologies. Modern EHR standards such as ISO13606 allow, on the one hand the representation of any EHR entry and on the other hand the meaningful definition of the clinical information models present in the systems.

Our approach includes a set of tools to deal with the interoperability of clinical research systems and EHRs based on archetypes. Archetypes are used to build a conceptual layer of the kind of a virtual health record (VHR) over the EHR whose contents need to be integrated and reuse, associating them with structural and terminology-based semantics.

Our methodology uses data transformations at different levels (such as normalization and abstraction) in order to provide adequate and useful data inputs to client applications. Our tooling (the LinkEHR platform) provides advanced data transformation tools to translate legacy EHR data into normalized EHR extracts and to perform data abstractions. The proposed methodology provides means, for instance, to access and combine EHR data (phenotype) with genetic risk prevention models and research (genotype). We exemplify this platform with legacy data transformations linked to publicly available EMBL-EBI web services such as Expression Atlas or QuickGO

#084

Web tools for the analysis of genomic data and the discovery new disease genes.

Alejandro Alemán[1], Fransico Salavert-Torres[1], Mercedes Medina[1], Francisco García-García[1], Jose Carbonell[1], Marta Hidalgo[1], Alicia Amadoz[1], Cankut Çubuk[1], Asunción Gallego[1], and Joaquín Dopazo[1]

[1]*Centro de Investigación Príncipe Felipe (CIPF). Spain*

The continuously increasing data production capability of sequencing technologies has shifted the bottleneck of the discovery process from the production to the data analysis phase. Our contribution to bridge the gap between genomic data production and its biological interpretation consists on the generation of a set of web-based tools used in different large-scale projects (MGP, CIBERER etc.). These include tools for gene prioritization, as BiERapp [1] (http://bierapp.babelomics.org/), which only during the last year was used for the analysis of more than 1000 exomes of patients of more than 70 different inherited pathologies, or TEAM [2] (http://team.babelomics.org/), designed for the efficient management of NGS targeted sequencing data for diagnostic. The CIBERER Spanish Variant Server (http://csvs.babelomics.org) is a public resource that provides information about the variability of the Spanish. We are in our fifth version of Babelomics [3] (http://babelomics.org), a general purpose platform for the analysis of Transcriptomics, Proteomics and Genomics data with advanced functional profiling, with more than 2000 registered users and about 2000 analysis carried out per month. and we are proud to present our new generation of precision medicine web tools that allow exploring disease mechanisms in the context of signaling pathways, hiPathia (http://hipathia.babelomics.org/), and the PathAct (http://pathact.babelomics.org/), an interactive framework to study of the consequences that KOs or over-expressions. More than 40,000 analyses were carried out in our tools during 2015.

[1] Alemán et al. Nucleic acids research, (2014), p. gku407.
[2] Alemán et al. Nucleic acids research, (2014), vol. 42, no W1, p. W83-W87.
[3] Alonso et al. Nucleic acids research, (2015), vol. 43, no W1, p. W117-W121.

#088

disgenet2r: An R package to explore the molecular underpinnings of human diseases

Alba Gutiérrez-Sacristán[1], Janet Piñero[1], Núria Queralt-Rosinach[1], and Laura I. Furlong[1]

[1]*Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Department of Experimental and Health Sciences, Universitat Pompeu Fabra. Spain*

DisGeNET [1] is a discovery platform designed to answer questions concerning the molecular mechanisms underlying human diseases. DisGeNET data can be explored using a suite of tools which includes a web interface, a Cytoscape plugin [2], and a SPARQL endpoint [3]. In this contribution, we present disgenet2r, an R package for exploring DisGeNET. disgenet2r contains a variety of functions for leveraging DisGeNET using the powerful visualization and statistical capabilities of the R environment. disgenet2r is specially designed to harness the large amount of information contained in DisGeNET, facilitating its analysis and interpretation. The package offers different types of visualization of DisGeNET data, such as heatmaps and networks, and it is especially well suited to explore the genetic basis of diseases as well as disease comorbidity. Furthermore, to allow answering more sophisticated research questions that need the interrogation of multiple, heterogeneous and disparate resources, the disgenet2r package permits benefiting of the potential of the Semantic Web technologies, without the need of special expertise in this area. This is achieved through a set of functions that connect DisGeNET with other resources present in the Linked Open Data, covering different information such as gene expression, drug activity, and biological pathways, just to mention a few examples. The disgenet2r package also expedites the integration of DisGeNET data with other R/Bioconductor packages, and allows the construction of complex bioinformatic workflows. We illustrate the functionality of disgenet2r through several use cases to show how the package can be applied to aid particular user's needs. The source code and documentation of disgenet2r package are available at https://bitbucket.org/albags/disgenet2r.

[1] J. Piñero, et al. Database (2015) 2015:bav028–bav028.
[2] A. Bauer-Mehren et al. Bioinformatics 26 (2010) 2924–6.
[3] N. Queralt-Rosinach et al. Cold Spring Harbor Labs Journals (2015) doi:10.1101/032961.

#089

Drug Enrichment Analysis as a tool for drug repositioning and functional analysis

Héctor Tejero[1], Jon Sánchez-Valle[1], Rafael Tabarés-Seisdedos[2], Alfonso Valencia[1], and Fátima Al-Shahrour[1]

[1]*Centro Nacional de Investigaciones Oncológicas. Spain*
[2]*Universidad de Valencia. Spain*

Some methods of drug repositioning are based on testing the similarity of a given drug with some characteristics (chemical structure or expression signature) of other of known indication or on looking for drugs with an opposite expression signature with respect to a given disease. In both cases is obtained a list of drugs ranked according to the similarity with the drug or the disease. In the last years, this kind of analysis have been favored by data projects as Cmap or LINCS.

We present deaR, an R package to carry out drug set enrichment analysis based on the pre-ranked GSEA or overrepresentation analysis of drug sets based on the hypergeometric test. Using three predefined manual curated collections of drug sets (Anatomical Therapeutic Classification, U.S. Pharmacopeia and KEGG Target-based classification) representing structurally and functionally related groups of drugs, deaR returns which drug sets are significantly enriched in an input list of drugs allowing the interpretation of the results and new hypothesis generation. deaR has been designed in order to interact with web resources as iLINCS or lincscloud.

Using deaR we show how an Alzheimer's Disease (AD) gene expression signature has similar effects to tyrosine kinase inhibitors transcriptional response signature, which could partially explain the observed inverse comorbidity between cancer and AD. We also show that HDAC inhibitors are significantly dissimilar to AD expression signature and, thus, could be used as therapeutic agents against AD.

deaR is implemented as a publicily available R package at
[https://github.com/htejero/deaR].

Scientific Session:

Methods matter & Reproducible Science (C).

Phylogeny & Evolutionary Bioinformatics (E).

#006

Breaking-Cas – Interactive design of guide RNAs for CRISPR-Cas experiments for ENSEMBL genomes

Juan Carlos Oliveros[1], Mònica Franch[1], Daniel Tabas[1], David San León[1], Lluis Montoliu[1], Pilar Cubas[1], and Florencio Pazos[1]

[1]*Centro Nacional de Biotecnología (CSIC). Spain*

The CRISPR/Cas9 technology is enabling targeted genome editing in multiple organisms with unprecedented accuracy and specificity by using RNA-guided nucleases. A critical point when planning a CRISPR/Cas9 experiment is the design of the guide RNA (gRNA) which directs the nuclease and associated machinery to the desired genomic location. This gRNA has to fulfil the requirements of the nuclease and lack homology with other genome sites that could lead to off-target effects. Here we present the Breaking-Cas system for the design of gRNAs for CRISPR/Cas9 experiments. The server has unique features not available in other tools, including the possibility of using all eukaryotic genomes available in ENSEMBL (currently more than 600), and can be freely accessed at:
[http://bioinfogp.cnb.csic.es/tools/breakingcas].

#078

Interplatform consistency using estimations of signaling pathway activity from transcriptomic data

Daniel Crespo[1], Marta Hidalgo[1], Alicia Amadoz[1], Cankut Çubuk[1], José Carbonell[1], and Joaquín Dopazo[1]

[1]*Centro de Investigación Príncipe Felipe. Spain*

The lack of direct correlation between gene expression platforms of different generation (microarray and RNAseq) limits data recycling such as the use of legacy microarray data to validate RNA-seq results [1]. To overcome this lack of correlation, we propose a pathway-level approach - that provides extra functional information by itself - rather of gene level. Using HiPathia method - a derivative method from Pathiways[2] - in toxicogenomics data sets such those obtained from SEQC-MAQC[3], we observe a nearly linear correlation between microarrays and RNAseq platforms. Our results open the door to the extensive reuse the valuable legacy of older platform data.

[1] Su et al., Genome Biology (2014) 15:523
[2] Sebastian-León P, et al. Nucleic Acids Research 41 (2013) 213-217.
[3] SEQC/MAQC-III Consortium. Nature Biotechnology 32 (2014) 903-914.

#100

RECONSTRUCTOR: a new tool that mixes reference guided and de novo assembly strategies to create a private genome sequence.

Irantzu Anzar[1], Andreu Paytuví[1], Riccardo Aiese Cigliano[1], and Walter Sanseverino[1]

[1]*Sequentia Biotech SL. Spain*

One important goal in genomics is to determine the genetic differences among individuals and to understand their relationship with their phenotype. Current methods for identifying genetic variations rely on either assembling whole genomes de novo, or identifying the changes between a species and a reference sequence. Only few attempts have been made to combine the two approaches.

We have developed Reconstructor, an automatic in silico approach, which aims at the generation of a full genome sequence of an individual starting from a reference genome and resequencing data. Reconstructor is based on Iterative Read Mapping and de novo assembly. The first step is based on SUPERW (Simply Unified Pair-End Read Workflow, doi:10.1093/molbev/msv152), a dynamic and fast tool that is used recursively to identify SNPs and structural variations (SVs) which are then added to a reference genome. The second step uses the unmapped reads to perform a de novo assembly. Through the use of paired-reads and split reads, Reconstructor integrates the de novo contigs into the genome in order to obtain a new individual-specific genome sequence. Finally, with a lift-over approach, gene annotations are transferred from the reference genome to the new obtained sequence.

Reconstructor, with its combined approach, is an ideal tool for detecting genomic variations (including SVs) and generating customized genomic sequences to be used in downstream analyses.

#021

Evolution of nested endosymbiosis in Tremblaya: bulls in a China shop

Rosario Gil[1], Sergio López-Madrigal[1,2], Ayelén Rojas[1], Carlos Vargas[1], Andres Moya[1,3], and Amparo Latorre[1,3]

[1]*ICBiBE, Universitat de Valencia. Spain*
[2]*Biologie Fonctionnelle, Insectes et Interactions, INSA-Lyon. France*
[3]*Área de Genómica y Salud, FISABIO – Salud Pública, València. Spain*

Insects with restricted diets live in obligate mutualistic symbiosis with endosymbiotic bacteria that provide them essential nutrients [1]. Mealybugs (Hemiptera: Pseudococcidae) from subfamilies *Phenacoccinae* and *Pseudococcinae* maintain an obligate symbiosis with betaproteobacteria of genus *Tremblaya*. The majority of phenacoccids present *T. phenacola* as a single endosymbiont, while in pseudococcids a nested symbiosis has been detected, in which each *T. princeps* cell contains several cells of a gammaproteobacterium [2]. The metagenomic analysis of the T. princeps - Moranella endobia consortium found in *Planococcus citri* [3-5], its comparison with the genome of *T. phenacola* PAVE (single endosymbiont of Phenacoccus avenae) [6], and the genetic analysis of endosymbiotic systems identified in other mealybugs from both subfamilies [2, 7], revealed the relationship between the presence of a nested endosymbiosis and the atypical genomic reduction in *T. princeps*. This includes the extreme loss of both informational and metabolic essential functions, or the presence of genomic duplications undergoing concerted evolution. The level of complementation is such that it would be more appropriate to consider the consortium as a new composite living form. The recent genome project of *T. phenacola* PPER, single endosymbiont of *Phenacoccus peruvianus*, provided new and striking surprises in the evolutionary history of *Tremblaya*.

[1] Moya et al. Nature Reviews Genetics 9 (2008) 218.
[2] López-Madrigal et al. Frontiers in Microbiology 6 (2015) 642.
[3] McCutcheon and von Dohlen. Current Biology 21 (2011) 1366.
[4] López-Madrigal et al. Journal of Bacteriology 193 (2011) 5587.
[5] López-Madrigal et al. BMC Microbiology 13 (2013) 74.
[6] Husnik et al. Cell 153 (2013) 1567.
[7] López-Madrigal et al. Frontiers in Microbiology 5 (2014) 449.

#079

Understanding the frequency distribution of human polymorphic inversions

Isaac Noguera[1], David Castellano[1], Sergi Villatoro[1], and Mario Cáceres[1]

[1]*Universitat Autònoma de Barcelona. Spain*

Chromosomal inversion polymorphism has been a paradigm in evolutionary biology. Since early on it was shown that inversions have adaptive effects in different organisms, but very little is known about the action of selection on inversions, especially in humans. A key evolutionary effect of inversions is that they suppress recombination as heterozygotes due to the generation of lethal unbalanced gametes. It is also known that there are two main generation mechanisms of inversions in humans (associated maybe to different mutation rates): mediated and non-mediated by inverted repeats (IRs). Thus, it is essential to assess the role of mutation, drift and selection in the population behavior of these two types of inversions. In this work, we took advantage of a large-scale genotyping effort of 44 inversions in 550 individuals from seven populations to carry out a global analysis of inversion frequency in humans. First, we built generalized linear mixed models to predict how the frequency varies according to the presence and size of IRs at their breakpoints, inversion positional effects, such as (distance to closest gene, number of captured genes, gene location et.), and features associated with the effect of inversions in recombination, such as inversion length and local recombination rate. Next, we compared the observed frequency against that predicted by our models in order to identify outliers and therefore inversion candidates to be under selection. Our models fit better the data when we distinguish inversions according to the presence (~30% of the variation explained) or absence (~50% of the variation explained) of IRs in their breakpoints, with inversion genetic length and inversion positional effects, respectively, as main and secondary factors affecting the variation in inversion frequencies. Inversion physical length is negatively correlated with both local recombination rate and inversion frequency (while controlling for each other and gene content). Moreover, inversions affecting coding regions are at significantly lower frequency than intergenic or intronic inversions. These results suggest that human polymorphic inversions are under strong purifying selection due to its role in promoting the generation of unbalanced gametes. Finally, we report two inversions that show clear signs of positive selection that deserve further molecular and phenotypic characterization.

#091

Prokaryote Protein Co-evolution points to Cell Response to Environment

David Juan[1], and Alfonso Valencia[1]

[1]*Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre – CNIO, Madrid, Spain*

Protein co-evolution has become an important source of information for protein structure and protein interaction prediction [1]. Recent methodological breakthroughs have shown that high quality detection of residue co-evolution, when possible, can provide very accurate predictions of inter-residue contacts [2, 3]. Similarly, protein co-evolution can provide helpful predictions of protein-protein interactions for prokaryote proteomes. Here, we present Global ContextMirror, the first methodology for protein interaction prediction based on co-evolution that takes advantage of the detection of direct interactions by correlation matrix inversion. Our results show that Global ContextMirror clearly outperforms existing methods based on co-evolution, such as MirrorTree [4] or ContextMirror [5]. We have performed a comparative analysis of the results of Global ContextMirror for 22 species from 8 different taxonomic groups. This analysis shows that lineage-specific co-evolution points to different functionally related protein pairs from the same macromolecular systems. In particular, our results indicate that protein co-evolution in different taxonomic groups concentrates on systems related to energy production, flagellum and transmembrane transport. As a whole, these molecular systems are devoted to cellular response to different environmental conditions, suggesting that protein co-evolution plays a key role on the evolutionary adaptation to changing environments.

[1] Juan, D., Pazos, F. & Valencia, A. Nature Reviews Genetics, 14 (2013), 249–261.
[2] Weigt, M et al. Proceedings of the National Academy of Sciences of the United States of America, 106 (2009), 67–72.
[3] Ovchinnikov, S., Kamisetty, H. & Baker, D. eLife, 3 (2014), e02030.
[4] Pazos, F. & Valencia, A. Protein engineering, 14 (2001), 609–614.
[5] Juan, D., Pazos, F. & Valencia, A. Proceedings of the National Academy of Sciences of the United States of America, 105 (2008), 934–939.

#116

DOMINO: Development of informative molecular markers IN non-model organisms using NGS data or pre-computed alignments

José F. Sánchez-Herrero[1], Cristina Frías-López[1], Sara Guirao-Rico[2], Joel Vizueta[1], Angel Blanco-García[1], Miquel A. Arnedo[3], Alejandro Sánchez-Gracia[1], and Julio Rozas[1]

[1]*Departament de Genètica, Microbiologia i Estadística; Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona. Spain*
[2]*Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Barcelona.Spain*
[3]*Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals; Institut de Recerca de La Biodiversitat (IRBio), Universitat de Barcelona. Spain*

The development of highly informative molecular markers is one of the most important challenges in current phylogenomics and genome-wide population genetics studies, especially in non-model organisms. In recent years, next generation sequencing (NGS) technologies have facilitated obtaining vast amounts of suitable markers with low associated costs. In this context, we have developed DOMINO, a bioinformatics tool for informative markers discovery and/or selection using either raw NGS data or pre-computed multiple sequence alignments (MSA) in various formats (including MSA from RAD data analyses). The software combines popular NGS tools with new developed utilities in a highly versatile pipeline that implements two marker development modules. The first module (marker identification module) allows identifying candidate markers after filtering, assembly and mapping NGS reads across a group of taxa. Alternatively, DOMINO can select the most informative markers (marker selection module) among a set of user-supplied MSAs (in various formats, e.g. fasta, phylip or PyRAD and STACKS output alignments). Under both modules, user can pre-define a large number of desired marker features, as the marker length, the minimum levels of variation or the minimum number of taxa included in the marker, among others.

We evaluated the performance of DOMINO in detecting informative markers by means of computer simulations. In particular, we simulated NGS reads emulating a reduced-representation library (RRL) experiment with four taxa and under a large set of parameter options, including sequencing platform, read depth, marker density, RRL fragment size and nucleotide divergence across taxa.

DOMINO is open cross-platform software that can be run either using an easy-to-use graphical user interface (GUI) or under an extended command-line version. Although the current version of DOMINO handles with data from most commonly used NGS technologies and file formats, the modular structure of the application allows its easy adaptation to the new sequencing platforms and NGS data.

Posters: Methods matter & Reproducible Science (C).

#016

Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package

Sonia Tarazona[1,2], Pedro Furió-Tarí[1], David Turrà[3], Antonio Di Pietro[3], María José Nueda[4], Alberto Ferrer[2], and Ana Conesa[1,3]

[1] *Genomics of Gene Expression Lab, Centro de Investigación Príncipe Felipe, Valencia, Spain*
[2] *Department of Applied Statistics, Operations Research and Quality, Universidad Politécnica de Valencia, Valencia, Spain*
[3] *Department of Genetics, Universidad de Córdoba, Córdoba, Spain*
[4] *Statistics and Operational Research Department, Universidad de Alicante, Alicante, Spain*
[5] *Microbiology and Cell Science Department, Institute for Food and Agricultural Sciences, University of Florida, USA*

As the use of RNA-seq has popularized, there is an increasing consciousness of the importance of experimental design, bias removal, accurate quantification and control of false positives for proper data analysis. We introduce the NOISeq R-package for quality control and analysis of count data. We show how the available diagnostic tools can be used to monitor quality issues, make pre-processing decisions and improve analysis. We demonstrate that the nonparametric NOISeqBIO efficiently controls false discoveries in experiments with biological replication and outperforms state-of-the-art methods. NOISeq is a comprehensive resource that meets current needs for robust data-aware analysis of RNA-seq differential expression.

#035

Dealing with management and integration of biological information according to Data Quality Dimensions

Ana León[1], Francisco Valverde[1], and Óscar Pastor[1]
[1]*PROS. Spain*

Genomics is a very complex and heterogeneous field, and each data repository was created to meet certain requirements. Due to this heterogeneity each repository stores the information from a specific domain following a particular schema. The lack of schema standards for representing the genomic information is a noteworthy issue when trying to integrate data from different repositories, hindering and slowing the bioinformaticians' daily work.

Performing an analysis of some well-known information repositories such as GenBank, ClinVar or Ensemble, we observed that most of the problems that we have found are not unique to Genomics domain. In fact, they have been already studied in the Information Systems development research field. This community highlights the need of sound conceptual models as a prior step to support data quality. Following this reasoning line, if high quality data (i.e. without "noise" values) are not properly organized it won't be straightforward to take advantage of them. and the other way around, a high quality Information System can be implemented but if the stored data have low quality, it won't be useful. .

There are some proposals [1][2] to assess data quality, using metrics classified by dimensions. Additionally, Moody [3] proposes the use of the ISO/IEC 9126 standard to assess the quality of the underlying conceptual models to represent the data. Taking these approaches as foundations, we have applied them to three genomic data repositories (RefSeq, ClinVar and Ensembl), to categorize the data problems found and to offer real examples from each one of them. Some of these quality problems are: different names for the same concept, redundant information, different structure formats to represent the same concept, lack of correct metadata, etc.

This analysis and classification highlights the need of having a sound conceptual model to represent the genomic information, so many of these problems would no longer exist or would be reduced greatly. In this regard, the specification of conceptual domain models and the development of bioinformatics software from them are presented as possible solutions.

[1] Wand, Y. et al. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86–95.
[2] Batini, C. et al. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 1–52.
[3] Moody, D. L. (2005). Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. Data & Knowledge Engineering, 55(3), 243–276.

#051

Interaction challenges in web collaborative environments supporting cognitive processes: An example from the genomics domain

Carlos Iñiguez[1], Oscar Pastor[2], and Francisco Valverde[2]

[1]*Escuela Politécnica Nacional. Ecuador*
[2]*Universitat Politècnica de València. Spain*

Genetic analysis is a domain that requires collaborative coordination between clinicians of several fields in order to identify patterns to justify or discard genetic anomalies. Information Technologies have expressed its great interest in this area, providing web-based tools to support the genetic analysis. As a result of genetic practice, many public and private data repositories with heterogeneous characteristics have been created around the world. Additionally, with the wide use of Big Data technologies, the analysis of huge amounts of data to transform it into knowledge has become a challenge. The analysis of data is still a manual procedure in which human cognitive curation and active collaboration of several stakeholders is required.

On the other hand, the increasing availability of mobile devices, tablet devices or smart screens makes a reality to propose novel interaction mechanisms in collaborative environments. As web technologies have become a standard in such devices, now it is possible to develop multi-device applications. However, current web interaction proposals do not sup-port both the collaborative and cognitive concerns in an efficient and effective way.

These novel collaborative scenarios are specifically relevant for the genomic analysis community in which geneticists, in order to perform an accurate diagnosis, have to work collaboratively identifying the suitable sources of genomic data, dealing simultaneously with thousands of variations, establishing hierarchies or priorities and sharing data.

The cognitive analysis perspective together with a web collaborative environment presents an attractive scenario in the context of the genome analysis domain. The contributions of this work are to describe the interaction challenges ahead, to analyze the underlying cognitive process of a genetic analysis and to design a web collaborative solution that meet both requirements.

#054

BioNetDB: A STORAGE ENGINE for THE INTEGRATION and ANALYSIS of BIOLOGICAL NETWORKS

Daniel Perez-Gil[1], Pedro Furio-Tari[2], Antonio Fabregat-Mundo[3], Felipe Javier Chaves-Martinez[1], Pablo Marin-Garcia[4], and Ignacio Medina[5]

[1]*INCLIVA. Spain*
[2]*Centro de Investigacion Principe Felipe. Spain*
[3]*EMBL-EBI. United Kingdom*
[4]*IMEGEN. Spain*
[5]*University of Cambridge. United Kingdom*

A key aim of biomedical research is to understand biological processes as a consequence of the complex interactions between their molecular components. These biological processes such as gene regulation, signal transduction, protein-protein interaction or metabolic pathways are often represented in the form of networks. Their modelling, analysis and visualization are essential for uncovering important properties of the underlying biological system. In addition, biological networks are not isolated, but there are a lot of different and complex interactions between them. Only a database that embraces relationships as a core aspect of its data model is able to store, process and query connections efficiently. Therefore, analyzing biological networks with graph concepts provides us the opportunity to describe a network of thousands of interacting components and gives us clues about how their organization and connection influence their function and dynamic responses.

Here we present BioNetDB, a storage engine to work with biological networks using a NoSQL graph database. BioNetDB integrates relevant biological network information from well-known data sources and provides access to this data through a comprehensive RESTful web services API or using the command line interface.

#071

CHEMDNER: named entity recognition of drugs and chemical names.

Martin Krallinger[1], Obdulia Rabal[2], Miguel Vazquez[1], Julen Oyarzabal[2], and Alfonso Valencia[1]

[1]*Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain*
[2]*Center for Applied Medical Research (CIMA) - University of Navarra. Spain*

There is an increasing need to facilitate automated access to information relevant for chemical compounds and drugs described in text, including scientific articles, patents or health agency reports. A number of recent efforts have implemented natural language processing (NLP) and text mining technologies for the chemical domain (ChemNLP or chemical text mining). Due to the lack of manually labeled Gold Standard datasets together with comprehensive annotation guidelines, both the implementation as well as the comparative assessment of ChemNLP technologies is opaque. Two key components for most chemical text mining technologies are the indexing of documents with chemicals (chemical document indexing - CDI) and finding the mentions of chemicals in text (chemical entity mention recognition - CEM). These two tasks formed part of the chemical compound and drug named entity recognition (CHEMDNER) task introduced at the fourth BioCreative challenge, a community effort to evaluate biomedical text mining applications. for this task, the CHEMDNER text corpus was constructed, consisting of 10,000 abstracts containing a total of 84,355 mentions of chemical compounds and drugs that have been manually labeled by domain experts following specific annotation guidelines. This corpus covers representative abstracts from major chemistry-related sub-disciplines such as medicinal chemistry, biochemistry, organic chemistry and toxicology.

A total of 27 teams -- 23 academic and 4 commercial ones, comprised of 87 researchers -- submitted results for this task. 26 of these teams provided submissions for the CEM subtask and 23 for the CDI subtask. Teams were provided with the manual annotations of 7,000 abstracts to implement and train their systems and then had to return predictions for the 3,000 test set abstracts during a short period of time. When comparing exact matches of the automated results against the manually labeled Gold Standard annotations, the best teams reached an F-score of 87.39% for the CEM task and of 88.20% for the CDI task. This can be regarded as a very competitive result when compared to the expected upper boundary, the agreement between to human annotators, at 91%. In general, the technologies used to detect chemicals and drugs by the teams included machine learning methods (particularly CRFs using a considerable range of different features), interaction of chemistry-related lexical resources and manual rules (e.g., to cover abbreviations, chemical formula or chemical identifiers). By promoting the availability of the software of the participating systems as well as through the release of the CHEMDNER corpus to enable implementation of new tools, this work fosters the development of text mining applications like the automatic extraction of biochemical reactions, toxicological properties of compounds, or the detection of associations between genes or mutations to drugs in the context pharmacogenomics.

#077

MetaGenyoPol: A web tool for meta-analysis of genetic association studies

Jordi Martorell Marugán[1], Daniel Toro Domínguez[1,2], Luis Javier Martinez González[3], Marta E. Alarcón Riquelme[2,4], and Pedro Carmona Sáez[1]

[1]*Bioinformatics Unit. Centre for Genomics and Oncological Research (GENYO), PTS, Granada. Spain*
[2]*Medical Genomics. Centre for Genomics and Oncological Research (GENYO), PTS, Granada. Spain*
[3]*Genomics Unit. Centre for Genomics and Oncological Research (GENYO), PTS, Granada. Spain*
[4]*Institute for Environmental Medicine, Karolinska Institutet. Stockholm, Sweden*

Genetic association studies estimate the statistical association between genetic variants and a given phenotype, usually complex diseases, in order to identify those genetic variants involved in susceptibility to that disease. Meta-analysis of genetic association studies integrates results from different studies in order to increase statistical power of the individual studies and to isolate real genetic associations to diseases from false positives. The amount of published meta-analysis has increased in the recent years. In 2011, there was a 64-fold increase in genetics-related meta-analysis compared to 1995 [1].

In this context, there are several methods to perform this type of meta-analysis [2]. However, nowadays there is not any user-friendly tool to perform such analysis and programming or scripting skills are required. In this work we have developed MetaGenyoPol, a web tool for genetic association studies meta-analysis. This web tool can be accessed by the URL http://hades.genyo.es/metagenyopol. MetaGenyoPol integrates different R packages including *HardyWeinberg* [3], meta [4] and metafor [5] to perform a complete genetic association meta-analysis. RStudio's Shiny [6] was used to implement the entire workflow into an interactive and easy-to-use web interface. We think that this can be a very useful tool for the research community that will allow researchers to perform a complete association meta-analysis in a user-friendly environment.

[1] Ioannidis et al. PloS ONE 8(6) (2013).
[2] Lee Annals of Laboratory Medicine 35(3) (2015) 283-287.
[3] Graffelman Journal of Statistical Software 64(3) (2015).
[4] Schwarzer R Package (2015).
[5] Viechtbauer Journal of Statistical Software 36(3) (2010) 1-48.
[6] Chang et al. R Package (2016).

#099

HiCloud: a new pipeline for a more inclusive way to analyze Hi-C data

Andreu Paytuví[1], Riccardo Aiese Cigliano[1], Walter Sanseverino[1], Irantzu Anzar[1], Aurora Ruiz-Herrera[2], and Covadonga Vara[2]

[1]*Sequentia Biotech SL. Spain*
[2]*Universitat Autònoma de Barcelona. Spain*

Attention for studying chromatin structure is rising due to its implications in gene regulation. Erez Lieberman-Aiden, et al. (2009) described Hi-C, a genome-wide approach that allows to identify any contact between a pair of loci by coupling proximity-based ligation (cross-linking with formaldehyde and chromatin digestion) with sequencing. Bioinformatics approaches to analyze Hi-C data consists of: 1.1) reads preprocessing to remove chimeric parts of a read followed by mapping or 1.2) iterative mapping where the unmapped reads, whose ends are trimmed 5-10nt each round, are realigned; 2) filtering read-pairs that are not close to any cleavage site from the restriction enzyme used at the chromatin digestion step; 3) filtering out possible artifacts -inward and outward pairs- and duplicates; 4) contact map creation, which is a square matrix of non-overlapping bins across the genome having each cell the number of pairs connecting 2 bins; 5) contact map normalization; 6) visual representation by heatmap. However, current tools for analyzing Hi-C data fail at the time to perform visual or statistical analysis of two or more contact maps. We are designing a fast and easy-to-use tool that will allow -using web technologies- a more interactive way to explore Hi-C results, including integrative analyses among replicates and conditions. Compared with HiCUP, which is one of the most used pipelines to analyze Hi-C data, our pipeline uses iterative mapping with STAR instead of reads preprocessing yielding faster mapping and higher efficiency. Our pipeline also sets distance thresholds between inward and outward mates for removing artifacts by learning from expected/observed distributions.

#118

Multiscale model to recapitulate breast cancer invasion phenotypes

Arnau Montagud[1,2,3], Margriet M. Palm[4], Vanessa Benhamo[1,5], Laurence Calzone[1,2,3], Andrei Zinovyev[1,2,3], Dirk Drasdo[4], Anne Vincent-Salomon[1,5], and Emmanuel Barillot[1,2,3]

[1]*Institut Curie, France*
[2]*INSERM U900, France*
[3]*Mines ParisTech, France*
[4]*INRIA, France*
[5]*INSERM U830, France*

Background: Understanding tumour invasion mechanisms is crucial to improve prognosis and develop new cancer treatment strategies, but this is hindered by the lack of understanding of detailed molecular determinants of this process and their interactions leading to different ways cancer cells invade the surrounding tissues. Tumour invasion varies from individual to collective cell movement or if proteases facilitate their migration.

Method: We devised a multi-scale mathematical model that incorporates information of a series of traits, cellular and environmental, that output in a set of invasion modes. For this, the model incorporates different intracellular and signalling pathways and the resulting influence network has been translated into a mathematical model using discrete logical modelling. We have taken advantage of continuous time Boolean modelling based on Markovian stochastic process defined on the model state transition graph to simulate intracellular molecular processes determining individual cellular properties. We have embedded this Boolean model in a lattice-free individual cell population model to cope with interaction between cells and microenvironment affecting cell properties, leading to various patterns of collective cell behaviour.

Results: The model is now tuned to recapitulate major breast cancer invasion phenotypes such as mesenchymal single cell invasion, solid strand multicellular invasion and bulk growth tumour. Present work is part of a collaborative effort to model tumour invasion in order to identify treatment strategies and to understand underlying properties of metastasis.

Posters: Phylogeny & Evolutionary Bioinformatics (E).

#004

Phylogenetic tree reconstruction through metabolic networks

Daniel Gamermann[1], J. Alberto Conejero[2], Arnau Montagud[3], and Cristina Loureiro[4]

[1] *UFRGS. Brazil*
[2] *IUMPA - Universitat Politècnica de València. Spain*
[3] *UPV. Spain*
[4] *ETS Ingeniería Agronómica y del Medio Natural. Universitat Politècnica de València. Spain*

More and more, sciences that traditionally have followed a more qualitative approach are adopting quantitative methodologies. A clear example is the development of biotechnology and the study of genetics and molecular biology. Many systems in biology seem to form interconnected networks of its parts and therefore, the application of graph theory to these systems comes naturally. We study the metabolic network of organisms and, in this work, we evaluate the application of a graph theoretical approach in order to reconstruct phylogenic trees based only on the abstract graph that represents an organism's metabolism. Phylogenic trees are a way to pictorially represent the evolutionary distances between different organisms. Nowadays, these trees are built based on alignment scores for homologous genetic sequences. We apply graph theoretical concepts in order to define a graph distance between two networks in a set and then, using the Kruskal algorithm we build a dendrogram that represent the evolutionary distances for a set of organisms. These network generated trees are then compared with different trees inferred by genetic and proteomic sequence alignments and the differences between all sets of trees are evaluated using the Robinson–Foulds metric.

#014

Ancestral Protein Reconstruction with Selection on Structural Stability with ProtASR

Miguel Arenas[1,2,3,4], Claudia Weber[5], David Liberles[4,5], and Ugo Bastolla[3]

[1]*Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP). Portugal*
[2]*Instituto de Investigação e Inovação em Saúde (i3S), University of Porto, Portugal*
[3]*Centre for Molecular Biology Severo Ochoa (CBMSO), Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain.*
[4]*Department of Molecular Biology, University of Wyoming, Laramie, USA.*
[5]*Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, USA.*

The reconstruction of ancestral proteins is essential to study past biological events and it is currently used in relevant areas such as biomedicine and biotechnology. Current available tools for ancestral sequence reconstruction (ASR) of proteins are commonly based on empirical amino acid substitution models of evolution that assume that all sites evolve under the same model. However, protein evolution can be highly heterogeneous where sites often evolve under different rates of change. Indeed, large improvements in evolutionary inferences have been derived from the consideration of structural constraints [1]. Here, we present a new ASR framework, called ProtASR, to infer ancestral protein sequences while accounting for selection on the stability of the protein structure. In a first step, ProtASR implements mean-field structurally constrained substitution models (MF) [2] that consider both unfolding and misfolding states to generate site-specific exchangeability matrices. We already showed that MF models outperform empirical amino acid substitution models, as well as other structurally constrained substitution models, through both maximum-likelihood and amino acid distribution across sites [2]. In a second step, ProtASR applies an adapted version of a well-established ML ASR computer program to infer ancestral proteins under MF models. ProtASR incorporates a large variety of options such as folding temperature, configurational entropies of unfolded and misfolded states and diverse MF and empirical models. We evaluated ProtASR by analyzing, through extensive computer simulations, the structural stability of ancestral sequences of diverse protein families. We found that the MF model generates ancestral proteins with folding stability closer to that of the ancestral simulated -and extant- proteins than those generated with the empirical model. We also analyzed the protein folding thermodynamics of the ancestral proteins derived from real data and we found that the evolution of protein folding stability in the studied families is heterogeneous over time and differs depending on genes and gene groups. ProtASR is freely available from https://github.com/miguelarenas/protasr, includes detailed documentation and ready-to-use examples. It runs in seconds/minutes depending on the protein length and alignment size.

[1] D.A. Liberles et al. Protein Sci 21 (2012) 769.
[2] M. Arenas et al. Molecular Biology and Evolution 32 (2015) 2195.

#058

A Colless-like balance index for multifurcating phylogenetic trees

Lucia Rotger[1], Francesc Rossello[1], and Arnau Mir[1]

[1]Dept. of Mathematics and Computer Science, University of the Balearic Islands, Palma de Mallorca. Spain.

The Colless index [1] is one of the most popular balance indices for rooted binary phylogenetic trees. It is defined as the sum, over all internal nodes *v* of the tree, of the absolute value of the difference between the number of descendant leaves of the two children of v. But, unlike other balance indices [2,3], it cannot be used as it stands on multifurcating phylogenetic trees, and the solutions proposed so far in the literature (like, for instance, Shao's and Sokal's proposal [4] of not taking into account the multifurcating nodes in the sum defining the index) can easily lead to nonsensical results.

In this work we define a general balance index for multifurcating rooted phylogenetic trees *T* as follows. Let f:$\mathbf{N}\rightarrow\mathbf{R}^+$ be a mapping, let D be any dissimilarity defined on vectors of real numbers, and, for every node *v*, let *deg(v)* be its out-degree. Now, for every node *v* in *T*, let δ_f *(v)* be the sum, over all its descendant nodes *w*, of *f(deg(w))* — this measures the size, relative to *f*, of the subtree rooted at *v* — and let $D_f(v)$ be the *D*-dissimilarity D(δ_f(v$_1$),..., δ_f(v$_k$)) of the values δ_f(v$_1$),..., δ_f(v$_k$), where v_1,...,v_m are the children of *v*. Let finally $Col_{D,f}(T)$ be the sum, over all nodes *v* of *T*, of their $D_f(v)$ values.

We show that, taking *f(x):=log(x+e)*, and as *D* the standard deviation or the mean deviation from the median, the resulting index Col$_{D,\log}$ has the following nice properties:

- When restricted to binary phylogenetic trees, it yields their Colless index up to a non-zero constant factor.
- It takes its minimum value, 0, exactly on the *fully symmetric trees* (which can be defined recursively as those trees whose subtrees rooted at the root's children are topologically isomorphic and fully symmetric, considering a single node as a fully symmetric tree)
- For every *n>0*, it takes its maximum value among the rooted trees with *n* leaves exactly at the caterpillars, which are usually considered as the most unbalanced trees.

In our presentation we shall also show that replacing *log(x+e)* by simpler functions *f* does not yield a well-behaved balance index, and we shall report on some numerical experiments with this index Col$_{D,\log}$ on TreeBase as well as on sets of phylogenetic trees generated under Ford's alpha-gamma distribution [5]

[1] D. H. Colless, Sys. Zool, 31 (1982), 100-104.

[2] M. J. Sackin, Sys. Zool, 21 (1972), 225-226.
[3] A. Mir, F. Rosselló, L. Rotger, Math. Biosc. 241 (2013), 125-136.
[4] K.T. Shao, R. Sokal, Sys. Zool, 39 (1990), 226-276.
[5] B. Chen, D. Ford, M. Winkel, Electron. J. Probab 14 (2009), 400-430.

#062

Evolutionary analysis of the miR-15/16 family and the need for a revised nomenclature

Aida Arcas[1], and M. Ángela Nieto[1]

[1]*Instituto de Neurociencias de Alicante (CSIC-UMH). San Juan de Alicante, Spain*

microRNAs (miRNAs) are small noncoding RNA molecules about 22 nt long that regulate gene expression post-transcriptionally by pairing with the 3’ UTR of target mRNAs.

The miR-15/16 family, characterized by the AGCAGC seed sequence, is involved in biological processes such as cell-fate determination, cell differentiation, development or disease [1].

The current annotation of the miRNA 15/16-family suggests the existence of many subfamilies, the presence of distinctive miRNAs in equivalent loci in different species, and the absence of certain miRNAs in various organisms and phyla. Although extensive experimental work has been conducted in particular miRNAs from the family, data on its evolutionary history are scarce [1].

Our objective is to study the emergence and evolution of the 15/16-family and to assess whether the observed differences in nomenclature and existence are actual variations among taxa.

We searched miRBase [2] and Rfam [3] for precursor miRNAs (pre-miRNAs) containing the AGCAGC seed sequence to retrieve already known and putative novel miR-15/16 family members.

Our analysis of the evolutionary history and of synteny of the 15/16-family shows that (i) their classification is simpler than that suggested by the current annotation, (ii) most of the clusters are widely conserved among vertebrates, (iii) there are potential pre-miRNAs in organisms where miRNAs have not yet been identified and (iv) confirms that both phylogenetic information on the pre-miRNAs sequences and synteny should be used as the base for classification and nomenclature of miRNAs.

We propose a revised nomenclature for the miR15/16 family that simplifies the existing one and truly reflects the phylogenetic relationships among the miRNAs, helping in the identification of orthologs and paralogs to better design and interpret intra- and interspecies functional studies.

[1] Finnerty JR et al. J Mol Biol. 2010 Sep 24;402(3):491-509.
[2] Kozomara A, Griffiths-Jones S. Nucleic Acids Res. 2014 Jan;42(Database issue):D68-73.
[3] Nawrocki EP et al. Nucleic Acids Res. 2015 Jan;43(Database issue):D130-7.

#068

A comprehensive description of the variation landscape in the genome of Drosophila melanogaster.

Sergi Hervás[1], Miquel Ràmia[1], and Antonio Barbadilla[1]

[1]*Genomics, Bioinformatics and Evolution Group, Institut de Biotecnologia I de Biomedicina (IBB) and Department de Genètica i Microbiologia, Universitat Autònoma de Barcelona. Spain*

High-throughput sequencing technologies are allowing the description of genome-wide variation patterns in the genomes of a growing number of organisms. However, we still lack a comprehensive understanding about the relative amount of different types of variation, their phenotypic effects, and how to properly detect and quantify distinct selection regimes acting on genomes. The Drosophila Genome Nexus project [1] is an incredible resource for population genomic analyses, consisting in more than 600 worldwide wild-derived Drosophila melanogaster genome sequences from 36 populations out of 17 countries. These genomes have been assembled using a common pipeline to reduce the potential bias due to methodological differences. Here, we present a complete characterization of the variation landscape in the genome of Drosophila melanogaster by analyzing the Drosophila Genome Nexus data. The diverse geographical origin of the samples allows us inferring the impact of demographic and environmental variables on the distinctive variation patterns along the genome. Specifically, we have estimated, at different zoom scales, summary measures of nucleotide diversity for both, nucleotide and structural variants (polymorphism and divergence metrics), linkage disequilibrium, historical recombination, and applied neutrality tests. A population genetics oriented web-browser based on JBrowse software is going to be implemented for easy and intuitive visualization, exploration and retrieval of the data [2,3]. This work provides both a global view of evolutionary forces shaping the genome variation patterns in D. melanogaster, and a novel reference tool to the research community for future population genomic studies.

[1] Lack et al., Genetics 199 (2015) 1229-1241.
[2] Skinner et al., Genome Research 19(9) (2009) 1630-1638.
[3] Ràmia et al., Bioinformatics 28 (2012) 595-596