VII Jornada de Bioinformàtica i Genòmica

Organitzada per:
Secció de Bioinformàtica i Biologia Computacional de la SCB
Secció de Genòmica i Proteòmica de la SCB
Associació Bioinformatics Barcelona – BIB
Spanish National Bioinformatics Institute (INB)

Amb el suport de:

PROGRAMA

Auditori edifici Vèrtex

Campus Nord UPC

Plaça Eusebi Güell 6, Barcelona

17 de desembre de 2019

COMITÈ ORGANITZADOR:	SUPORT:
Gabriel Valiente (UPC)	Mariàngels Gallego (SCB)
Mònica Bayés (CNAG-CRG)	Maite Sánchez (SCB)
Julio Rozas (UB)	Simón Perera (BIB)
Mario Cáceres (ICREA, UAB)	
Roderic Guigó (CRG-UPF)	
Ana Ripoll (UAB, BIB)	

8:30–9:15 Registration

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

Session I. Chair: Julio Rozas (UB)

9:30–10:15 **Invited Lecture: David Posada** (University of Vigo). Understanding tumor evolution within patients.

10:15–10:30 **Jon Lerga-Jaso** (UAB). Comprehensive analysis of the influence of human inversions on gene expression, epigenetic changes and phenotypic variation.

10:30–10:45 **Marta Coronado-Zamora** (UAB). Mapping natural selection through the Drosophila melanogaster life-cycle.

10:45–11:00 **Alejandro Sánchez-Gracia** (UB). Comparative genomics and transcriptomics in onychophorans and tardigrades shed light on the origin and evolution of arthropod chemosensory gene families.

11:00–11:30 Coffee Break

Session II. Chair: Mònica Bayés (CNAG-CRG)

11:30–11:45 **Virginia Díez-Obrero** (ICO-IDIBELL). Gene expression and splicing regulation in the colon helps to explain the genetic heritability of many complex traits and diseases.

11:45–12:00 **Jara Cárcel Márquez** (IR Sant Pau). Genomics and epigenomics: An integromic approach in stroke omics.

12:00–12:15 **Manuel Solís-Moruno** (UPF). Genetic load of somatic variants in primary immunodeficiency diseases.

12:15–13:00 **Invited Lecture: Mihaela Zavolan** (University of Basel). The role of RNA 3' end processing in defining mammalian cell types.

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

Session III. Chair: Ana Vivancos (VHIO)

14:30–14:45 **Miranda D. Stobbe** (CNAG-CRG). Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer.

14:45–15:00 **Claudia Arnedo-Pac** (IRB Barcelona). Oncodrive-CLUSTL: A sequence-based clustering method to identify cancer drivers.

15:00–15:15 **Anna Pedrola Gómez** (IDIBAPS). PCIG: A web-based application to infer immunological and genomic determinants across cancer types.

15:15–15:30 **Juan A. Subirana** (UPC). Non-coding satellites in bacteria: Their eventual role in nucleoid stabilization.

15:30–15:45 **Silvia Galan** (CNAG-CRG). Definition of "structural alphabets" for determining the relationship between structural patterns and genomic features.

15:45–16:00 **Mar González-Ramírez** (CRG). Histone modifications at enhancers are good predictors of gene expression.

16:00–16:30 Coffee Break

Session IV. Chair: Gabriel Valiente (UPC)

16:30–16:45 **Julien Lagarde** (CRG). An assessment of methods for third-generation long-read transcriptome sequencing.

16:45–17:00 **Carlos Ruiz Arenas** (ISGlobal). Historical recombination variability contributes to deciphering the genetic basis of phenotypic traits.

17:00–17:15 **Jordi Leno-Colorado** (CRAG). GSAW: A graphical interface package for population genomic analyses using high-throughput sequence data.

17:15–18:00 **Invited Lecture: Giorgio Valentini** (University of Milan). Machine learning for computational biology and precision medicine.

18:00–19:00 Poster viewing with authors and cocktail

19:00–19:15 Genes award to the best oral communication and poster and end of the symposium

Oral Presentations

Comprehensive analysis of the influence of human inversions on gene expression, epigenetic changes and phenotypic variation

Jon Lerga-Jaso, Marta Puig, Alejandra Delprat, Marina Laplana, Sergi Villatoro, Alba Vilella-Figuerola, Teresa Soos, Claudia Ramírez, Clara Vizuete, Roser Zaurín, Mario Cáceres

Presenting author: Jon Lerga-Jaso, Institut de Biotecnologia i de Biomedicina (IBB), Universitat Autònoma de Barcelona (UAB)

Structural variation contributes substantially to genetic diversity, but its association with complex traits and diseases deserves further characterisation. In particular, inversions have been often set aside due to the presence of repetitive sequences at their breakpoints and their balanced nature. Thanks to the unique methods developed by the InvFEST Project, we analysed the functional consequences of 111 polymorphic human inversions that have been accurately genotyped in a large number of individuals from diverse populations, representing the most complete resource to date for these variants. First, we measured their impact on gene expression using transcriptome data across different tissues or cell lines. Strikingly, half of the studied inversions act as lead eQTL or are in high linkage disequilibrium (LD) with top eQTLs, which suggests an enrichment of functional effects. Examples include inversions maintaining differentiated haplotypes, disrupting or reorganising gene structures, and creating novel fusion transcripts. Next, we identified 12 inversions modulating chromatin accessibility, DNA methylation and histone marks in LCLs, providing insights into their regulatory action through epigenetic patterns. Finally, we found that inversions show an excess of GWAS signals in their surrounding area, supporting their potential implication in human phenotypes. Moreover, 14 of them are in high LD with variants associated with neurological disorders, diabetes, anthropometric traits, atrial fibrillation or immune conditions, which constitutes a significant proportion taking into account the reduced number of inversions with highly linked SNPs. Interestingly, several inversions have clear effects at different levels, like HsInv0124, an inversion that regulates the expression of IFITM genes through histone modification patterns and has a pervasive effect on immune-related gene expression under infection, indicating that it may play an important role in antiviral defense. These findings highlight the functional impact of inversions on the human genome and reveal previously missing variants responsible for phenotype variability.

Mapping natural selection through the Drosophila melanogaster life-cycle

Marta Coronado-Zamora, Irepan Salvador-Martínez, David Castellano, Antonio Barbadilla, Isaac Salazar-Ciudad

Presenting author: Marta Coronado-Zamora, Universitat Autònoma de Barcelona (UAB)

In contrast to the genome of an organism, the transcriptome is a phenotype that varies during the lifetime and across different body parts. Studying a developmental transcriptome from a population genomic and spatio-temporal perspective is a promising approach to understand the genetic and developmental basis of the phenotypic change. We have carried out two different studies integrating the patterns of genomic diversity with multiomics layers across developmental time and space. In the first study, we give a global perspective on how natural selection acts during the whole life cycle of D. melanogaster. In the second study, we draw an exhaustive map of selection acting on the complete embryo anatomy of D. melanogaster. Taking all together, our results show that genes expressed in mid- and late-embryonic development stages exhibit the highest sequence conservation and the most complex genetic structure. Selective constraint is pervasive, particularly on the digestive and nervous systems. On the other hand, earlier stages of embryonic development are the most divergent, which seems to be due to the diminished efficiency of natural selection on maternal-effect genes. Additionally, genes expressed in these first stages have on average the shortest introns, probably due to the need for a rapid and efficient expression during the short cell cycles. The phenotypes that show evidence of adaptation are the immune and reproductive systems. Finally, genes that are expressed in one or a few different anatomical structures are younger and have higher rates of evolution, unlike genes that are expressed in all or almost all structures. The new developmental transcriptome of D. melanogaster at the single-cell level, allows increasing the resolution of our map to the cellular level, which is our next goal: detecting selection at each cellular expression profile.

Comparative genomics and transcriptomics in onychophorans and tardigrades shed light on the origin and evolution of arthropod chemosensory gene families

Joel Vizueta, Paula Escuer, Cristina Frías-Lopez, Sara Guirao-Rico, Georg Mayer, Julio Rozas, Alejandro Sánchez-Gracia

Presenting author: Alejandro Sánchez-Gracia, Universitat de Barcelona (UB)

Chemosensory perception is a fundamental biological process with great interest in basic and applied arthropod research. However, apart from insects, there is very little knowledge of the specific molecules involved in this system, and this is restricted to very few and phylogenetically uneven lineages. From an evolutionary point of view, onychophorans and tardigrades are of especial importance for studies on arthropods since they are the closest living relatives of this phylum, all three composing the so-called Panarthropoda clade. To get insights into the evolutionary origin and the diversification pattern of the arthropod chemosensory gene families, we have carried out a comparative genomics study across representative species of the three major Panarthropoda subclades, which included the sequencing of tissue specific transcriptomes and the implementation of a new bioinformatics solution to assist the comprehensive identification and annotation of new family members in these and other non-model species. The analysis uncovered key differences in the repertory of chemosensory genes across species, some of which could be related to the specific adaptations of tardigrades to survive in extreme environments. Besides, our results permitted to trace the origin and evolutionary y history of some of the major arthropod chemosenosry gene families and, more importantly, hypothesize the moment at which their members were co-opted to perform a chemosensory function.

Gene expression and splicing regulation in the colon helps to explain the genetic heritability of many complex traits and diseases

Virginia Díez-Obrero, Ferrán Moratalla-Navarro, Robert Carreras-Torres, Graham Casey, Víctor Moreno

Presenting author: Virginia Díez-Obrero, Institut Català d'Oncologia (ICO), Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)

The functional role of genomic regions identified in genome-wide association studies (GWAS) of complex traits and diseases is poorly understood and need to be characterized. It is hypothesized that these loci are enriched by variants with tissue-specific regulatory roles, i.e. Quantitative Trait Loci (QTLs). In this study, we analyzed a novel RNA-Seq dataset of about 200 colon tissue biopsies from healthy people to i) profile alternative splicing (AS) and gene expression, ii) assess their association with SNPs to identify colon-specific sQTLs and eQTLs, and iii) perform functional annotation and enrichment analysis at regulatory and trait-associated genomic regions. We profiled 27,226 AS events from seven distinct AS categories, among them 1,049 AS events in 676 genes are sQTLs associate with 898 unique SNPs. Besides, we identified 3,210 eQTLs comprising 3,148 unique SNPs associated with the expression of 3,203 genes. These s/eQTLs are mostly intronic, independent and enriched at open chromatin, transcription factor binding sites and active enhancer regions specific for normal and cancerous colon tissue. Then, we found that colon-specific s/eQTLs explain a proportion of genetic heritability of psychiatric and neurodegenerative diseases, inflammatory bowel diseases, colorectal cancer, cognitive and anthropometric traits, among others. We replicated our results using GTEx colon transverse data and made an interactive web resource to explore the results. Overall, our findings provide evidences of the regulation of alternative splicing and gene expression in the colon as potential underlying mechanisms of genetic susceptibility SNPs found at GWAS, and link colon tissue to traits and diseases not directly affecting it.

Genomics and epigenomics: An integromic approach in stroke omics

Jara Cárcel Márquez, Nuria Paz Torres-Águila, Elena Muiño, Caty Carrera, Natalia Cullell, Cristina Gallego-Fabrega, Israel Fernández-Cadenas

Presenting author: Jara Cárcel Márquez, Institut de Recerca Hospital de la Santa Creu i Sant Pau

In a cohort of 6.321 ischemic stroke (IS) patients and controls we performed a case/control GWAS. A locus located in GENE1 was genome-wide significant (top SNP p-value = 4.87·10-8). This result was replicated in lacunar stroke vs controls GWAS in two independent cohorts: a Spanish Cohort (n = 282) and an international cohort (n = 466.160). RTqPCR revealed overexpression of the RNAm of GENE1 (p-value = 0.04), besides the evaluation of the enzymatic activity of GENE1-protein showed an enhanced activity in patients of lacunar stroke compared to controls (p-value = 0.02). Using the significant (p-value < 5·10-8) and independent (LD < 0.8) SNPs, we performed a Summary-data-based Mendelian Randomization (SMR) using a public blood mQTLs database (n = 1980). The predicted causal CpG-sites were evaluated in a cohort of IS patients and controls (n = 357) of which we have blood methylation data. We performed generalized linear models adjusting covariates: smoking status, age and sex, for the analysis of IS, atherothrombotic, cardioembolic or lacunar vs control. 1 SNP was selected. The SMR analysis revealed 12 methylation sites to be causative of IS risk. We performed the validity test for 11 of them. None of the 11 CpG-sites were Bonferroni significant (p-value < 4.54·10-3) in IS vs control, atherothrombotic vs control and stroke vs control analysis. The 11 CpG-sites were significant (p-value < 0.05) in lacunar vs control analysis, 7 of them Bonferroni-corrected significant and had the correct direction predicted by SMR analysis. 2 of the validated CpG-sites were located in GENE1, 2 located in GENE2 and 3 in GENE3. Methylation of 7 CpG-sites is causally associated to the risk of suffering a lacunar stroke. These CpG-sites are located in GENE1, a previously described gene associated to lacunar stroke, and in GENE2 and GENE3 two novel genes associated to lacunar stroke risk. Further research is needed to understand the role of GENE1, GENE2 and GENE3 in lacunar stroke risk.

Genetic load of somatic variants in primary immunodeficiency diseases

Manuel Solís-Moruno, Anna Mensa-Vilaró, Laura Batlle-Masó, Irene Lobón, Tomàs Marquès-Bone, Juan I Aróstegui, Ferran Casals

Presenting author: Manuel Solís-Moruno, Universitat Pompeu Fabra (UPF)

There are increasing evidences for the contribution of somatic genetic variants to non-cancer diseases. However, their detection using massive parallel sequencing methods still presents important challenges. We performed whole-exome sequencing (WES) in 16 samples from people known to harbour a pathogenic somatic variant causing primary immunodeficiency. We first tested the ability of different variant callers to detect them. Those variant callers that allow establishing low frequency read thresholds were able to detect most of them, even at very low frequencies in the tissue (2-3%). Next, with the aim to measure the load of somatic coding variants in whole blood, we selected 460 candidates from the total detected. The selection criteria included stringent filters for sequencing/mapping quality and mutation features. The 460 variants were analysed by amplicon-based deep sequencing (mean coverage above 20,000X) and 1 or 2 somatic coding variants were validated per individual, except for one case with 11, clustering around 21% and 4% of variant allele frequencies. Our results show that somatic genetic variants at intermediate-high sequencing depths are detectable with specific variant callers. On the basis of the current ability to detect somatic variants and their involvement in the pathogenesis of Mendelian diseases, we recommend considering this particular mechanism when analysing genetic tests and, also, revisiting previous massive parallel sequencing data in patients with negative results. Moreover, we have described a low load of somatic coding variation in most of the analysed individuals in a validation experiment with 460 candidate variants from WES data.

Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer

Miranda D. Stobbe, Gian A. Thun, Andrea Diéguez-Docampo, Meritxell Oliva, Justin P. Whalley, Emanuele Raineri, Ivo G. Gut

Presenting author: Miranda D. Stobbe, Centre Nacional d'Anàlisi Genòmica (CNAG-CRG)

The sheer size of the human genome makes it improbable that identical somatic mutations at the exact same position are observed in multiple tumours solely by chance. The scarcity of cancer driver mutations also precludes positive selection as the sole explanation. Therefore, recurrent mutations may be highly informative of characteristics of mutational processes. To explore the potential, we use recurrence as a starting point to cluster over 2,500 whole genomes of a pan-cancer cohort. We describe each genome with 13 recurrence-based and 29 general mutational features. Using principal component analysis we reduce the dimensionality and create independent features. We apply hierarchical clustering to the first 18 principal components followed by k-means clustering. We show that the resulting 16 clusters capture clinically relevant cancer phenotypes. High levels of recurrent substitutions separate the clusters that we link to UV-light exposure and deregulated activity of POLE from the one representing defective mismatch repair, which shows high levels of recurrent insertions/deletions. Recurrence of both mutation types characterizes cancer genomes with somatic hypermutation of immunoglobulin genes and the cluster of genomes exposed to gastric acid. Low levels of recurrence are observed for the cluster where tobacco-smoke exposure induces mutagenesis and the one linked to increased activity of cytidine deaminases. Notably, the majority of substitutions are recurrent in a single tumour type, while recurrent insertions/deletions point to shared processes between tumour types. Recurrence also reveals susceptible sequence motifs, including `TT[C>A]TTT` and `AAC[T>G]T` for the POLE and 'gastric-acid exposure' clusters, respectively. Moreover, we refine knowledge of mutagenesis, including increased C/G deletion levels in general for lung tumours and specifically in midsize homopolymer sequence contexts for microsatellite instable tumours. Our findings are an important step towards the development of a generic cancer diagnostic test for clinical practice based on whole-genome sequencing that could replace multiple diagnostics currently in use.

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

Claudia Arnedo-Pac, Loris Mularoni, Ferran Muiños, Abel Gonzalez-Perez, Nuria Lopez-Bigas

Presenting author: Claudia Arnedo-Pac, Institut de Recerca Biomèdica (IRB Barcelona)

One of the key objectives in oncogenomics research is the identification of the genomic alterations that drive tumor development. Cancer driver genes have been computationally detected using methods based on signals of positive selection, which are acquired during tumor evolution. These signals –recurrence, clustering and high impact of somatic mutations– have been shown to be complementary in the detection of driver genes, thus highlighting the need of combining different up-to-date methods. However, these algorithms face the challenge of accurately calculating the expected mutation rates to detect positive selection. Interestingly, recent work showed that mutation rates can be modeled locally –region wise– using the probabilities of k-nucleotide context substitutions, avoiding genome-wide covariates and extending method's applicability to the non-coding regions of the genome. Until now, no clustering-based method using this model was available. Here we present OncodriveCLUSTL, a new sequence-based clustering algorithm to detect significant clustering signals across genomic regions. OncodriveCLUSTL is based on a local background model derived from the tri- or penta-nucleotide context substitutions extracted from the cancer cohort under analysis and can be applied to coding and non-coding regions from any species using whole exome and whole genome sequencing data. Our method is able to identify known clusters and bona-fide cancer drivers in coding regions, outperforming the existing OncodriveCLUST and complementing other methods based on different signals of positive selection. OncodriveCLUSTL also highlights different non-coding regions with significant clustering signals for further characterization.

PCIG: a web-based application to infer immunological and genomic determinants across cancer types

Anna Pedrola Gómez, Sebastià Franch Expósito, Roger Esteban Fabró, Laia Bassaganyas, Jordi Camps

Presenting author: Anna Pedrola Gómez, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS)

Cancer genomes play an important role in defining tumor immune features and therefore may determine the response to immunotherapy. Moreover, recent data suggest that genomic alterations modulate tumor immune phenotypes depending on their nature and histology, highlighting a complex interplay between the cancer genome and the immune system, which still remains poorly understood. The recently published Pan-Cancer Analysis of Whole Genomes (PCAWG) international project from The Cancer Genome Atlas (TCGA) Research Network has provided the most comprehensive landscape of genomic features in primary tumors. Here, we aimed at using the PCAWG data to investigate the interaction between genomic alterations and the tumor immune phenotype across different cancer types. To do so, an SQL database was used to join the different datasets provided by PCAWG, and a PanCancer ImmunoGenomics (PCIG) web-based application was developed using the Python-based open-source Django framework. PCIG provides an extensive immuno-genomic analysis using whole genome sequencing data available for more than 2,800 samples spanning up to 40 different cancer types and three levels of specimen classification: organ, tumor type and histology. We used these data to survey somatic mutations, copy number alterations (CNA), structural variants, gene expression quantifications and clinical classification, and to perform integrative analyses with additional estimated variables of high interest in immuno-genomics, such as tumor mutational load, tumor immune composition, and broad and focal CNA burdens. The ultimate goal of PCIG is to provide clinical research groups a user-friendly tool for visualizing the relationships between cancer genomic traits and immune-related phenotypes to better interpret tumor immunogenicity.

Non-coding satellites in bacteria: Their eventual role in nucleoid stabilization

Juan A. Subirana, Xavier Messeguer

Presenting author: Juan A. Subirana, Universitat Politècnica de Catalunya (UPC)

Tandem repeats (Satellites) are very abundant in many eukaryotic genomes. Prokaryotes have a very dense genome and are not expected to have satellites in intergenic regions. Occasionally they have been reported to be present in some prokaryotes, but here we present for the first time a complete general study. We describe the distribution and properties of satellites in the 12,333 bacterial genomes for which a complete sequence is available. We have detected 121,638 satellites. Our results demonstrate that the presence of satellites in the genome is not an exclusive feature of eukaryotes. Their distribution is very variable: for our study we have selected 1,241 genomes with 20 or more satellites. In particular we have searched for families of non-coding satellites in intergenic regions: we have only found them in 85 genomes in a few bacterial groups. Interestingly there are only three types of satellites, depending on their repeat sizes: 20-23, 40-44 or 52 nt. The presence of a limited number of sizes clearly indicates a function for these satellites. An intriguing feature is the constant size of the repeats in each genome, whereas their sequence shows little conservation. We conclude that they may be involved in the stabilization of the nucleoid through interaction with specific proteins. This situation is reminiscent of the alfa satellite found in the centromeres of eukaryotic chromosomes. Finally we will consider a particular, separate case: the spirochaete Leptospira interrogans. Its genome contains non-coding repeats of a different size, only found in this species; they may be related to the stabilization of its peculiar elongated nucleoid.

Definition of “structural alphabets” for determining the relationship between structural patterns and genomic features

Galan, Kai Kruse, Noelia Díaz, Juan M. Vaquerizas, Marc A. Marti-Renom

Presenting author: Silvia Galan, Centre Nacional d’Anàlisi Genòmica (CNAG-CRG)

Multiple studies have shown that dynamic changes (e.g. loops, Topological Associated Domains (TADs)) in the three-dimensional organisation of chromatin are associated with essential biological processes, such as transcription, replication and development. Nevertheless, the identification of these changes in an unsupervised and structure-specific manner is very challenging. We extended an algorithm called CHESS (Comparison of Hi-C Experiments using Structural Similarity) to be able to identify and classify differences between two Chromosome Conformation Capture (3C)-based experiments, such as Hi-C and Capture-C. The detection of the differences is not limited to specific structural features, such as TADs or loops, but instead is unsupervised and CHESS retrieves specifically the identified differential changes between two 3C-based maps. We developed a workflow to first obtain the significant different regions between datasets and then process the matrices to get the features that are specifically different. This was accomplished by applying different algorithms widely used in photography, such as binarization, closing morphology, smoothing and finally labelling, which allowed the detection and extraction of the differentially interacting regions.

Histone modifications at enhancers are good predictors of gene expression

Mar González-Ramírez, Enrique Blanco, Francesca Mugianesi, Cecilia Ballaré, Luciano Di Croce

Presenting author: Mar González-Ramírez, Centre de Regulació Genòmica (CRG)

Promoters and enhancers are genomic elements that co-ordinately regulate gene expression. Distinct configurations of histone modifications at these regulatory regions have been shown to be associated to transcriptional activation or repression. However, whereas the ChIP-seq signal of histone modifications at promoters is reported to be a good predictor of gene expression in different cellular contexts, this question has not been addressed for enhancers yet. Unfortunately, previous work applied to promoters cannot be used in the context of enhancers. Therefore, here we present a new methodology that we developed based on the combination of chromatin segmentation and linear regression to infer gene expression using epigenomic data at both, enhancers and promoters. The appropriate model system is one in which active and repressed regulatory regions can be identified. For this reason, we chose mouse Embryonic Stem Cells (mESC). It has been shown that mESCs contain active enhancers and repressed (poised) enhancers, which respectively co-ordinate with active promoters and repressed (bivalent) promoters to regulate gene expression. Moreover, since bivalent promoters and poised enhancers can either be activated or remain repressed during later stages of differentiation, we applied our framework to predict gene expression in differentiated cells. We have also compared the results obtained when using different approaches to assign enhancers to target genes. We found that histone modifications at enhancers, as well as at promoters, can predict gene expression of their target genes. HiC data was shown to be the best method to associate target genes to enhancers in our predictive model. Remarkably, bivalent promoters but also poised enhancers were found to be good predictors of gene expression in later stages of differentiation. Finally, we concluded that our strategy is universal since we were able to predict gene expression in a particular cell type using a model trained in another one.

An assessment of methods for third-generation long-read transcriptome sequencing

Julien Lagarde, Silvia Carbonell, Carme Arnan, Roderic Guigo

Presenting author: Julien Lagarde, Centre de Regulació Genòmica (CRG)

Long-read, third-generation sequencing (TGS) techniques such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences Single Molecule Real-Time sequencing (PacBio SMRT/IsoSeq), hold the promise to revolutionize transcriptomics. However, their relatively high sequencing error rate, together with their bias towards sequencing shorter molecules, have raised some doubts as to their suitability for gene and transcript annotation. In addition, the RNA/cDNA library preparation methods TGS relies upon often fail at representing transcripts faithfully in their complete, intact form. In an effort to improve the human and mouse GENCODE annotation, we have assessed the merits and weaknesses of competing TGS platforms on the one hand, and various library preparation protocols on the other. Our extensive assessment of the ONT and PacBio IsoSeq platforms reveals that the latter greatly outperforms the former in terms of transcript model accuracy. Additionally, by comparing various TGS-coupled RNA and cDNA library preparation methods (ONT direct RNA, 5'cap-trapping, SMARTer(c) and TeloPrime(c)), we show that 5'cap-trapping consistently provides the most reliable results. We also developed bioinformatic techniques to mitigate the shortcomings of TGS experimental workflows. These results show that complete, full-length transcriptome sequencing is within reach, with a wide range of potential applications in eukaryotic genome annotation and transcriptomics in general.

Historical recombination variability contributes to deciphering the genetic basis of phenotypic traits

Carlos Ruiz Arenas, Alejandro Cáceres, Marcos López, Dolors Pelegrí, Josefa Gonzalez, Juan R. Gonzalez

Presenting author: Carlos Ruiz Arenas, Institut de Salut Global de Barcelona (ISGlobal)

Recombination is a main source of genetic variability. However, the potential role of the variation generated by recombination in phenotypic traits, including diseases, remains unexplored as there is currently no method to infer chromosomal subpopulations based on recombination patterns differences. We developed recombClust, a method that uses SNP-phased data to detect differences in historic recombination in a chromosome population. We validated our method by performing simulations and by using real data to accurately predict the alleles of well known recombination modifiers, including common inversions in Drosophila melanogaster and human, and the chromosomes under selective pressure at the lactase locus in humans. We then applied recombClust to the complex human 1q21.1 region, where non-allelic homologous recombination produces deleterious phenotypes. We discovered and validated the presence of two different recombination histories in these regions that significantly associated with the differential expression of ANKRD35 in whole blood and that were in high linkage with variants previously associated with hypertension. By detecting differences in historic recombination, our method opens a way to assess the influence of recombination variation in phenotypic traits.

GSAW: a graphical interface package for population genomic analyses using high throughput sequence data

Jordi Leno-Colorado, Luca Ferretti, Emanuele Raineri, Giacomo Marmorini, Joan Jené, Gonzalo Vera, Sebastian Ramos-Onsins

Presenting author: Jordi Leno-Colorado, Centre de Recerca en Agrigenòmica (CRAG)

GSAW (Genome-Sequence Analysis using sliding Windows) is an innovative software package for the analysis of genome variability of multiple populations. The package can work with multiple format alignment files (gVCF, Tfasta, fasta, ms), annotation files (GFF, GTF) and additional filter files. A common problem using High Throughput Sequence data is to deal with a large quantity of missing data, which can substantially restrict the analysis of variability to few regions, or perform it with a reduced number of statistics. GSAW calculates an extended number of population and variability-related statistics and neutrality test accounting for missing data. In addition, optimal neutrality tests are for first time implemented. A number of file outputs are available, including simple text tables, Site Frequency Spectrum for one or several populations, output for other software (e.g., dadi), sliding window or fully extended formats. Finally, filtering options are available to obtain the desired output. GSAW provides a user-friendly graphical user interface (GUI), which helps to the user to include all the necessary functions and flags for the analysis of variability. The final result is a file with all the statistics obtained from the performed analysis, which is showed in a tab-format table in the interface. The interface is divided in three main sections: (i) pre-processing: where sequence format converters, annotation files and additional filtered options are managed to be ready for the analysis, (ii) analysis: calculation of all statistics given the desired options of the user, and including all sequence and annotation files, and (iii) post-processing: the obtained statistics are showed and can be filtered from the whole output file(s) and released in text tab separated tables. A window including all the selected commands are visualized at the bottom of the interface, ready to be copied and run on a terminal.

Posters

Poster 1

Integrative web interface for the visualization of complex planarian RNA-seq datasets

Sergio Castillo-Lara, Eudald Pascual-Carreras, Josep F. Abril

Presenting author: Sergio Castillo-Lara, Universitat de Barcelona (UB)

Over the last decade, a huge amount of transcriptomic and genomic data for the planaria Schmidtea mediterranea has been generated, thanks to its new genome assembly and especially to the advent of single-cell RNA-seq technologies. Developing interfaces for dealing with such data is of crucial interest to the research community. In order to bridge the gap between transcriptomic, genomic, and interactomic data we have developed PlanExp, a web-application to explore and visualize expression data from several experiments. PlanExp integrates tools for creating multiple interactive plots, tables, and visualizations; incorporating functional annotations performed both at the transcript and the genome level. Additionally, a prediction of gene regulatory networks has been performed in order to aid researchers to understand the complex planarian biology, and these predictions have been incorporated in the application, together with a full network editor and expression mapper powered by cytoscape.js.

Poster 2

The role of chromatin-associated proteins in genome topology

Francesca Mugianesi, Luciano Di Croce, Marc A. Mart-Renom

Presenting author: Francesca Mugianesi, Centre Nacional d'Anàlisi Genòmica (CNAG-CRG)

During embryonic development, proper orchestration of gene expression programs is accompanied by precise epigenetic and topological rearrangements of chromatin, tightly regulated in both space and time. Unfortunately, the link between gene expression, epigenetic modifications and genome architecture is still poorly understood. To get a unified vision of epigenetic and topological features of the genome, we aim to determine how chromatin-associated proteins contribute to genome architecture by the development and application of 3DepINT, a new computational method that integrates chromatin structure data from Hi-C interaction matrices and epigenetic data form ChIP-seq profiles of chromatin factors. Explicitly, we define an epigenetic coefficient for each pair of chromatin loci, weighted on their three-dimensional physical interaction in the cell nucleus. Thanks to this, we automatically detect loops and three-dimensional (3D) clusters associated to a comprehensive set of proteins and epigenetic marks, to characterize their role in genome structure. We also plan to test whether the folding of the genome allows for combinations of proteins and chromatin modifications that could contribute to gene regulation. Investigating dynamic changes of 3D clusters of chromatin factors during mouse embryonic stem cell differentiation in quantitative and genome-wide manner will help deciphering the complex relationship between genome function, its 3D architecture and the epigenome.

Poster 3

Bijective encoding of proteins in a scalable distributed deep learning framework

Angela Lopez-del Rio, Maria Jesus Martin, Alexandre Perera-Lluna, Rabie Saidi

Presenting author: Angela Lopez-del Rio, Universitat Politècnica de Catalunya (UPC)

Deep learning protein-based prediction models have gained great popularity in recent years. For these models, protein sequences are usually encoded into feature vectors. However, these encoding features are generally aggregative and not bijective, or require sequences to be alignable, thus decreasing the generalisation capability of the models. The use of raw amino acid sequences as models input is now gaining popularity. Padding is usually applied to get different length proteins to be within the same dimension, but little is known on how this addition could affect to the model performance. On the other hand, state-of-the-art Deep Learning models are not yet taking advantage of big data frameworks and distributed computation. Although there have been some approaches towards this integration, there are still no stable solutions. Overcoming this gap is crucial for getting the maximal potential out the growing public biological datasets. In this work, we build an scalable Deep Learning model by integrating big data and deep learning frameworks. We then analyse different protein bijective encodings in a protein function prediction problem and study the impact that the padding has on the performance of the model. Our results provide good practices on distributed computing protein-based deep learning models.

Poster 4

Predicting the DNA-binding preferences of C2H2-ZF proteins combining structural and experimental data

Alberto Meseguer, Filip Arman, Ruben Molina, Oriol Fornés Crespo, Jaume Bonet, Baldo Oliva

Presenting author: Alberto Meseguer, Universitat Pompeu Fabra (UPF)

Cis2 His2 zinc finger (C2H2-ZF) proteins are the largest family of transcription factors in human and higher metazoans. They are involved in many biological processes as well as diseases and they provide a platform for genome editing. However, the DNA-binding preferences of many of these proteins remain unknown. We developed a computational method to predict these DNA-binding preferences. Our method combines structural data coming from PDB structures of C2H2-ZF proteins with results from bacterial one-hybrid (B1H) experiments to compute statistical potentials scoring functions. These scoring functions can be applied to compute theoretical position weight matrices (PWMs) taking as input the structure of the C2H2-ZF protein of interest. We have validated our method by: 1) differentiating binding from non-binding events between C2H2-ZF proteins and DNA binding sites; 2) predicting PWMs for individual zinc fingers and comparing them to PWMs obtained from B1H experiments; and 3) predicting PWMs for entire C2H2-ZF proteins and comparing them to their corresponding experimental PWMs from the JASPAR database. Further, we applied our method to predict the DNA-binding preferences of CTCF.

Poster 5

An ensemble learning approach for modeling the systems biology of drug-induced injury in human liver

Joaquim Aguirre-Plans, Terezinha Souza, Janet Piñero, Giulia Callegaro, Steven J. Kunnen, Narcis Fernandez-Fuentes, Laura I. Furlong, Baldo Oliva, Emre Guney

Presenting author: Joaquim Aguirre-Plans, Universitat Pompeu Fabra (UPF)

Drug-induced liver injury (DILI) has a relatively high incidence rate, estimated to affect around 20 in 100,000 inhabitants worldwide each year. Many drugs ranging from pain killers to anti-tuberculous treatments can cause DILI. Despite DILI being one of the leading causes of acute liver failure, the pathophysiology of DILI is poorly understood and pinpointing the toxicity of compounds in human liver remains non-trivial. Accordingly, several methods have been proposed to predict the hepatotoxicity of compounds. Among these, machine learning models trained using drug estructural features have shown a good performance. Furthermore, the incorporation of gene- and pathway-level signatures from transcriptomics data has shown a high predictive accuracy using Deep Neural Networks. In this work, to predict DILI, we investigated combining gene expression data from the Connectivity Map (CMap), target binding information and chemical similarity of drugs upon drug treatment into ensemble learning methods using random forest classifiers and gradient boosting machines.

Poster 6

Quality control of sequencing files deposited at the European Genome Phenome Archive (EGA)

Dietmar Fernandez Orth, Aina Jene Cortada, Claudia Vasallo Vega, Babita Singh, Jordi Rambla de Argila

Presenting author: Jordi Rambla de Argila, Centre de Regulació Genòmica (CRG)

“A picture is worth a thousand words.” This adage has been extensively used to define how a complex idea can be conveyed with just a single image, that is how a simple graph can show large amounts of information. The European Genome-phenome Archive (EGA), as a repository of sequencing, variation array and phenotypic data for biomedical research projects, fully supports that proverb. One of the main purposes of EGA is allowing scientists and clinicians to get useful data available for their own analyses at a glance. Thus, EGA is making an effort to facilitate such exploration and is currently implementing some initial quality control (QC) tools for all sequencing data (aligned BAM) and Variant Call Format (VCF) stored files. Raw data (fastq) QC is also analysed by using the FastQC tool. Within the File QC Portal users can visualize several graphs and hence are allowed to check the main characteristics of the file and get an overall idea about its quality and reusability before downloading these. As an example, information regarding the proportion of mapped reads, duplicates, quality distribution can be checked, with an explanation on how to interpret each graph. Moreover, summary statistics and non-sensitive graphs are included, as well. Finally, with the QC Portal we aim to provide universal access to the main characteristics of the data through an interactive and intuitive graphical user interface.

Poster 7

Long-read based assembly and synteny analysis of a reference Drosophila subobscura genome reveals signatures of structural evolution driven by inversions recombination-suppression effects

Charikleia Karageorgiou, Rosa Tarrío, Francisco Rodríguez-Trelles

Presenting author: Charikleia Karageorgiou, Grup de Genòmica, Bioinformàtica i Biologia Evolutiva (GGBE), Universitat Autònoma de Barcelona (UAB)

Drosophila subobscura has long been a central model for evolutionary research on chromosomal inversion polymorphisms. Yet, research using this system has been limited due to the lack of a reference genome. Here we used PacBio long-read technology, together with the available wealth of genetic marker information, to assemble and annotate a high-quality nuclear and complete mitochondrial genome for the species. A highly-contiguous 129 Mb-long nuclear genome was obtained, consisting of six pseudochromosomes corresponding to the six chromosomes of a female haploid set, along with a complete 15,764 bp-long mitogenome. Additionally, we provide an account of their numbers and distributions of codifying and repetitive content. All 12 identified paracentric fixed inversions differences, with some associated duplications, but no evidence of direct gene disruptions by the breakpoints. Between lineages, inversion fixation rates were 10 times higher in continental D. subobscura than in the two small oceanic-island endemics D. guanche and D. madeirensis. Within D. subobscura, we found contrasting ratios of chromosomal divergence to polymorphism between the A sex chromosome and the autosomes. Our findings generally support genome structure evolution in this species being driven indirectly, through the inversions' recombination-suppression effects in maintaining sets of adaptive alleles together in the face of gene flow. The resources developed will serve to further establish the subobscura subgroup as model for comparative genomics and evolutionary indicator of global change.

Poster 8

Point mutations under positive selection in tumours

Ferran Muiños, Francisco Martinez-Jimenez, Oriol Pich, Abel Gonzalez-Perez, Nuria Lopez-Bigas

Presenting author: Ferran Muiños, Institut de Recerca Biomèdica (IRB Barcelona)

Somatic cells can accumulate thousands of somatic mutations during tumour development, a natural selection process whereby the tumour cell population undergoes a sequence of selective sweeps. But not all observed mutations are tumorigenic, even in genes known to drive tumorigenesis. Here we address the problem of distinguishing between tumorigenic and passenger mutations in driver genes, with a view towards understanding how the interplay between positive selection and background mutagenesis shapes the mutational landscape in tumours. We present a framework that aims to identify likely tumorigenic point mutations by learning gene and tissue specific models from a large cohort of more than 27,000 tumours comprising 61 cancer types (`https://intogen.org`). Aberrant mutation patterns and intrinsic characteristics of mutation sites at DNA or protein level can be used to ascertain the mechanisms of tumorigenesis. We retrieved a collection of mutational features from driver discovery outputs and public databases. Then we trained a classifier (boostDM) to score the chances that any given SNV is involved in tumorigenesis. Furthermore, we exploit the model's internal architecture to infer feature explanations for each individual prediction. Both the predictions and explanations provided by boostDM are used to conduct a site-specific driver potential analysis for all observed mutations in our dataset. We examine the relationship between predicted tumorigenic mutations and site-specific mutation rates explained by the background mutational processes of the tumour.

Poster 9

Brain transcriptomic profiling reveals common patterns across neurological and neuropsychiatric disorders

Iman Sadeghi, Emilio Palumbo, Manuel Munoz, Silvia Pérez-Lluch, Juan Domingo Gispert, Roderic Guigo, Natalia Vilor-Tejedor

Presenting author: Iman Sadeghi, Centre de Regulació Genòmica (CRG)

Neurological and neuropsychiatric disorders (NPDs) are multifactorial, polygenic and complex behavioral phenotypes caused by complicated brain abnormalities that share similar symptoms. Several large efforts have been carried out to identify the causal genes for a large number of NPDs. However, the underlying molecular pathogenesis is yet to be known. We used transcriptomic profiling of 2608 brain samples from eight groups of patients with Alzheimer's disease (AD), Parkinson's disease (PD), Progressive Supranuclear Palsy, Pathological Aging, Autism Spectrum Disorder (ASD), Schizophrenia (Scz), Major Depressive Disorder, and Bipolar Disorder-in comparison with 2019 brain samples from matched control subjects, to investigate cross-disease shared molecular signatures. Moreover, we examined the wide characterization of cortical regions across disorders, and explored cell-type specific patterns for each disorder, highlighting brain regions with common and specific transcriptome changes across diseases. The top transcriptome similarities were observed between AD-PD, Scz-ASD, ASD-PD and also between other phenotypes showing common patterns across neurological and psychiatric disorders. In addition, cortical specific comparisons revealed shared transcriptome across different NPDs. Using co-expression network analysis, we also identified fourteen gene modules differentially expressed which demonstrated expression specificity for eight brain cell types including neurons, astrocytes and oligodendrocytes. These in-depth analyses highlight the overlap and unique molecular structure of brain phenotypes.

Poster 10

Rational design of protein dynamics

Joan Planas-Iglesias, Gaspar Pinto, Andrea Schenkmayerova, David Bednar, Jiri Damborsky

Presenting author: Joan Planas-Iglesias, Loschmidt Laboratories, Masaryk University, Czech Republic

A key aspect to understand the function and evolution of proteins is deciphering their structural interactions, restraints, and dynamics. This is a particularly valid for enzymes, in which structure-function-dynamics relationships are particularly constrained. Apprising structural fluctuations on proteins is still challenging due to intrinsic technical limitations of experimental methods, and yet computational techniques can help surmounting these hindrances. Rational protein design aims to exploit the structure-function relationships for tailoring different aspects of enzymatic activity. Due to their lesser evolutionary constraints and distance to the catalytic centre, recent design efforts have specifically targeted loops – particularly dynamic aperiodic regions flanked by regular secondary structures. However, loop design approaches still rely more on empirical sampling than on rational design, hinting the need for wider quantitative knowledge about loops flexibility. A particularly challenging task in loop design is transferring a desired property between two proteins by means of loop grafting. We have recently shown that a successful activity transfer via loop transplantation requires of a precise geometric overlay of the target structure and meeting dynamical requirements for the engineered property. To address this problem we are developing a computational framework to assess loops flexibility, to compare their geometry and dynamics on different proteins and to propose viable solutions for loop grafting. Our newly developed strategy will be applicable to a wide range of protein families.

Poster 11

CNVfilteR: Identification of false positives generated by CNV calling tools from NGS data

José Marcos Moreno-Cabrera, Bernat Gel, Jesús Del Valle, Eli Castellanos, Lidia Feliubadaló, Eduard Serra, Gabriel Capellà, Conxi Lázaro

Presenting author: José Marcos Moreno-Cabrera, Institut Germans Trias i Pujol (IGTP), Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)

Germline copy number variants (CNVs) are one of the genetic causes of multiple hereditary diseases. Several tools for germline CNV detection from next-generation sequencing (NGS) data have been published. However, available benchmarks show that all CNV calling tools produce false positives. We developed CNVfilteR, an R package for identifying false positives generated by CNV calling tools from NGS data. To achieve this, CNVfilteR uses the germline single nucleotide variant calls (SNVs) frequently obtained in germline NGS pipelines. A scoring model based on fuzzy logic is used to identify false duplication CNVs calls. We evaluated CNVfilteR against two in-house NGS panel datasets containing a total of 542 samples. The number of false positives decreased by 15% and 12.5%, and no CNV was wrongly identified as false positive. CNVfilteR specifically supports VCFs generated by VarScan2, Strelka/Strelka2, freeBayes, HaplotypeCaller (GATK), and UnifiedGenotyper (GATK). Although further work is convenient to assess the performance, CNVfilteR can be used to improve the specificity of CNV calling tools. CNVfilteR is freely available at Bioconductor site `https://bioconductor.org/packages/CNVfilteR`.

Poster 12

Genetic characterization of secondary hemophagocytic lymphohistiocytosis patients

Laura Batlle-Masó, Laura Viñas-Giménez, Clara Franco, Mónica Martínez-Gallo, Roger Colobran, Ferran Casals

Presenting author: Laura Batlle-Masó, Universitat Pompeu Fabra (UPF)

Hemophagocytic lymphohistiocytosis (HLH) is a rare, severe disease caused by dysregulation of the immune cells. It is characterized by an exaggerated inflammatory response that can lead to severe complications and/or death. It is considered an autosomal recessive condition caused by mutations at PRF1, UNC13D, STX11 and STXBP2. However, there are also secondary cases in which an external agent such as an infection or a malignancy triggers the disease. Our aim is to study the genetic mechanisms underlying secondary HLH in late-onset patients. For that, we performed whole-exome sequencing and bioinformatics analysis of 31 secondary HLH cases. The project was divided into three main phases: the first phase is focused on exploring genetic variation in PRF1, UNC13D, STX11 and STXBP2. The second phase is intended to study possibly pathogenic variants in other candidate genes (immunodeficiency genes or blood genetic diseases). Finally, the main objective of the third phase is to extend the analysis to the whole exome data, exploring other kinds of variation such as somatic events and structural variation. In this communication, we report the results of the first phase, which shows that 16 patients (51.61%) carry rare variants in HLH candidate genes: ten monoallelic variants, one compound heterozygote and four double heterozygotes. We also corroborate the significant enrichment of PRF1_p.A91V and UNC13D_p.A59T in the HLH cohort compared to healthy controls. In addition to likely pathogenic variants in two-four patients, we suggest that secondary HLH may be caused by an accumulation of functional alterations in cytotoxic pathway genes. These hypotheses will be corroborated in silico with the results of second and third phases and functionally validated afterwards.

Poster 13

Repair PolyPurine Reverse Hoogsteen hairpins act without producing off-target effects in the genome

Alex Jimenez Felix, Veronica Noe, Carlos Ciudad

Presenting author: Alex Jimenez Felix, Universitat de Barcelona (UB)

A Repair PolyPurine Reverse Hoogsteen hairpin (repair-PPRH) consists of a (i) PPRH hairpin core that binds to a polypyrimidine target sequence in the dsDNA and (ii) an extension sequence attached to the 5' end of the molecule, which is homologous to the DNA sequence to be repaired, but containing the corrected nucleotide instead of the mutation. By using different Repair-PPRHs we were able to correct different nonsense mutations caused by a single substitution in the endogenous locus of the adenosyl phosphoribosyl transferase (aprt) gene. Surviving colonies were obtained by applying the +AAT (adenine-aminoptherine-thymidine) selection and were analyzed by DNA sequencing, mRNA expression and enzymatic activity. In addition, we assessed the possible off-target effects generated in the genome due to the treatment with the repair-PPRH. We performed Whole-Genome Sequencing (WGS) analyses to check if there was any major difference in the genome of the repaired cells in comparison with the original mutant cells. To do so, we looked at (i) the number of total variants and (ii) if there was any evidence of the insertion of the repair-PPRH in other genomic locations. We compared the number of variants (insertions, deletions and single nucleotide variants) in both samples and we did not detect any major discrepancy. Therefore, there was no evidence of a major increase in the number of variants in the repaired cells. To study the possible integration of the repair-PPRH in different genomic regions, reads with similarity to the construct were scrutinized. We found out that all the reads were mapped in the target region. No unmapped reads or mapped anywhere else with similarity were found. These results demonstrate that Repair-PPRHs can correct point mutations in the dsDNA of mammalian cells without off-target effects.

Poster 14

GCAT genome: A comprehensive catalog of genetic variability of the Iberian population

Jordi Valls-Margarit, Iván Galván-Femenía, Dani Matias-Sánchez, Mario Cáceres-Aguilar, Rafael de Cid-Ibeas, David Torrents-Arenales

Presenting author: Jordi Valls-Margarit, Barcelona Supercomputing Center (BSC-CNS)

One of the major challenges of the scientific community is to understand the genetic variation behind the wide range of the phenotype variation in humans. For this purpose, international and local initiatives, such as the 1000 Genomes project, Haplotype Reference Consortium (HRC), Genomes of the Netherlands (GoNL) or the UK10K, have joined their efforts in the present decade. Among other goals, these projects provide large reference panels of genetic variation using Whole Genome Sequencing (WGS) technologies. Currently, specific reference panels are commonly used for downstream genetic studies, requiring a comprehensive genome wide information through high quality imputation methods. Thus, increasing the chance of discovering disease risk variants and enlightening the knowledge of complex diseases in human populations. However, most of these reference panels only cover a part of the total genetic variation, probably due to the low (3 12X) coverage used, leading to an inefficient characterization of Structural Variants (SVs), which remain poorly understood in relation with common diseases. In this context, the GCAT Genomes For Life project, which is designed to integrate and assess the role of epidemiological, environmental and omic factors in the development of diseases in general population. This project comprises a living cohort of 20,000 participants between 40 to 67 years old linked to public electronic health records. A core of the cohort (GCATcore n=5,000) includes a deep characterization with SNP-array genotypes and metabolomic data, cancer exome sequencing (n=200, 400X coverage) and WGS (n=808) at high coverage (30X). We here present, preliminary results of a deep characterization of SNPs and SVs across WGS of 808 GCAT participants. Furthermore, we show the methodological pipeline used to complete the first comprehensive catalogue of genetic variation within the iberian population. This valuable genetic resource is expected to allow, among other, the discovery of the role of SVs within complex diseases.

Poster 15

A normalization-free analysis of stoichiometric change for RNA-seq and metagenomics

Thom Quinn, Cedric Notredame, Ionas Erb

Presenting author: Ionas Erb, Centre de Regulació Genòmica (CRG)

Compositional data analysis (CoDA) has emerged quite recently as a unified framework within which sequencing data (RNA-seq expression or metagenomics data) can be analyzed. The main advantage of CoDA is that samples do not need to be normalized. Instead of trying to guess absolute abundances from relative sequencing data, all analyses are carried out with respect to an internal reference, e.g. in form of one or various genes or operational taxonomic units (OTUs). The resulting "theory of relativity" casts a new light on the notions of differential expression (DE) and correlation: the same differentially expressed gene ratio can be formulated as a result of both, one, or none of the genes being DE themselves, depending on the correlation the reference has with the genes. Taking this idea to its logical extreme, we show that the DE analysis of all gene ratios enables us, unlike for all commonly available DE packages, to uncouple significance of DE from normalization procedures. This new perspective makes biological sense since the stoichiometry of gene expressions determines the phenotype of the cell. (Similarly, abundance relationships of OTUs determine metabolic profiles.) We show that it also makes computational sense, with our efficient implementation in the propr R package effectively controlling the false discovery rate. Finally, we show that due to the self-normalizing property of our approach, despite the huge variation across brain tissues and along development, we can make a simple comparison between male and female brains across more than 500 samples that reveals gene sets changing in a highly coordinated manner.

Poster 16

Detection and validation of G-quadruplex motifs to silence Thymidylate synthase using Polypurine Reverse Hoogsteen Hairpins

Eva Aubets, Alex J. Félix, Miguel G Garavís, Anna Aviñó, Ramon Eritja, Véronique Noé, Carlos J. Ciudad

Presenting author: Eva Aubets, Universitat de Barcelona (UB)

G-quadruplex motifs are structures that may regulate translation and transcription. In this work, we explored putative G4-forming sequences that could modulate Thymidylate synthase (TYMS), used as an anti-cancer target because of its role in the synthesis "de novo" of dTTP. Traditional treatments (5-FU) can lead to drug resistance through an autoregulatory mechanism of TYMS inhibiting its own translation. We found 9 potential G4 sequences in the TYMS mRNA using the Quadruplex forming G-Rich Sequences (QGRS) mapper. The formation of the G4 structure with the highest G-score in DNA and mRNA was confirmed by circular dichroism or by NMR. We designed a Polypurine Reverse Hoogsteen hairpin to modulate TYMS expression interfering with a G4-forming sequence. We demonstrated the binding of the PPRH (HpTYMS-G4-T) to its target sequence in the complementary strand of the G4 sequence by gel-shift assays. In addition, TYMS protein bound to this target sequence, both as dsDNA or ssDNA, whereas 2 negative control proteins (DHFR and BSA) did not produce any binding. Additionally, we observed that HpTYMS-G4-T and purified TYMS compete with each other for the binding to the target sequence in the dsDNA. Surprisingly, TYMS protein was also able to bind to the mRNA corresponding to the G4-forming sequence. These results uncover the relevance of this region for the regulation of TYMS expression. HpTYMS-G4-T was cytotoxic in a dose dependent manner. This PPRH could be altering TYMS transcription and the G4 structure formed. In fact, incubation with HpTYMS-G4-T decreased TYMS mRNA and protein levels. In conclusion, our results show that the designed PPRH binds to its target sequence of TYMS and decreases HeLa cells viability. Therefore, PPRHs can be considered as a new type of molecules to modulate TYMS expression and overcome the resistance produced by traditional treatments.

Poster 17

UMI4Cats: an R package for analyzing UMI-4C contact data

Marc Subirana Granés, Mireia Ramos Rodríguez, Lorenzo Pasquali

Presenting author: Marc Subirana Granés, Institut Germans Trias i Pujol (IGTP)

Three-dimensional chromatin structure is essential for gene regulation and was shown to have dynamic proprieties during differentiation, in malignant processes and under certain stimuli, such as hormones or pro-inflammatory cytokines. Different methods, such as 3C, 4C or HiC were successfully applied to capture the 3D chromatin structure of different cell types. Nevertheless, an amplification step shared by all these techniques, limits the quantitative comparison of the contact intensities detected in different cell types or conditions. In order to tackle this limitation, Schawartzman et al. developed UMI-4C, a new method using targeted chromosome conformation capture (4C) and unique molecular identifiers (UMI), in which the sonication of the ligated fragments allows collapsing PCR duplicates. However, a state-of-the-art software that deals with the specific challenges of processing such data is currently missing. To answer this need, we have developed UMI4Cats (UMI-4C Analysis Turned Simple), an R package that deals with the quality control, processing, analysis and differential testing of UMI-4C data. Unlike other available methods, UMI4Cats allows the analysis of UMI-4C experiments generated using any restriction enzyme with any viewpoint length. It performs all necessary steps to collapse ligation fragments with the same UMIs, thus removing PCR duplicate bias. Additionally, it normalizes UMI-4C data taking into account the number of UMIs in the region of interest and the distance from the viewpoint and it integrates the visualization of UMI-4C data with genomic annotations. Finally, it includes different methods for testing differential contacts, including a sliding window approach using a Fisher's Exact Test with multiple-testing correction and DESeq2 Wald's Test In summary, we developed UMI4Cats, a simple, robust and user-friendly R package for the analysis of UMI-4C data.

Poster 18

Improving the RD-Connect GPAP usability to facilitate rare disease diagnosis and gene discovery

Alberto Corvo, Leslie Matalonga, Steven Laurie, Gemma Bullich, Davide Piscia, Marcos Fernandez, Daniel Picó, Carles Hernandez, Anastasios Papakonstantinou, Ivo Gut, Sergi Beltran

Presenting author: Alberto Corvo, Centre Nacional d'Anàlisi Genòmica (CNAG-CRG)

More than 400 million people worldwide are affected by rare diseases, of which most of them have uncertain diagnosis and unmet needs. The consequences for an individual are a decreased quality of life, which is often compromised due to a lack of knowledge of effective treatments. Therefore, the International Rare Disease Research Consortium vision for 2027 is "to enable all people living with a rare disease (RD) to receive an accurate diagnosis, care, and available therapy within one year of coming to medical attention". The RD-Connect Genome-Phenome Analysis Platform (GPAP) is contributing to facilitate this vision through a user friendly system that enables data collation, processing, sharing, analysis and interpretation of integrated phenotypic and genomic data. The GPAP is part of the European Joint Programme on Rare Diseases (EJP-RD) and Solve-RD, two of the major EU projects on Rare Diseases. As part of this collaborative work, we are re-designing the Graphical User Interface and are implementing new features to improve the system's usability. One of the new functionalities being developed is a visual analytics module to interactively create and analyze sub-cohorts based on the individual's clinical and phenotypic data. The application is web-based and implemented in React and D3.js, two Javacript libraries that offer high flexibility in design, usability, and interactive visualization.

Poster 19

FOBI: An ontology to represent food intake data and associate it with metabolomic data

Pol Castellano-Escuder, Raúl González-Domínguez, David S. Wishart, Cristina Andrés-Lacueva, Alex Sánchez-Pla

Presenting author: Pol Castellano-Escuder, Universitat de Barcelona (UB)

Nutrition research can be conducted by using two complementary approaches: 1) traditional self-reporting methods or 2) via metabolomics techniques to analyze food intake biomarkers in biofluids. However, the complexity and heterogeneity of these two very different types of data often hinder their analysis and integration. To manage this challenge, we have developed a novel ontology that describes food and their associated metabolite entities in a hierarchical way. This ontology uses a formal naming system, category definitions, properties and relations between both types of data. The ontology presented is called FOBI (Food-Biomarker Ontology) and it is composed of two interconnected sub-ontologies. One is a “Food Ontology” consisting of raw foods and prepared foods while the second is a “Biomarker Ontology” containing food intake biomarkers classified by their chemical classes. These two sub-ontologies are conceptually independent but interconnected by different properties. This allows data and information regarding foods and food biomarkers to be visualized in a bidirectional way, going from metabolomics to nutritional data or vice versa. Potential applications of this ontology include the annotation of foods and biomarkers using a well-defined and consistent nomenclature, the standardized reporting of metabolomics workflows (e.g. metabolite identification, experimental design), or the application of different enrichment analysis approaches to analyze nutrimetabolomic data. FOBI is freely available in both OWL (Web Ontology Language) and OBO (Open Biomedical Ontologies) formats at the project’s Github repository (`https://github.com/pcastellanoescuder/FoodBiomarkerOntology`) and FOBI visualization tool is available in `https://polcastellano.shinyapps.io/FOBI_Visualization_Tool/`.

Poster 20

Integrative transcriptome analysis of brain metastases immune microenvironment

Sara Hijazo-Pechero, Ania Alay, Noelia Vilariño, Noemi Vidal, Jordi Bruna, Ernest Nadal, Xavier Solé

Presenting author: Sara Hijazo-Pechero, Institut Català d'Oncologia (ICO), Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)

Brain metastases (BM) are a frequent complication in cancer and are associated with quality of life deterioration and dismal prognosis. In the recent years, immunotherapy has become the standard of care in some tumors. However, the efficacy of these therapies is known to be conditioned by the immune contexture which has not been well characterized in BM. To characterize the immune infiltrate pattern of a set of BM samples (76 breast cancer, 34 melanoma, 28 NSCLC) using bioinformatics approaches; to define immune groups based on the cellular composition of that infiltrate and to characterize these groups from a molecular point of view to find potential specificities. Gene Set Variation Analysis algorithm was employed to compute the relative abundances of 20 immune cells and 4 brain specific populations in 138 BM samples from 7 different publicly available microarray gene expression datasets. Groups with similar immune infiltrate composition were defined using an unsupervised clustering method. Finally, statistical analyses were used to find associations between molecular characteristics and the previously defined immune groups. BM were classified into three groups based on their immune infiltrate composition. BM within the group with a greater degree of immune cell infiltration showed an increasing relative abundance of most cell types, especially macrophages M2, microglia, or regulatory T cells for which an immunosupressive role has been reported in the context of BM. Cytotoxic T lymphocytes might be turning into a more exhausted phenotype due to the expression of immunosuppressive markers. This work describes an immune landscape in BM and defines three groups based on the cellular composition of their immune infiltrate. These results might be relevant to understand how BM respond to immunotherapy and may guide the development of novel therapeutic strategies.

Poster 21

Characterization of High Risk Neuroblastoma group: Potential epigenetic biomarkers for treatment response assessment

Alícia Garrido Garcia, Soledad Gómez-González, Sara Pérez-Jaume, Laura Garcia-Gerique, Mariona Suñol, Óscar Muñoz, Cinzia Lavarino

Presenting author: Alícia Garrido Garcia, Institut de Recerca Sant Joan de Déu

Neuroblastoma (NB), the most frequent extracranial pediatric solid tumor, accounts for 15% of cancer-related deaths in children. Probability of cure varies according to patient's age, extent of disease and tumor biology. High risk NB (HR-NB) are a heterogeneous group of tumors, whereby patients can display response to treatment and long-term outcome or develop early progressive, chemorefractory disease with poor outcome (ultra-high risk; UHR-NB). Genetics underlying this aggressive subgroup is still greatly unknown and no reliable biomarkers have been reported. Our aim was to define and characterize UHR-NB subgroup using DNA methylation (DNAme) and gene expression (GE) profiling, and investigate the effects of core (epi)genetic alterations on regulatory pathways. Thereby, identify potential biomarkers and therapeutic targets. We analyzed DNAme and GE datasets of ¿100 HR-NBs. DNAme data was annotated according to gene location, CpG islands and chromatin state categories. GE data was used to determine the affected pathways in the UHR-NB tumors. Potential biomarkers were identified by Cox-regression models. Survival curves were analyzed by Kaplan-Meier method and log-rank test. Validation was performed by bisulfite sequencing, pyrosequencing and immunohistochemistry. We observed two differential DNAme patterns within HR-NB group associated with divergent clinical evolution and defining a subgroup of patients with rapidly progressing, chemo-refractory tumors. Mining the DNAme patterns, we identified a reduced set of differentially methylated cytosines that defined the UHR-NBs. Functional genomic analysis of UHR-NB unveiled differential DNAme patterns associated with development pathways and metabolism. Crosschecking information of DNAme and GE differential analysis also confirmed changes in metabolism and purine biosynthesis. These results provided potential therapeutic targets. Our findings reveal that UHR-NB is characterized by specific DNAme changes. We have identified potential biomarkers that define this subgroup of NB tumors and identified aberrant activation of biological pathways that could be relevant for targeted therapeutic options.

Poster 22

Towards a network-driven biological knowledge integration

Adria Fernandez-Torras

Presenting author: Adria Fernandez-Torras, Institut de Recerca Biomèdica (IRB Barcelona)

Biological data is steadily growing while its integration is becoming an increasingly cumbersome process. We have created a gigantic heterogeneous network (more than 700k nodes and 77M edges) that harmonizes and connects biomedical data from multiple kinds and sources. Overall, 14 types of biological entities (e.g. genes, diseases, drugs) were linked by 65 types of relationships (e.g. drug treats disease, gene interacts with gene). To enable fast and flexible querying of such a complex resource, we stored the full network as a graph database, appropriately handling the ontologies and vocabularies encountered throughout the more than 200 sources. Finally, we explored the graph to uncover popular and neglected nodes among sources, find relevant relationships between entities and unveil correlations between datasets.

Poster 23

BacFITBase: A database to assess the relevance of bacterial genes during host infection

Javier Macho Rendón, Benjamin Lang, Gian Gaetano Tartaglia, Marc Torrent Burgas

Presenting author: Javier Macho Rendón, Universitat Autònoma de Barcelona (UAB)

Bacterial infections have been on the rise world-wide in recent years and have a considerable impact on human well-being in terms of attributable deaths and disability-adjusted life years. Yet many mechanisms underlying bacterial pathogenesis are still poorly understood. Here, we introduce the BacFITBase database for the systematic characterization of bacterial proteins relevant for host infection aimed to enable the identification of new antibiotic targets. BacFITBase is manually curated and contains more than 90 000 entries with information on the contribution of individual genes to bacterial fitness under in vivo infection conditions in a range of host species. The data were collected from 15 different studies in which transposon mutagenesis was performed, including top-priority pathogens such as Acinetobacter baumannii and Campylobacter jejuni, for both of which increasing antibiotic resistance has been reported. Overall, BacFITBase includes information on 15 pathogenic bacteria and 5 host vertebrates across 10 different tissues. It is freely available at `www.tartaglialab.com/bacfitbase`.

Poster 24

Polygenic risk scores to identify genetic risk of ADHD

Sofía Aguilar Lacasaña

Presenting author: Sofía Aguilar Lacasaña, Institut de Salut Global de Barcelona (ISGlobal)

There are multiple environmental and genetic factors operating in the etiology of Attention-Deficit/Hyperactivity disorder (ADHD) symptoms, although the highly heritable estimates of this disorder, suggest a strong genetic contribution. Genome-wide association studies (GWAS) findings suggest that multiple common genetic variants of small effect size contribute to ADHD, implying a highly polygenic architecture. Thus, polygenic risk scores (PRS) are increasingly being used to index genetic susceptibility of ADHD, which also provide substantially greater predictive power by aggregating the whole set of genome-wide information. There has been exponential growth in the literature of PRS studies, which challenges the standardization of analytical methods in this field. In this work, we give a complete up-to-date account of PRS studies on ADHD, which serves as a reference catalog for researchers working with PRS and ADHD. We searched MEDLINE/Pubmed and identified 43 articles that met our eligibility criteria. We noticed that the most used tools for PRS calculation in ADHD were PLINK (58%) and PRSice (30%). PLINK uses a linear scoring system for calculating PRS, in which the quality control of the data needs to be performed before running the computations. PRSice is similar, but also includes a step in which ambiguous SNPs can be removed (clumping). In addition, PRSice allows the selection of different p-values (thresholds) to obtain the best fit scores for the data. Moreover, we showed that PRS are used for both testing the genetic overlap between characteristics of ADHD, as well as, testing the genetic overlap among ADHD and other neurodevelopmental disorders. These studies agreed that polygenic architecture may help identify sets of weak variants that otherwise remain undetected using traditional approaches. However, PRS calculation depend greatly on the power of the GWAS used to describe the genetic architecture of ADHD. Hence, further computational efforts are still required in the field.

Poster 25

Transcriptomic analyisis reveals immunological processes associated with the response to abatacept in rheumatoid arthritis

Irene Bonafonte Pardàs, Maria Lopez Lasanta, Antonio Gómez, Raimon Sanmarti, Carlos Marras Fernandez Cid, José Manuel Pina Salvador, Susana Romero-Yuste, Raul Maria Veiga Cabello, Pilar Navarro, Carme Moragues Pastor, Silvia Martinez Pardo, Javier de Toro-Santos, Amalia Sánchez, Dacia Cerda, Alejandro Prada, Alba Erra, Jordi Monfort, Ana Urruticoechea-Arana, Núria Palau, Raquel M. Lastra, Raül Tortosa, Andrea Pluma Sanjurjo, Sara Marsal, Antonio Julià

Presenting author: Irene Bonafonte Pardàs, Vall d'Hebron Institut de Recerca (VHIR)

The T cell costimulation modulator abatacept (CTLA4-Ig) has proven effective for the treatment of rheumatoid arthritis (RA). However, 30–40% of patients do not show a significant clinical improvement after treatment. The objective of the present study was to characterize the biological basis of this differential response. A total of n = 57 RA patients were recruited for this study. The primary clinical response to abatacept was defined at week 12 using the EULAR criteria. Good and moderate responders were aggregated into a single response group and compared to the no response group. Blood RNA was collected from all patients at baseline and, for a subgroup of patients (n=31), also at weeks 12, 24 and 48 of treatment. Gene expression levels were determined using paired-end RNA-seq (Illumina). Differential gene expression, association to biological processes, longitudinal association analysis and building of the multigenic predictor were performed using the R software and the specialized Bioconductor libraries. From the 57 patients treated with abatacept, n = 34 (59.5%) were classified as responders and n = 23 (40.5%) as non-responders. Biological process analysis identified two significantly distinct biological profiles between responders and non-responders. In responders, we found an association to pathways associated with the effector phase of T cells (e.g. interleukin-15 and 2 signalling, $P < 0.05$). Non-responders showed instead a strong association to biological processes associated with antigen presentation and activation of T cells ($P < 0.005$). Using the baseline gene expression profiles, we built a multigenic predictor of response to abatacept with a ROC AUC = 75%. The analysis of blood RNA profiles of RA patients has enabled the identification of specific biological processes associated with the lack of response to abatacept. Blood expression profiles can be predictive of the response to the drug at week 12 of therapy. Funded by Bristol-Myers Squibb.

Poster 26

Translocations in CDKN2A locus evidence the necessity of its loss for MPNST development

Miriam Magallón-Lorenz, Juana Fernández-Rodríguez, Meritxell Carrió, Edgar Creus, Conxi Lázaro, Eduard Serra, Bernat Gel

Presenting author: Miriam Magallón-Lorenz, Institut Germans Trias i Pujol (IGTP)

Malignant peripheral nerve sheath tumor (MPNST) is an aggressive type of soft tissue sarcoma with a bad prognosis. MPNST develops sporadically or in the context of neurofibromatosis type 1 (NF1). NF1 patients bear a germline inactivation of one NF1 allele and have an 8–13% lifetime risk of developing an MPNST. In the context of NF1, MPNSTs may arise from pre-malignant nodular lesions, atypical neurofibromas (aNF), usually associated with a pre-existing benign plexiform neurofibroma (pNF). pNFs arise by the complete inactivation of NF1 in a Schwann cell precursor and show no additional alterations. In addition to the mutation of both NF1 alleles, aNFs present the recurrent loss of the CDKN2A/B locus, mainly via sub-arm or focal deletions. In contrast, MPNSTs have highly rearranged hyperploid genomes with somatic copy number alterations affecting most chromosomes. According to literature, only about 70% of MPNSTs lose CDKN2A and the rest of them seem to have this locus unaltered. We analyzed the CDKN2A/B locus in 8 different MPNSTs cell-lines and 12 tumors using WES, RNA-seq, and SNP-array data. WES data from different cell-lines allowed us to identify a novel inactivation mechanism of CDKN2A in MPNSTs. We found inter-chromosomal translocation break ends in an intronic part of CDKN2A in MPNSTs cell-lines and tumors. All translocations cluster in a putative 200bp hotspot region that translocates to different chromosomes. We validated these translocations by PCR and Sanger sequencing. Moreover, RNA-seq data of those cell-lines bearing the translocation evidenced the presence of an altered expression pattern of the CDKN2A gene. A specific multiplex PCR assay was developed to screen other MPNSTs for translocations involving this hotspot. Our conclusion so far is that the complete inactivation of both NF1 and CDKN2A is required for MPNSTs to arise, at least in the context of NF1.

Poster 27

Genomic structural alterations as a driving force in MPNST development

Bernat Gel, Miriam Magallon, Ernest Terribas, Elisabeth Castellanos, Inma Rosas, Ignacio Blanco, Juana Fernández-Rodríguez, Conxi Lázaro, Eduard Serra

Presenting author: Bernat Gel, Institut Germans Trias i Pujol (IGTP)

Malignant peripheral nerve sheath tumors (MPNST) are soft tissue sarcomas with bad prognosis and lack of curative treatments. NF1 patients have an 8–13% lifetime risk of developing an MPNST. MPNST may arise from a preexisting benign plexiform neurofibroma (PNF), often after the formation of a pre-malignant distinct nodule termed atypical neurofibroma (aNF). We generated genomic structure (SNP-array), mutational (exome), transcriptomic (RNA-seq, microarray) and epigenomic (DNA methylation) data from a set of 15 MPNSTs, and also collected available data on MPNST, plexiform and atypical neurofibromas. We performed an integrative bioinformatic analysis this data to infer the mechanisms of MPNST development. Regarding the genomic structure, PNFs have no structural alterations except those affecting chromosome 17q involved in the somatic inactivation of NF1. aNFs also present recurrent losses of the CDKN2A/B locus. In contrast, MPNSTs have hyperploid and highly rearranged genomes with somatic copy number alterations (SCNAs) affecting most chromosomes. However, MPNST genome structure is highly stable over time. MPNSTs have a very low number of point mutations, with no clear recurrently affected genes. Most point mutations appear to be acquired after the genome reorganization. This data suggests a model for MPNST origin, with a first progression towards a proliferative cell with reduced senescence due to the loss of NF1 and CDKN2A/B, followed by one or more random catastrophic events of genomic alteration and the selection of a viable stable genomic combination. SCNA have a profound impact on gene transcription levels and create regions with an accumulation of over- and under- expressed genes, transcriptional imbalances (TI). TIs mostly capture passenger gene expression but allow identification of genes with SCNA-independent expression regulation. In conclusion, gross genomic structural alterations are a driving force in MPNST biology and their genomic stability suggest a catastrophic event mediated by loss of senescence capacity as a probable origin.

Poster 28

Worldwide distribution and dating of signatures of recent adaptation in our genomes

Aina Colomer, Jesús Murga-Moreno, Antonio Barbadilla, Sònia Casillas

Presenting author: Aina Colomer, Universitat Autònoma de Barcelona (UAB)

Since the divergence with chimpanzees, and especially when migrating across the globe, our species has faced frequent social and environmental challenges that have acted as selective pressures. In response to these, natural selection has shaped our genomes, leaving signatures that are preserved in our present-day genetic variation. In this context, PopHumanScan was conceived as a collaborative database amassing 2859 putatively selected genomic regions to facilitate their subsequent analysis. This catalogue encompasses the 22 non-admixed human populations of the 1000 Genomes Project phase 3 and pinpoints signatures of putative selective processes at different historical ages based on a combination of eight different population metrics. Seeking to achieve a better understanding of the Earth colonization process by providing insights into the patterns of adaptation among populations and across time, here we present our current methodological approach consisting in: (i) recalculate the integrative haplotype score (iHS) metric genome-wide using the most recent and finest recombination map available in the human genome; (ii) define putatively selected regions based on the empirical distribution of iHS values; (iii) scan the candidate regions to pinpoint, in each case, the most likely favored mutation by combining different evidences of selection and functional annotations; and (iv) date the time to the most recent common ancestor for a beneficial allele (TMRCA) for each selected favored mutation. In addition to help us explain the evolutionary mechanisms behind the variation patterns in particular regions of our genomes, our approach is allowing us to tackle questions of remarkable interest such as identifying sweeps that spread concurrently in our recent evolutionary history.

Poster 29

Histone H1 depletion in cancer cells promotes changes on genome architecture related to gene expression deregulation

Albert Jordan, Núria Serna

Presenting author: Núria Serna, Institut de Biologia Molecular de Barcelona (IBMB-CSIC)

Histone H1 binds to the linker DNA at the nucleosome, participating in the formation of higher-order chromatin structures. Human somatic cells may contain up to seven members of the histone H1 family contributing to the regulation of nuclear processes, apparently with certain subtype specificities. We have previously shown that in T47D breast cancer cells, the combined knock-down of H1.2 and H1.4 subtypes (multi-H1 KD) has a strong deleterious effect, deregulates many genes, promotes the appearance of accessibility sites genome-wide and triggers an interferon response via activation of heterochromatic repeats. Now, through the integration of chromatin immunoprecipitation followed by sequencing (ChIP-Seq), RNA sequencing (RNA-Seq) and high-throughput chromosome conformation capture (Hi-C) techniques, we aim to elucidate the biological role of different H1 subtypes in the interplay between genome architecture and gene expression. Our results support that histone H1 variants are differentially distributed in topologically associating domains (TADs) and A/B compartments. For instance, TADs located within compact and GC-poor genomic regions were characterized by a high H1.2/H1X content ratio and overlapped with the B compartment of the 3D genome. Multi-H1 KD increased TAD border definition and intra-TAD contacts, while decreased inter-TAD interactions. Moreover, TADs enriched in histone H1.2 showed major transitions from B to A compartment and changes in interactions. Multi-H1 depletion also promoted genes deregulation in 40% of total TADs. Specifically, up-regulated genes accumulated within TADs presenting high H1.2/H1X ratios and low gene richness, while the opposite occurs in TADs containing down-regulated genes. Within affected TADs, the frequency of deregulated genes compared to total gene count was higher in those with a high H1.2/H1X ratio. In conclusion, our data suggest that the equilibrium between distinct histone H1 variants is involved in maintaining the topological organization of the genome and the proper expression of particular gene programs.

Poster 30

Hereditary non-polyposis colorectal cancer may be explained by accumulation of genetic risk variants

Nuria Bonifaci, Pilar Mur, Anna Díez-Villanueva, Matilde Navarro, Marta Pineda, Gabriel Capellá, Victor Moreno, Laura Valle

Presenting author: Nuria Bonifaci, Institut Català d'Oncologia (ICO)

Most of the familial colorectal cancer cases cannot be linked to pathogenic variants in known cancer-predisposing genes, being the management of those families based mainly on family history. A personalized cancer risk assessment is required to perform a more precise management of each individual. Polygenic risk scores (PRS) based on the risk effects of multiple common genetic variants have been proposed for individual risk assessment on a population level. Our aim was to investigate the applicability of the PRS for risk prediction in hereditary CRC patients. A weighted PRS (wPRS) based on 92 previously known CRC risk single nucleotide polymorphisms (SNPs) (Huyghe et al. Nature genetics 2018) was determined for 504 familial CRC individuals: serrated polyposis syndrome patients (SPS, N = 82), and genetically-unexplained hereditary/early CRC (CRC-X, N = 422). Moreover, 1,642 hospital-based consecutively recruited cancer-free controls, and 1,077 sporadic cases were used. Logistic regression was used to model the risk. Two-sided t-test was applied to compare the risk between control samples and familial CRC or sporadic cases. Individuals showing high PRS (over threshold= median +2SD standard deviation from control samples) were well characterized. Additionally, we grouped individuals (CRC-X and controls) into wPRS quantiles and odds ratios (OR) were estimated referring to the median quantile. The median of wPRS was significantly higher for CRC-X and sporadic CRC cases compared with control samples ($p < 2.22e-16$). These differences were not observed in SPS samples. CRC-X individuals with higher wPRS values (N = 32) were males (83%), over 50 years of age (84%), fulfilling Bethesda criteria (87%). A 4 fold-increase in CRC risk was identified for subjects in the highest quantile (20th) respect to the reference (OR = 4.34; 95%CI 2.31-8.14; P-value = 5.0529e-06). In conclusion, our results show higher genetic risk in CRC-X patients and suggested its potential utility for disease risk stratification in CRC families with unknown germline pathogenic mutation.

Poster 31

xcmsQCtools: Quality metrics for liquid and gas chromatography mass spectrometry data sets

Pol Solà-Santos, Sergi Picart-Armada, Alexandre Perera-Lluna

Presenting author: Pol Solà-Santos, Universitat Politècnica de Catalunya (UPC)

Metabolites, low weight chemical molecules involved in cellular reactions, constitute a functional fingerprint complementary to the upstream information obtained through genetics, transcriptomics and proteomics. The metabolome, full set of metabolites, is critically dependent on environmental factors. In consequence, actions that provide confidence to the quality of the experiments (i.e. quality assurance, QA) and measures to quantify and report the latter (i.e. quality control, QC) are fundamental to ensure high-standard research. Targeted studies, i.e. focus on reduced subset of the metabolome, have established guidelines for QA/QC. In contrast, untargeted studies, i.e. measure as many metabolites as possible, have a lack on community consensus for both QA and QC. Although recent efforts to standardize QA and QC through instrumental solutions, quantitative measures remain heterogeneous within the metabolomics community. We propose an array of existent and novel metrics to quantify the most typical sources of variance induced in-between the whole experimental process. First, confounding effects are quantified through linear models. Second, unknown and atypical variances are monitored through multivariate statistical control processes based on principal component analysis under a sliding window strategy. Third, system suitability is characterized, among others, through peak shape characterization. Fourth, signal instability is tested by monitoring the sample intensity mean along injection order. Finally, sub-optimal peak extraction (e.g. background integration, phantom peaks) is quantified with Gaussian Mixture Models. Due the lack of true references in untargeted metabolomics there is not a reliable way to validate quality metrics thresholds. To overcome this, metrics are characterized through a batch of annotated experiments of the Metabolights platform. This results are established as benchmark for quality control in untargeted metabolomics. In addition we present xcmsQCtools, an R package to implement all the proposed metrics and generates a summarized QC report.

Poster 32

Machine learning mapping of layperson medical terminology into the Human Phenotype Ontology

Enrico Manzini, Jon Garrido-Aguirre, Alexandre Perera-Lluna

Presenting author: Jon Garrido-Aguirre, Universitat Politècnica de Catalunya (UPC)

The precise analysis of the clinical phenotypes of an individual is known as deep phenotyping, a methodology with the potential to improve the identification of disease with prognostic and therapeutic implications. An essential tool for deep phenotyping is the Human Phenotype Ontology (HPO), a standardized vocabulary of human phenotypic abnormalities. There exists a huge terminological gap between patients (i.e. laypeople) and the technical language of HPO, hindering its use for deep phenotyping outside clinical and academical contexts, for instance in patient-driven initiatives (e.g. patient research platforms). The aim of this work is to exploit deep learning techniques in order to fill in this gap. We use a two-step method to translate from layperson medical terms into HPO terminology. First, we create a vector space to represent HPO. Then, the HPO embedding is used as the output space for a neural network model that combines convolutional and recurrent layers. For both training and testing phases we used layperson terms and other textual descriptors included in HPO as inputs. The inputs were codified in a word embedding layer trained with the model. Different output embeddings (HPO-specific, generic, and combined) were built, tested and analyzed using an ontology-specific similarity function. In general, the performance using different embeddings is similar, with median similarity of 0.94. In addition, 43% of the terms were identified exactly (sim = 1), and 80% of the terms approximately ($0.7 < \text{sim} \leqslant 1$), on average. Importantly, the output embeddings provide meaningful information, reflected on an increase in the overall performance of the models as compared with random embeddings at the output (0.59 median similarity; 22% exact; 46% approximate). We show that combining neural network models with different word embeddings meaningful layperson-HPO mappings can be learned.

Poster 33

The mutational footprints of cancer therapies

Oriol Pich, Ferran Muiños, Abel Gonzalez-Perez, Nuria López-Bigas

Presenting author: Oriol Pich, Institut de Recerca Biomèdica (IRB Barcelona)

Some cancer therapies damage DNA and cause mutations in both cancerous and healthy cells. Therapy-induced mutations may underlie some of the long-term and late side effects of treatments, such as mental disabilities, organ toxicity and secondary neoplasms. Nevertheless, the burden of mutation contributed by different chemotherapies has not been explored. Here we identify the mutational signatures or footprints of six widely used anticancer therapies across more than 3,500 metastatic tumors originating from different organs. These include previously known and new mutational signatures generated by platinum-based drugs as well as a previously unknown signature of nucleoside metabolic inhibitors. Exploiting these mutational footprints, we estimate the contribution of different treatments to the mutation burden of tumors and their risk of contributing coding and potential driver mutations in the genome. The mutational footprints identified here allow for precise assessment of the mutational risk of different cancer therapies to understand their long-term side effects.

Poster 34

Novelty detector a quality control in methylation machine-learning based predictors

Joshua Llano, Soledad Gómez, Cinzia Lavarino, Alexandre Perera-Lluna

Presenting author: Joshua Llano, Universitat Politècnica de Catalunya (UPC)

The automatic classification of medical data for diagnosis or prognosis using online tools are increasingly used today. Misclassifications in diagnosis or prognosis have an undesired impact on the patient's treatment. One of the issues yielding to misclassification in predictive models is the departing of the query data from the domain from which the predictive model was built with. We have developed a classifier for the Platform of Epigenetic Classifiers in Oncology from Sant Joan de Déu Hospital with the aim the detection of whether a sample belongs to a known category from the medulloblastoma subgroup domain or it does not belong to the domain of the predictive model was built with. We have analyzed 1500 450k DNA methylation micro-arrays published in GEO. We have divided the data into a training set of 200 medulloblastoma samples and a validation set of 1300 samples (medulloblastoma and non-medulloblastoma samples). Both data sets were normalized before being used for the one-class SVM training. Also, a grid search was performed in order to find the optimal nu parameter of the one-class SVM. Using a one-class SVM approach, we developed and validated a novelty detection classifier for a predictor of medulloblastoma type based on six epigenetic biomarkers (AUC = 0.95). Our findings show that one-class SVM classifier represent a simplified approach for a novelty detection as a samples filter that can be included in the Platform of Epigenetic Classifiers in Oncology analysis pipeline.

Poster 35

Gene expression and blood cells in long-distance races

Pol Gil, Emma Roca, Maria Maqueda, José Manuel Soria, Alexandre Perera-Lluna

Presenting author: Pol Gil, Universitat Politècnica de Catalunya (UPC)

Running in long-distance races can have a great impact on health during and after the activity. Numerous studies on blood cell dynamics during physical exercise have been carried out. However, the number of researches on differential expression in this field is not large. There are even a few published works attempting to find variations in differential expression not related to variations in the blood cell count. The objective of this study is to try to explain the variation in total blood differential expression while controlling the variation of the values of the blood cell count. For this purpose, genetic expression data (HuGene2.0st microarrays), complete blood count data and other biological variables (sex, age, performance) are collected from runners before and immediately after the competition "Volta a la Cerdanya 2013". Runners are stratified by three distance categories; A) 14 km ($N = 20$, $Age = 36.5 \pm 9.6$ years), B) 35km ($N = 26$, $Age = 35.1 \pm 8.3$ years) and C) 55km ($N = 13$, $Age = 36.2 \pm 4.8$ years). Blood count and expression data variation after the race are computed normalized to the pre-race values of data. Descriptive statistics for blood count are provided (distributions and differences between categories). A linear regression model was fitted to each TC variation in expression data. The 4 first principal components of blood count indicators, Pace (m/s) and Distance (km) have been used as explanatory variables. Finally, gene enrichment analysis has been computed over the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway. An average increase in erythrocytes (9.8%), platelets (21.7%), eosinophils (24.4%), basophils (188.5%), segmented neutrophils (205.2%) and monocytes (47.9%) and an average fall in lymphocyte levels (-41.7%) has been observed. 95 genes are reported as statistically significant (independent of the cell counts), 1105 in PC3 (related to cell counts), 90 in PC4 (related to cell counts) and 8 in Distance.

Poster 36

TALKIEN: Crosstalk bipartite network, a web-based Shiny interface to analyze crosstalk networks

Ferran Moratalla-Navarro, Victor Moreno, Rebeca Sanz-Pamplona

Presenting author: Ferran Moratalla-Navarro, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)

Molecular and cellular changes in particular cell types could be explained by the crosstalk with surrounding neighbors (i.e. stromal content). Cellular crosstalk is mainly explained by the interaction between secreted molecules in a tissue/cell and receptor proteins in other tissue/cells. This event is taking more attention in different research fields like cancer, in which an active communication exists not only between different cell communities within the tumor bulk but also between the tumor and the healthy surrounding tissue. We present TALKIEN: crossTALK bIpartitE Network, a simple and intuitive online R/Shiny application to analyze crosstalk between two lists of genes. This program takes plasma membrane receptor and secreted proteins annotated by Human Protein Atlas v13, combined with Protein-Protein interaction STRING database v10, to achieve high confidence network interactions between and/or within these two kind of proteins. TALKIEN computes basic global network analysis, local centrality measures, allows users to select four different interactive network graphical layouts and performs a further enrichment analysis for genes found in networks. With this tool, users are able to quickly detect important protein protein interactions that could improve the understanding of a particular biological question, demonstrate changing effects in different situations or help generating new in-silico predictions of cell-cell communication. TALKIEN is freely available at `https://shiny.snpstats.net/talkien/`.

Poster 37

The mechanisms of drug resistance in the emergent pathogen Candida glabrata

Miquel Àngel Schikora Tamarit, Ewa Ksiezopolska, Toni Gabaldón

Presenting author: Miquel Àngel Schikora Tamarit, Barcelona Supercomputing Center (BSC-CNS) and Institut de Recerca Biomèdica (IRB Barcelona)

Advances in medicine (such as chemotherapy, antibiotics or transplants) have allowed extending the life expectancy of patients which used to be doomed to death. Many of these patients have an impaired immune system due to treatment or disease condition, which generates a population of patients that are highly susceptible to infections. Among them, fungal pathogens have become a major source of life-threatening agents, which kill as many people as malaria or tuberculosis. To make it worse, there are very few families of antifungal drugs, and resistance towards them is increasingly reported, particularly for emerging species such as the yeast Candida glabrata. A key step towards solving the problem is understanding the molecular mechanisms of resistance, which are likely generated by adaptive mutations. We have investigated this by in vitro evolving Candida glabrata populations exposed to several of these drugs, followed by whole-genome sequencing. We have performed these experiments with fluconazole, anidulafungin and the serial combination of both, which mimic standard clinical therapy. We find mutational signatures of each drug that are consistent with previous work in pathogenic yeasts. As an example, FKS1/FKS2 mutations are widely associated to anidulafungin resistance, while PDR1/ERG11 changes appear in fluconazole. In addition, we predict a novel phenomenon of cross-resistance between these drugs through mutations in a component of ergosterol biosynthesis. Furthermore, we find events of loss of the resistance trait upon changing the treatment, which are associated with truncation of the proteins that previously conferred the resistance. All in all, this work represents a comprehensive evaluation of the evolutionary processes that confer drug resistance to Candida glabrata.

Poster 38

Microbiome and colorectal cancer: Applications for diagnosis

Olfat Khannous, Jesse Willis, Ester Saus, Toni Gabaldón

Presenting author: Olfat Khannous, Barcelona Supercomputing Center (BSC-CNS) and Institut de Recerca Biomèdica (IRB Barcelona)

Colorectal cancer (CRC) is the third most common cancer and the fourth leading cause of cancer deaths worldwide. Nowadays, one of the most used screening strategies to detect CRC is Fecal Immunochemical Test (FIT). If positive, a colonoscopy is carried out. However, most derived cases are negative after colonoscopy. Furthermore, colonoscopy is invasive, expensive, and time-consuming. For these and other reasons there is a strong interest in developing complementary approaches that will enhance the accuracy of current pre-colonoscopy diagnosis tools. The present project is framed within a larger one that aims to perform a risk model for CCR using various types of data, including those obtained in a colorectal screening program. In particular, in this project we integrate the information from the microbiome present in the samples and other metadata (diagnosis of the samples, age, sex, etc) to investigate whether individuals with a positive result from FIT and or/colonoscopy tests have differences in their microbiome profile in order to develop a risk model for CRC. Our analysis of microbiome profiles that combined with additional metadata, allow us to identify significant differences of some taxonomic ranks according to different variables. For instance, we found 26 species to have significantly different abundances depending on diagnosis. We also used genome-content inference to predict from taxonomic abundances the differential abundance of orthologous gene families that can be linked to specific pathways and metabolic modules that in many cases have been related to the pathogenesis of CRC in previous studies. On the other hand, we explored different machine learning approaches and evaluated their performance applied in our data. Integrating some of these results, a risk model will be built and will be validated in an independent set of samples.