POSTER PRESENTATIONS (updated March 22, 2016)

Posters will be up for the duration of the conference with formal presentation on Monday, April 4 during the poster and networking reception.

P01 Combinatorial Properties of Dual Graphs and Identification of Pseudoknots in RNA Secondary Structures

Louis Petingi, College of Staten Island - City University of New York US
Tamar Schlick, New York University US

In this paper we illustrate combinatorial properties of dual-graphs which have been applied to study RNA secondary structures. In particular we present a technique to partition a dual graph into connected subgraphs called blocks to detect pseudoknots, or intertwined based-paired regions. We show that a block contains a pseudoknot if and only if the block has a vertex of degree three or more. This characterization allows us to isolate smaller RNA fragments, via a new linear-time algorithm, and classify each region as either pseudoknotted or pseudoknot-free, while keeping these sub-structures intact. New topological properties of dual graphs are discussed to validate the further analysis and classification of pseudoknots based on graph-theoretic representation of RNAs.

P02 Fast and accurate computation of differential splicing in plants and animals across multiple conditions

Juan Carlos Entizne, Pompeu Fabra University, Barcelona ES
Miha Skalic, Pompeu Fabra University, Barcelona ES
Cristiane Calixto, University of Dundee UK
Runxuan Zhang, University of Dundee UK
Amadis Pages, Pompeu Fabra University, Barcelona ES
Juan Luis Trincado, Pompeu Fabra University, Barcelona ES
John Brown, University of Dundee UK
Eduardo Eyras, Universitat Pompeu Fabra, Barcelona ES

Alternative splicing plays an essential role in many cellular processes in eukaryotes and bears major relevance in development and disease. High-throughput RNA sequencing allows genome-wide analyses of splicing. However, the increasing number of available data sets represents a major challenge in terms of computation time and storage requirements.

We report on SUPPA, a computational tool to calculate relative inclusion values of alternative splicing events. SUPPA achieves accuracies comparable or higher than current methodologies thousand times faster [1]. We extended SUPPA to study alternative splicing in plants, which present frequent overlapping events and different properties of exon-intron structures compared to animals. We validate its high accuracy by using RT-PCR for over 40 events at different conditions. We further show a new method to calculate differential splicing across multiple conditions with biological replicates. We applied this method to RNA-Seq data for a time-course of Arabidopsis plants transferred from 20°C to 4°C to examine the effects of low temperature and determine new patterns in circadian clock genes. Additionally, SUPPA uses a novel density-based clustering algorithm to determine groups of events with similar patterns across conditions. We apply this to data from different stages of neuronal differentiation in human and mouse to uncover novel brain-specific splicing regulatory networks. Assessing the accuracy in terms of the choice of annotation shows the importance of current efforts aimed at completing the transcript annotation in plants and animals [2].

SUPPA calculates alternative splicing profiles and differential splicing patterns across multiple conditions and from a large number of samples at a much higher speed than existing methods without compromising accuracy, thereby facilitating the systematic analysis of very large data sets with limited computational resources [3]. SUPPA is available at https://bitbucket.org/regulatorygenomicsupf/suppa.

1. Alamancos et al. RNA 21(9):1521-31.
2. Zhang et al. New Phytol 208(1):96-101
3. Sebestyen et al. http://biorxiv.org/content/early/2015/08/02/023010

P03 NCBI---compliant genome submissions: tips and tricks to save time and money

Walter Pirovan, BaseClear B.V. NL
Marten Boetzer, BaseClear B.V. NL
Martijn Derks, Wageningen UR NL
Sandra Smit, Wageningen UR NL

Genome sequences nowadays play a central role in molecular biology and bioinformatics. These sequences are shared with the scientific community through sequence databases. The sequence repositories of the International Nucleotide Sequence Database Collaboration (INSDC, comprising GenBank, ENA and DDBJ) are the largest in the world. Preparing an annotated sequence in such a way that it will be accepted by the database is challenging because many

validation criteria apply. In our opinion, it is an undesirable situation that researchers who want to submit their sequence need either a lot of experience or help from partners to get the job done. To save valuable time and money, we list a number of recommendations for people who want to submit an annotated genome to a sequence database, as well as for tool developers, who could help to ease the process.

The corresponding publication is available at
http://bib.oxfordjournals.org/content/early/2015/12/10/bib.bbv104.full

---

P04 Universal phylo-Genomic phylo-Proteomic reference-annotation using BDGC: Big Data Genetic Code

Praharshit Sharma, Warsaw University of Life Sciences PL

A straight-forward application of 1-dimensional cellular automata to the sub-step of mRNA/cDNA-to-protein translation, within Central Dogma of Modern Biology, yields nucleotide cardinality of codon to be 0.9994 close to e= Napier's constant= base of natural logarithms= 2.718... Hence, we are left with, to explore exhaustively, to the order of R= (2^20)*(10^20) ~ 1.048576*10^26 Rules which may be characterized BOTH using algebraic/ boolean modalities. Interestingly, 0.9994 is fundamentally the correlation-coefficient of the almost linear scatter-plot of triplet-frequencies quantified from Watson- and Crick- strands of 'any' same single sequence- well in agreement with Chargaff's second parity rule. Hence, our logical-aim is to characterize the above R nodes (a Quantity the order of age of universe (in years)) as a EBDT (Entropy-Based Decision Tree) with Shannon probabilities re-computed from Phred+33 quality-scores culled from Universal datasets of raw FASTQ-reads available from SRA/ ENA/ DDBJ - for triple validation check. Many striking observations follow, such as EBDT path-traversals yielding reference genomes/MSA/PSA, extra-traversals as Structural-variants and clustered pre-/post-/in-/level-order traversals= Homo-/Para-/Ortho-/Xeno-logues.

P05 Mining the genetic background of clinical mastitis in dairy cattle using whole genome DNA sequences of 32 cows

Joanna Szyda, Wroclaw University of Environmental and Life Sciences PL
Magda Mielczarek, Wroclaw University of Environmental and Life Sciences PL
Magdalena Frąszczak, Wroclaw University of Environmental and Life Sciences PL
Tomasz Suchocki, Wroclaw University of Environmental and Life Sciences PL
Riccardo Riccardo Giannico, Fondazione Parco Tecnologico Padano IT
Giulietta Minozzi, Fondazione Parco Tecnologico Padano IT
Ezequiel Nicolazzi, Fondazione Parco Tecnologico Padano IT
Katarzyna Wojdak-Maksymiec, West Pomeranian University of Technology PL

Udder infections exhibit high prevalence in dairy cattle and pose a major problem for the industrial milk production process. The goal of our study was to utilise information on whole genome DNA sequences of 32 cows to better understand the genetic background of resistance to clinical mastitis, which is the most common infectious udder disease.

The sequenced animals were selected out of the data base of 991 Holstein-Friesian cows comprising individuals with clinical mastitis cases diagnosed by a veterinarian and their healthy herd-mates. The experimental design comprised 16 pairs of paternal half-sisters matched in terms of parity, production level and birth year, but differing in terms of mastitis resistance. The samples were sequenced on an IlluminaHiSeq 2000 Next Generation Sequencing platform. The total number of raw reads for a single animal varied between 164,984,147 and 472,265,620 with the corresponding genome-average coverage of 14.03. Raw reads were filtered and trimmed with Trimmomatic and aligned to the UMD3.1 reference genome using BWA-MEM. Variant calling was done using FreeBayes and Copy Number Variation (CNV) was identified by CNVnator.

Three approaches were incorporated to identify genomic regions corresponding to clinical mastitis resistance. First, we identified, separately for each half-sib pair, all single nucleotide polymorphisms (SNPs) whose genotypes differed between a resistant and a disease-prone half-sib, followed by the selection of SNPs consistently differing in all analysed pairs. Second, the same approach was applied to CNV type polymorphism. Third, was based on formal hypotheses testing of SNP allele frequency differences between resistant and prone group using the Odds Ratio test, followed by the multiple testing correction of resulting nominal P values accounting for linkage disequilibrium pattern. Finally regions identified in the analysis will be annotated to the UMD3.1 reference genome in order to reveal underlying genomic features with the focus on genes and regulatory sequences.

P06 Pathways Constrained Therapeutic Decision-Making Tool in Acute Myeloid Leukemia Personalized Medicine

Marco Manfrini, University of Bologna IT
Giorgia Simonetti, University of Bologna IT
Antonella Padella, University of Bologna IT
Daniel Remondini, University of Bologna IT
Italo Faria Do Valle, University of Bologna IT
Giovanni Marconi, University of Bologna IT
Cristina Papayannidis, University of Bologna IT
Gastone Castellani, University of Bologna IT
Giovanni Martinelli, University of Bologna IT

Personalized medicine will strengthen the prognosis and provide better, patient-specific drug identification. Massive parallel sequencing showed that a number of new discovered single

nucleotide variants (SNV) and insertion/deletions could affect Acute Myeloid Leukemia (AML) patients survival time and drug resistance. To exploit these informations for prognosis and personalized therapies we propose a web application which is based on a statistical model learned from data provided by NGS technology, patients cytogenetics predictive risk and clinical outcomes, which could be useful to predict drug sensitivity and prognosis for new diagnosed patients. DNA from a cohort of 65 AML patients from S. Orsola University Hospital, Bologna, Italy, were analyzed by Whole Exome Sequencing (WES). Briefly, pre-processing was carried out following GATK guidelines and re-calibrated alignment files were analyzed by MUTECT and VARSCAN2 algorithms in order to detect single nucleotide variants (SNV) and indels. An additional AML dataset containing WES data was downloaded from TCGA database in order to collect somatic variants and clinical data. Somatic mutations for each patients were mapped to Reactome and Oncotar databases. Binary matrix factorization was applied to the patients/pathways matrix in order to find pathways signatures capable to cluster patients. To model pathways constrained mutational data toward event free survival or drug resistance we employed a generalized linear model with binary outcome, including cytogenetics risk and presence/absence of cancer genes driver mutations. Regression model was applied with 10-fold cross validation to find statistically significant explanatory variable. AUCs were determined to assay classifier performances. Five pathways signatures were obtained from 2441 non-silent somatic mutations allowing clustering of patients in three groups. Regression analysis provided a model which allow classification of patients in responder/non-responder to standard chemotherapy and in event free survival low risk group. This approach is useful to selection of standard therapy versus experimental therapy.

P07 Identification of Loci related to Humoral Immunity in Chicken

Vahid Raeesi, Tarbiat Modares University IR
Alireza Ehsani, Tarbiat Modares University IR
Rasoul Vaez Torshizi, Tarbiat Modares University IR
Aliakbar Masoudi, Tarbiat Modares University IR
Rahim Dideban, Tarbiat Modares University IR

Background: The problems arising from diseases and infections are still a major challenge in livestock industry. Immunity related traits are heritable in chicken, therefore, it is possible to improve the inherent immunity by breeding programs. In this study using chicken 60k SNP chip, we performed genome-wide association study to determine candidate genes and loci responsible for 7 immunity related traits including lysozyme activity against Micrococus luteus bacteria, primary and secondary antibody mediated responses against Sheep Red Blood Cell (SRBC) including total serum immunoglobulin, total serum immunoglobulin Y, and total serum immunoglobulin M (IgM). An F2 design was used at the experimental farm in Tarbiat Modares University by reciprocal crosses between Arian and Urmia chickens, representing a commercial

meat type breed and an Iranian indigenous strain, respectively. Statistical analysis was based on a mixed linear model utilizing genomic prelateship matrix to prevent spurious associations. Correction for multiple testing was done by applying 5% and 10% chromosomal false discovery rates as significant and suggestive thresholds, respectively.

Results: This study has revealed 27 SNPs associated to seven immune traits in an F2 chicken. Ten significant and 17 suggestive SNPs were identified. Most of the SNPs that were suggestively associated with total serum immunoglobulins in primary response were significantly associated with this trait in secondary response too and three SNPs were located within a narrow region of 23kb on chromosome 16. Pathway analysis in database for annotation and visualization and the integrated discovery (DAVID) using candidate genes within 250,000 base pairs of significant SNPs showed that antigen processing and presentation is significantly associated with significant SNPs.

Conclusions: Existence of significant differences in the genome of chickens with different immune responses may imply that these differences could be use in selection indices to enhance immunity system.

P08 Statistical method to identify novel transcription factor mutual interactions

Nisar A. Shar, University of Leeds UK
M.S. Vijayabaskar, University of Leeds UK
David R. Westhead, University of Leeds UK

Transcription factors have important roles in the regulation of genes; they interact with each other to form complexes and bind to DNA where they influence gene expression through a variety of mechanisms. Identification of mutual interactions of transcription factors would help in understanding the regulation of genes, which ultimately will help us in understanding the erroneous regulation associated with some cancers. Here, we have developed a statistical method to identify interactions, using ENCODE ChIP-seq data, and found several novel interactions in the ENCODE cell lines. We found that transcription factor pairs occupy varying sizes of genomic region, often with one binding directly on the open chromatin while the other binding indirectly. We have validated our observations by various statistical methods, including randomization and the Poisson distribution. It was also found that shared binding sites for a particular transcription factor in multiple cells lines are more evolutionary conserved than binding sites unique to particular cell lines. Equally, overlapped binding sites for transcription factor pairs are more evolutionary conserved than binding sites for single transcription factors in most of the TF pairs. Evolutionary conservation gives an indication that transcription factor binding sites probably have important biological functions. Thus transcription factor mutual interactions should have some role in controlling gene expression. We found that genes which were located near the co-bound sites have higher expression levels than the genes which were

located near single bound sites.

P09 The assessment of whole NGS pipelines containing alignment and SNP calling tools

Magda Mielczarek, Department of Genetics, Wroclaw University of Environmental and Life Sciences PL
Joanna Szyda, Department of Genetics, Wroclaw University of Environmental and Life Sciences PL
Bernt Guldbrandtsen, Department of Molecular Biology and Genetics, Aarhus University DK

The analysis of data originated from next generation sequencing platforms (NGS) is highly dependent on software, which is generally applied for data editing, alignment, and variant calling. Unlike many studies devoted to comparing programmes for one particular step of NGS data analysis, in this study we compared whole pipelines consisting of aligners and variant callers, in order to account for possible in silico interactions between various programmes. In particular, these pipelines form eight different combinations of mappers (BWA-MEM, BWA-backtrack, Bowtie2, SMALT) and SNP callers (SAMtools, GATK). The overall quality of each pipeline was assessed by comparing a final calls of SNP genotypes based on sequence data (Illumina HIseq 2000) to SNP genotyping microarray (Illumina BovineHD BeadArray 770K SNP) genotype calls. Because the microarray based genotypes have a much lower technical error rate, they were as the reference in this comparison. This study was performed based on the whole genome DNA sequence reads of four traditional Danish Red Dairy Cattle bulls with an average coverage of 10X.

As a result we observed differences between pipelines, but their accuracies expressed by the comparison with microarray based genotypes were similar and high. Moreover, true SNPs and true SNP genotypes were mainly characterized by the high qualities of calls accompanying them, at the same time, confirming the high accuracy of pipelines. Nevertheless, the recommended pipeline consists of the BWA-MEM and GATK programmes which performed well together. GATK detected the highest number of true SNPs and one of the highest numbers of true genotypes for BWA based output. Also, the computation time of this pipeline was among the shortest.

P10 Combining multiple tools outperforms individual methods in gene set enrichment analyses

Monther Alhamdoosh, CSL Limited AU
Milica Ng, CSL Limited AU
Nicholas J. Wilson, CSL Limited AU

Julie M. Sheridan, Walter and Eliza Hall Institute of Medical Research AU
Huy Huynh, CSL Limited AU
Michael J. Wilson, CSL Limited AU
Matthew E. Ritchie, Walter and Eliza Hall Institute of Medical Research AU

Gene set enrichment (GSE) analysis allows researchers to efficiently extract biological insight from long lists of differentially expressed (DE) genes by interrogating them at a systems level. In recent years, there has been a proliferation of GSE analysis methods and hence it has become increasingly difficult for researchers to select an optimal GSE tool based on their particular data set. Moreover, the majority of GSEA methods do not allow researchers to simultaneously compare gene set level results between multiple experimental conditions.

The ensemble of gene set enrichment analyses (EGSEA) combines the enrichment analysis results of multiple methods and calculates collective gene set scores to improve the biological relevance of the highest ranked gene sets in RNA-sequencing experiments. EGSEA has multiple visualization capabilities that allow researchers to view gene sets at various levels of granularity. EGSEA has been tested on a number of human and mouse data sets and, based on biologists' feedback, consistently outperforms the individual tools that have been combined. Our evaluation demonstrates the superiority of the ensemble approach for GSE analysis, and its utility to effectively and efficiently extrapolate biological functions and potential involvement in disease processes from lists of differentially regulated genes.

P11 Workflow-BS an integrative workflow for RRBS and WGBS data. From the BS-seq to the DMR

Gaëlle Lefort, INRA FR
Marjorie Mersch, INRA FR
Sylvain Foissac, INRA FR
Frédérique Pitel, INRA FR
Céline Noirot, INRA FR

DNA methylation is an epigenetic mark that has suspected regulatory roles in a broad range of biological processes and diseases. The technology is now available for genome-wide methylation studies, at a high resolution and with possibly a large number of samples. Many specific aligners for BS-seq data exist, such as BSMAP and Bismark. Also, R packages (methylKit and DSS) were designed to detect differentially methylated cytosines (DMC) and differentially methylated regions (DMR). Methy-Pipe (Peiyong Jiang et al. 2014. PLOS one) fill the gap between those analyses by combining a complete pipeline from raw data to statistical outputs but it requires a specific cluster environment (SGE software). Here, we propose a workflow which deals with fastq files from BS-seq (WGBS and RRBS) and goes through all steps to provide bed files of DMC and DMR. It can support most distributed resource management systems (Condor, SGE, …).

Our pipeline uses standard software to i) clean data ii) align WGBS or RRBS reads to a reference genome iii) extract methylation and iv) identify DMC and DMR. Raw data are cleaned with Trim_galore and aligned with Bismark. The base-resolution methylation level is extracted by context and sample with methylKit. If a SNP file is provided, its polymorphic positions are removed from the analysis. Several tests to detect DMC and DMR are then performed, according to the experimental design supplied by the user. Statistics and graphics are also provided. As the pipeline is based on Jflow (Mariette et al. 2015. Bioinformatics), it can be used on command line or through a web server. Adding a new aligner or a new component has been made as simple as possible for future evolution of the tool.

We will present results obtained by using this pipeline on chicken and plant genomes.

P12 Profiling RNA editing in Human Single Cells

Ernesto Picardi, University of Bari & IBBE-CNR IT
Anna Maria D'Erchia, University of Bari & IBBE-CNR IT
Graziano Pesole, University of Bari & IBBE-CNR IT

A-to-I RNA editing is a widespread co/post-transcriptional modification affecting coding and non-coding human transcripts at hundred million sites. The deamination of adenosines to inosines has a plethora of biological effects depending on the RNA region involved in the modification. The genome-wide detection of A-to-I events has been largely facilitated by the advent of NGS technologies. Very recently, we have released a comprehensive human inosinome atlas profiling RNA editing in six different tissues and detecting overall 3,041,422 events (PMID: 26449202). As expected, we found that 97% of atlas events were in repetitive regions including ~90% in Alu elements, while only a limited amount of sites fell in non-repetitive regions (3%). The number of predicted A-to-I events differed greatly among samples because of sequencing depth variation, specific filters used to recover editing candidates and tissue specific roles of RNA editing.

Tissue samples consist of heterogeneous cell populations and, thus, investigated molecular effects recapitulate this molecular complexity. As a consequence, different cell types occurring in a tissue may bias molecular parameters as gene or transcript expression levels. To capture and characterize the complexity of RNA editing at single cell resolution, we investigated this phenomenon in single cells from adult human cortex obtained from living subjects in which transcriptome diversity was already surveyed by single cell RNA sequencing (scRNA-seq) (PMID: 26060301). Using our REDItools (PMID: 23742983) and a comprehensive collection of known RNA editing events, we explored inosinome profiles in 466 cortex cells. We found that the detection of A-to-I events is strongly correlated with the amount of RNA-seq reads and editing profiles are quite heterogeneous also inside the same cell population. However, the observed RNA editing profile is sufficient to discriminate major cell types as neurons, astrocytes and

oligodendrocytes, underlining the cell specific nature of RNA editing.

P13 Integration of Ixodes Ricinus genome sequencing with transcriptome and proteome annotation in the naïve

Wibke J. Cramaro, Luxembourg Institute of Health LU
Dominique Revets, Luxembourg Institute of Health LU
Oliver E. Hunewald, Luxembourg Institute of Health LU
Regina Sinner, Luxembourg Institute of Health LU
Claude P. Muller, Luxembourg Institute of Health LU

In Europe, Ixodes ricinus ticks are the most important vectors of diseases threatening humans, wildlife and companion animals. Nevertheless, genomic sequence information was missing and functional annotation of transcripts and proteins limited. This lack of information is restricting studies of the vector and its interactions with pathogens and hosts. We presented and integrated the first analysis of the I. ricinus genome with the transcriptome and proteome of the unfed I. ricinus midgut (Cramaro et al. 2015). The de novo assembly of 1 billion Illumina sequences to a reference genome of 393 Mb length provided an unprecedented insight into the I. ricinus genome. Homologous sequences to 89 % of the I. scapularis genome scaffolds indicate coverage of most genome regions, even if the coverage is partial. We identified moreover 6,415 putative new genes. In order to further elucidate the genetic structure, we now scaffolded the Illumina contigs with PacBio reads to 515 Mb genomic reference sequence.

14 % of European I. ricinus ticks are infected with members of the Borrelia burgdorferi sensu lato complex and Lyme borreliosis is the most important vector-borne disease in Europe. Interactions between I. ricinus and Borrelia in the midgut are essential for successful survival of the pathogen in the tick as well as its transmission to the host. Therefore, midgut proteins are important players in vector-pathogen interactions and potential targets for blocking feeding and transmission by vaccines.

By combining protein identification by mass spectrometry with RNA-sequencing, we annotated more than 10,000 transcripts and 285 proteins expressed in the midgut of unfed I. ricinus ticks for function, localization and biological processes.

This multiple-omics study provides important annotated data of I. ricinus, paving the way for further investigation of tick-pathogen interactions as well as for the identification of vaccine candidates to potentially control several vector-borne diseases.

P14 Short term exposure to oxygen and glucose deprivation induces widespread alterations in translation and synthesis of novel stress-specific proteoforms

Dimitry Andreev, Lomonosov Moscow State University RU
Patrick O'Connor, University College Cork IE
Alexandr Zhdanov, University College Cork IE
Ruslan Dmitriev, University College Cork IE
Ivan Shatsky, Lomonosov Moscow State University RU
Dmitry Papkovsky, University College Cork IE
Pavel Baranov, University College Cork IE

Oxygen and glucose metabolism play pivotal roles in many (patho)physiological conditions. Using time-resolved ribosome profiling we assessed events at the level of gene expression in neural cell line, PC12, during the first hour of complete OGD. The most substantial alterations were seen to occur within the first 20 minutes of OGD. While transcription of only about a hundred of genes was significantly altered during one hour of OGD, translation response affected about 3000 genes. This response involved reprogramming of initiation and elongation rates as well as stringency of start codon recognition. Genes involved in oxidative phosphorylation were affected the most. Detailed analysis of ribosome profiles revealed salient alterations of ribosome densities on individual mRNAs. The mRNA specific alterations include increased translation of upstream open reading frames (uORFs), site-specific ribosome pauses and production of stress-specific proteoforms with N-terminal extensions. The detailed analysis of ribosomal profiles also revealed two examples of dual coding mRNAs, where two protein products are translated from the same long segment of mRNA, but in two different reading frames. These findings uncover novel regulatory mechanisms of translational response to OGD in mammalian cells, different from the classical pathways such as Hypoxia Inducible Factor (HIF) signaling and also reveal sophisticated organization of protein coding information in certain genes.

P15 RiboSeq.Org for aligning, analyzing and exploring ribo-seq data

Audrey Michel, School of Biochemistry and Cell Biology, University College Cork, Cork IE
James P. A. Mullan, School of Biochemistry and Cell Biology, University College Cork, Cork IE
Stephen Kiniry, School of Biochemistry and Cell Biology, University College Cork, Cork IE
Patrick B. F. O'Connor, School of Biochemistry and Cell Biology, University College Cork, Cork IE
Pavel Baranov, University College Cork, Cork IE

The ribosome profiling (ribo-seq) technique enables the locations of actively translating ribosomes to be determined at a genome-wide level. On RiboSeq.Org (http://riboseq.org/) we provide freely available resources to help researchers analyse and explore ribo-seq data without having to use command-line tools. RiboGalaxy [1] is a Galaxy-based web server where researchers can pre-process, align, analyze and visualize their own ribo-seq data. GUI-based tools are provided to determine the strength of the triplet periodicity signal in ribo-seq data, generate metagene and ribosome profiles and carry out differential translation expression

analysis using the riboSeqR suite of tools in RiboGalaxy. The RUST suite of tools can be used to quickly characterize ribosome profiling datasets to assess their quality as well as analyze the relative impact of different mRNA sequence features on local decoding rates. The RiboTools suite provides functionality for exploring translation in alternative reading frames and stop codon readthrough events. GWIPS-viz [2] is an on-line genome browser which provides pre-populated ribo-seq and mRNA-seq tracks from many published studies. Users can use the GWIPS-viz workflows available in RiboGalaxy to explore and compare their own ribo-seq data to published data. In addition we provide help pages and a forum (http://gwips.ucc.ie/Forum/) where we encourage users to post their questions and feedback.

1. Michel AM, Mullan JPA, Velayudhan V, O'Connor PBF, Donohue CA, Baranov PV. RiboGalaxy: a browser based platform for the alignment, analysis and visualization of ribosome profiling data. RNA Biology (in press).

2. Michel AM, Fox G, M Kiran A, De Bo C, O'Connor PBF, Heaphy SM, Mullan JPA, Donohue CA, Higgins DG, Baranov P V. GWIPS-viz: development of a ribo-seq genome browser. Nucleic Acids Res 2014; 42:D859–64.

P16 sRNA-PlAn : profiling and annotating sRNAseq datasets in plants and animals

Olivier Rué, INRA Toulouse FR
Philippe Bardou, INRA Toulouse FR
Jerome Mariette, INRA Toulouse FR
Matthias Zytnicki, INRA Toulouse FR
Christine Gaspin, INRA Toulouse FR

http://gwips.ucc.ie/Forum/

The development of new high-throughput sequencing technologies has accelerated the discovery of very short non coding RNA (ncRNA), also called small non coding RNAs by reference to small RNAseq sequencing (sRNAseq) protocols. In eukaryotes, these small ncRNA generally are of size less than 30nt. In plants as in animals, high throughput sequencing resulted in a rapid increase of catalogued miRNA, siRNA and piRNA. It also enlarged the field of small RNA research by revealing the existence of many novel small RNA species such as short RNA products from rRNA, snoRNA, tRNA...

In sRNAseq data sets, only a very small fraction of reads can be assigned to a known functionnal family resulting in a lot of sRNAseq data orphan of functional annotation. The bias introduced by errors, the editing of some sequences but also the lack of similarities in existing ncRNA databases make challenging their structural and functional annotation. sRNAseq data analysis tools such as miRDeep[1], miRanalyzer[2] and others focus on microRNAs annotation and prediction, neglecting other types of RNAs. Recently, web tools such as DARIO[3], Ncpro[4] enlarged functional annotation.

We present sRNA-PlAn and sRNAbrowse, a sequel of software under development which aim at profiling, annotating and exploring as many sRNAseq data as possible, considering different ncRNA families and differential expression in multiple conditions and/or tissues. Preliminary results are accessible at http://ngspipelines.toulouse.inra.fr:9064.

References

[1] Friedländer et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol. 2008 ;26(4):407-15

[2] Hackenberg et al. Miranalyzer: a microrna detection and analysis tool for next-generation sequencing experiments. NAR 2009, 37, W68–W76

[3] Chen et al. Ncpro-seq: a tool for annotation and profiling of ncrnas in srna-seq data. Bioinformatics 2012, 28 (23), 3147–3149

[4] Fasold et al. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. NAR 2011, 39, W112-117

P17 Accurate detection of novel chimeric transcripts by integration of spanning and paired-end read information

Bernardo Rodríguez Martín, Super Computing Center, Barcelona ES
Emilio Palumbo, Centre for Genomic Regulation, Barcelona ES
Santiago Marco-Sola, Centro Nacional de Análisis Genómico, Barcelona ES
Paolo Ribeca, The Pirbright Institute UK
Graciela Alonso, Centro de Biología Molecular Severo Ochoa, Madrid ES
Alberto Rastrojo, Centro de Biología Molecular Severo Ochoa, Madrid ES
Begoña Aguado, Centro de Biología Molecular Severo Ochoa, Madrid ES
Roderic Guigó, Centre for Genomic Regulation, Barcelona ES
Sarah Djebali, Centre for Genomic Regulation, Barcelona ES

Chimeric transcripts are commonly defined as transcripts encoded by two different genes in the genome, and can be explained by various biological mechanisms such as genomic rearrangement, read-through or trans-splicing, but also by technical or biological artefacts. Several papers have shown their importance in cancer, others their role in cell pluripotency.

Several programs have been developed to identify chimeras from RNA-seq data (mostly fusion genes in cancer), however their outputs on the same dataset can be widely different, and tend to include many false positives. Other issues relate to unrealistic simulated datasets often restricted to fusion genes, real datasets with limited numbers of validated cases, result

inconsistency between simulated and real datasets, and gene rather than junction level assessment.

Here we present ChimPipe, a modular and easy-to-use RNA-seq pipeline performing an exhaustive mapping based on the GEM tools, integrating both spanning and paired-end reads, and using a stringent filtering module, to identify highly reliable chimeras. We have also produced realistic simulated datasets of three different read lengths, and enhanced two gold standard cancer datasets by associating exact junction points to validated gene fusions. Benchmarking ChimPipe together with four other state of the art tools on this data, showed ChimPipe to be the best program at identifying exact junction coordinates for both kinds of datasets, and the second most sensitive one at the gene level but with 1 to 1.6 order of magnitude less identified cases, therefore showing the best balance between sensitivity and precision.

Applied to 108 encode human RNA-seq samples, ChimPipe identified 33 highly supported chimeras connecting the coding sequence of their parent genes, of which six were attempted to be verified by RT-PCR, and of which three succeeded. Cloning and sequencing of these three cases revealed several new chimeric transcript structures, possibly encoding proteins with novel functions.

P18 ContrastRank: a new tool for ranking gene candidates in complex diseases

Jaume Sastre Tomas, University of the Balearic Islands (UIB) ES
Alexander Damia Heine Suñer, Idispa ES
Emidio Capriotti, University of Düsseldorf, Düsseldorf DE
Jairo Enrique Rocha Cárdena, University of the Balearic Islands (UIB) ES

Recent advances in next generation sequencing (NGS) have provided such a huge amount of data that is even beyond the analysis capacity of the scientific community. Therefore, it has become a necessity, the development of new bioinformatic tools for the detection of genetic variants associated with many genetic disorders.

An important factor in this analysis is the distinction between mutations that affect functionality and promote the onset of the disease (drivers) and mutations that do not have any harmful effect (passengers). Due to this need, emerge the development of a new prioritization method that assigns a score to each gene that corresponds to the possibility that it might be related to a disease.

The work carried out by Dr. Emidio Capriotti et al. [1], apply ContrastRank for the mutations prioritization in the development of important diseases such as lung, colon or prostate cancer. They obtained an ordered list of the most important related genes to these diseases.

Our work pretends to extend the Dr. Capriotti's applying ContrastRank in three new types of cancer, breast, ovarian and glioblastoma, as well as adapting the algorithm to work with any other complex disease, such as congenital heart disorders (CHD).

The method assigns a score to each gene of the exome according to their probabilities to be originators of the disease. As a reference, we used the allele frequencies from 1000Genomes project [2,3] and cancer data from (The Cancer Genome Atlas) and Congenital Heart Disease Genetic Network Study for CHD.

As a result, we will obtain an ordered list of genes related to the onset and development of the genetic disorders under study.

These candidate genes might be either known genes or new ones belonging to important signaling pathways for cancer or other genetic diseases and becoming good new drug targets for future research.

P19 gmove: Eukaryotic gene predictions using various evidences

Marion Dubarry, CEA/IG/Genoscope FR
Benjamin Noel, CEA/IG/Genoscope FR
Tsinda Rukwavu, CEA/IG/Genoscope FR
Sarah Farhat, CEA/IG/Genoscope FR
Corinne Da Silva, CEA/IG/Genoscope FR
Manuel Lebeurrier, CEA/IG/Genoscope FR
Jean-Marc Aury, CEA/IG/Genoscope FR

The NGS make the sequencing faster and cheaper, so affordable for more and more laboratories. Consequently, the number of sequenced genome explodes, and it becomes feasible to sequence more complex genomes (like large genomes, highly repeated genome) as well as non-model organisms. Pipelines of gene predictions have to get used to these technological improvements to keep going to improve the quality of their predictions, make easier the calibration step and handle the large amount of data available.

We present the Eukaryotic genome annotation pipeline used at Genoscope (the French national sequencing center), particularly the tool we use to predict coding genes, named Gmove (Gene Modelling using Various Evidence). It can use several source of data, like RNAseq, conserved proteic alignment and ab initio gene predictions. Gmove combines these data and finds a consensus without any prerequisite calibration. A graph is built where a node represents an exon and a vertex represents an intron, extracting paths are potential genes models. In these models, Gmove searches for an open reading frame based on existing protein alignments. We can select one or several isoform of a given gene. On the de novo annotation context, we already run this tool on a variety of current projects such as plant, fungus, insect and dinoflagellate (genomes not published yet). Moreover, Gmove could also improve existing gene

prediction by combining the former annotation with new data.

P20 RNAcentral: An international database of ncRNA sequences

Anton I. Petrov, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) UK

Rnacentral Consortium, RNAcentral Consortium UK

The field of non-coding RNA biology has been hampered by the lack of a comprehensive, up-to-date collection of accessioned RNA sequences. We begin to address this challenge by creating RNAcentral, a database of non-coding RNA sequences aiming to represent all non-coding RNA types from all organisms. RNAcentral serves as a single entry point for searching and accessing data from over twenty established RNA resources such as miRBase, RefSeq, GENCODE/Vega, and Rfam, with over nine million distinct sequences collected to date. The RNAcentral website (http://rnacentral.org) provides free interactive and programmatic access to data, faceted keyword search, and cross-references to external resources. Where possible, sequences are mapped onto reference genomes from key species and made available in an integrated genome browser as well as in Ensembl and UCSC genome browsers. RNAcentral also enables sequence similarity searches across multiple RNA databases using a web interface, which is a unique capability worldwide. RNAcentral brings a unified set of identifiers to the field of non-coding RNAs, thus providing a common framework upon which annotations can be applied, and a future goal for RNAcentral is to incorporate curated information about all non-coding RNAs as UniProt does for proteins. We began this next phase of development by importing nucleotide modification information from Modomics, and in the near future we plan to integrate functional annotations of non-coding RNAs, such as intermolecular interactions and high-quality secondary structures.

P21 Resequencing and assembly of seven complex loci to improve the Leishmania major (Friedlin strain) reference genome

Graciela Alonso, CBMSO ES
Alberto Rastrojo, CBMSO ES
Sara Lopez-Perez, CBMSO ES
Jose M. Requena, CBMSO ES
Begoña Aguado, CBMSO ES

Background: Leishmania parasites cause severe human diseases known as leishmaniasis. These eukaryotic microorganisms possess an atypical chromosomal architecture and the regulation of gene expression occurs almost exclusively at post-transcriptional levels. Accordingly, sequencing of the genome of Leishmania major and other related species, was paramount for

highlighting these peculiar molecular aspects. Recently, we performed an analysis of gene expression by massive sequencing of RNA in the L. major promastigote, and data derived from this analysis were suggestive of possible errors in the current genome assembly for this Leishmania species.

Results: During the analysis by RNA-Seq of the transcriptome for L. major Friedlin strain, 163,714 reads could not be aligned with the reference genome. Thus, de novo assembly with these reads was carried out and the resulting contigs were analyzed. After detailed homology searches in databases, it was postulated that 15 contigs might correspond to genomic sequences lost during the genome assembly of the L. major Friedlin strain. This was experimentally confirmed by PCR amplification, cloning and sequencing of the new genomic regions. As a result, we have identified seven regions of the L. major (Friedlin) genome that were lost during the sequence assembly. This led to the uncovering of six new genes (LmjF.15.1475, LmjF.15.0285, LmjF.24.0765, LmjF.14.0860, LmjF.19.0305, and LmjF.27.2035), and correction of the annotation for two others (LmjF.15.1480 and LmjF.27.2030). Our data suggest that these genomic regions probably collapsed during the genome assembly due to the existence of gene duplications and/or repeated regions surrounding the missed genes.

Conclusion: RNA-seq data helped to reconstruct some genomic regions misassembled during the L. major Friedlin genome assembly, which is otherwise quite robust. On the other hand, this study shows that data derived from massive sequencing approaches, including RNA-Seq, should be carefully inspected to improve current genome definition and gene annotations.

P22 A standalone tool for finding ORFs and reconstructing potential protein isoforms from RNA-Seq data

Manal Alsheri, University of Windsor CA
Abed Alkhateeb, University of Windsor CA
Iman Rezaeian, University of Windsor CA
Luis Rueda, University of Windsor CA

Translation, as the second step of central dogma in molecular biology, is a process for translating mRNAs into polypeptide chains. An open reading frame (ORF) is a continuous sequence of codons that begins with a start codon and ends with one of the different stop codons. Finding ORFs corresponding to a given mRNA transcript, is an important step in reconstructing protein isoforms, which is vital for better understanding of RNA alternative splicing effects in diseases like cancer.

There are some computational tools for finding ORFs such as ExPASy and ORF Finder. However, they suffer from several shortcomings. First of all, every mRNA, in general, contains only one main ORF, whereas the existing tools detect all regions between any start codon and stop codon in a given sequence without determining the actual ORF. Second, finding ORFs in a set of

transcripts is processed individually for each sequence. Lacking a batch mode for finding protein isoforms and process each transcript individually, when dealing with a large number of transcripts, makes the whole process time consuming and not convenient for the user.

In our approach, we have developed a tool for finding all six frames corresponding to a given mRNA sequence and identifying finding the corresponding ORFs from RNA-Seq data. The ability to process several transcripts in batch mode, makes this tool an ideal software, when dealing with protein isoform prediction for a large number of transcripts. Moreover, the proposed tool is able to find both known and novel protein isoforms from RNA-Seq data. Using this tool we have been able to identify several novel transcripts including an extra exon in the 5' UTR of WWP2. WWP2 has been shown to be related to several types of cancer.

P23 De novo identification, differential analysis and functional annotation of SNPs from RNA-seq data in non-model species

Hélène Lopez-Maestre, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Lilia Brinza, BIOASTER FR
Camille Marchet, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Janice Kielbassa, Synergie-Lyon-Cancer, Université Lyon 1 FR
Sylvère Bastien, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Mathilde Boutigny, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
David Monin, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Adil El Filiali, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Claudia Carareto, Department of Biology, UNESP - São Paulo State University, BR
Cristina Vieira, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Franck Picard, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Natacha Kremer, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Fabrice Vavre, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Marie-France Sagot, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Vincent Lacroix, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR

SNPs (Single Nucleotide Polymorphisms) are genetic markers whose precise identification is a prerequisite for association studies. Methods to identify them are currently well developed for model species, but rely on the availability of a (good) reference genome, and therefore cannot be applied to non-model species. They are also mostly tailored for whole genome (re-)sequencing experiments, whereas in many cases, transcriptome sequencing can be used as a cheaper alternative which already enables to identify SNPs located in transcribed regions.

In this paper, we propose a method that identifies, quantifies and annotates SNPs without any reference genome, using RNA-seq data only. Individuals can be pooled prior to sequencing, if not enough material is available for sequencing from one individual.

Using human RNA-seq data, we first compared the performance of our method with GATK, a well established method that requires a reference genome. We showed that both methods predict SNPs with similar accuracy.

We then validated experimentally the predictions of our method using RNA-seq data from two non-model species.

The method can be used for any species to annotate SNPs and predict their impact on proteins. We further enable to test for the association of the identified SNPs with a phenotype of interest.

The method is distributed as a pipeline: http://kissplice.prabi.fr/TWAS.

P24 Identification of rag genes mutations in human primary immunodeficiency diseases using exome sequencing

Reza Mahdian, Pasteur Institute of Iran
Hamzeh Rahimi, Pasteur Institute of Iran
Shirin Shahbazi, Tarbiat Modares University Iran

Background and Purpose: Primary immunodeficiencies (PIDs) are more than 250 rare diseases with similar clinical symptoms. They are caused by gene defects in proteins required for normal immune function and classified into nine categories based on immunological laboratory assay. We utilized the whole exome sequencing as it has demonstrated to be a potent and efficient tool for identifying causative genetic defects in rare mendelian disorder.

Methods: We applied whole exome sequencing to study patients with lymphohistiocytosis in a consanguineous family. The sequence analysis platform contain five steps; quality control, genome mapping and alignment, post-alignment processing, variant calling, variant annotation. The variants effects on protein function were investigated by Sift, Polyphen, Extasy, CADD, ClinVar.

Results: The data was annotated and subjected to quality control. Our results showed 49,454 variants and 19450 rarer SNPs with MAF less than 0.1. Following data filtering and gene prioritization, nine variants were candidate for structural analysis. By further processing, two RAG genes mutations were found as disease causing variants. Functional studies will describe the impact of these variants on protein expression and activation.

Conclusions: More than 150 genes have been implicated in PIDs including the genes that follow the mendelian inheritance. An efficient tool for mutation analysis in these pedigrees enhances genetic counselling and prenatal diagnosis. Our findings draw attention to the significance of whole exome sequencing as powerful and cost-effective strategy in the diagnosis of

heterogeneous disorders including PIDs.

P25 De novo assembly and preliminary analysis of the spider crab Maja brachydactyla transcriptome

Rosa M. Garcia-Junco, Grupo de Investigación en Bioloxía Evolutiva (GIBE), Departamento de Bioloxía Celular e Molecular, Facultade de Ciencias and Centro de Investigaciones Científicas Avanzadas (CICA), Universidade da Coruña ES

Andrés Martínez-Lage, Grupo de Investigación en Bioloxía Evolutiva (GIBE), Departamento de Bioloxía Celular e Molecular, Facultade de Ciencias and Centro de Investigaciones Científicas Avanzadas (CICA), Universidade da Coruña ES

Ana M. González-Tizón, Grupo de Investigación en Bioloxía Evolutiva (GIBE), Departamento de Bioloxía Celular e Molecular, Facultade de Ciencias and Centro de Investigaciones Científicas Avanzadas (CICA), Universidade da Coruña ES

Lars Kaderali, Universitätsmedizin Greifswald, Institut für Bioinformatik, Greifswald DE

The spider crab Maja brachydactyla is a brachyuran decapod crustacean occurring in the north-east Atlantic. This species is subject of a heavy exploitation pressure in the coastal area comprehended between Morocco and the north of France and ecological data suggest that there might be overexploitation problems in this region. Several studies about the spider crab have been carried out from an ecological and biochemical perspective but there is very little information on the molecular genetics aspects.

High-throughput sequencing technologies have notably changed our approach to genome and transcriptome analysis. Nowadays it is possible to generate large data sets of sequences at a relatively low cost in price and time. When working with a non-model organism the available genetic information is usually scarce. RNA sequencing (RNA-seq) is a very useful tool for gene expression profiling, gene identification and mining of molecular markers. Even lacking a reference genome, it is possible to assembly a transcriptome de novo as a starting point for gene expression studies. In the last few years RNA-seq technologies have made possible the transcriptome analysis of several crustacean non-model species, widening our knowledge about different biological aspects as signaling, reproduction and development.

In the present study we generate a dataset for the transcriptome of an adult female Maja brachydactyla. We performed Illumina paired-end sequencing of gut, gill, hepatopancreas and heart. The resulting sequences were assembled to contigs and analysed. We obtained 120698 unigenes, of which 29849 were successfully annotated. It has been possible to identify several growth-related genes and a new panel of SSR candidates is being developed, providing a starting point for further investigation in this species.

This research has been supported by a grant from Ministerio de Economía y Competitividad (MINECO), reference CTM2014-53838-R.

P23 De novo identification, differential analysis and functional annotation of SNPs from RNA-seq data in non-model species

Hélène Lopez-Maestre, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Lilia Brinza, BIOASTER FR
Camille Marchet, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Janice Kielbassa, Synergie-Lyon-Cancer, Université Lyon 1 FR
Sylvère Bastien, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Mathilde Boutigny, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
David Monin, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Adil El Filiali, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Claudia Carareto, Department of Biology, UNESP - São Paulo State University, BR
Cristina Vieira, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Franck Picard, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Natacha Kremer, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Fabrice Vavre, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Marie-France Sagot, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR
Vincent Lacroix, Laboratoire de Biométrie et Biologie Evolutive - Université Lyon 1 FR

SNPs (Single Nucleotide Polymorphisms) are genetic markers whose precise identification is a prerequisite for association studies. Methods to identify them are currently well developed for model species, but rely on the availability of a (good) reference genome, and therefore cannot be applied to non-model species. They are also mostly tailored for whole genome (re-)sequencing experiments, whereas in many cases, transcriptome sequencing can be used as a cheaper alternative which already enables to identify SNPs located in transcribed regions.

In this paper, we propose a method that identifies, quantifies and annotates SNPs without any reference genome, using RNA-seq data only. Individuals can be pooled prior to sequencing, if not enough material is available for sequencing from one individual.

Using human RNA-seq data, we first compared the performance of our method with GATK, a well established method that requires a reference genome. We showed that both methods predict SNPs with similar accuracy.

We then validated experimentally the predictions of our method using RNA-seq data from two non-model species.

The method can be used for any species to annotate SNPs and predict their impact on proteins. We further enable to test for the association of the identified SNPs with a phenotype of interest.

The method is distributed as a pipeline: http://kissplice.prabi.fr/TWAS.

P24 Identification of rag genes mutations in human primary immunodeficiency diseases using exome sequencing

Reza Mahdian, Pasteur Institute of Iran
Hamzeh Rahimi, Pasteur Institute of Iran
Shirin Shahbazi, Tarbiat Modares University Iran

Background and Purpose: Primary immunodeficiencies (PIDs) are more than 250 rare diseases with similar clinical symptoms. They are caused by gene defects in proteins required for normal immune function and classified into nine categories based on immunological laboratory assay. We utilized the whole exome sequencing as it has demonstrated to be a potent and efficient tool for identifying causative genetic defects in rare mendelian disorder.

Methods: We applied whole exome sequencing to study patients with lymphohistiocytosis in a consanguineous family. The sequence analysis platform contain five steps; quality control, genome mapping and alignment, post-alignment processing, variant calling, variant annotation. The variants effects on protein function were investigated by Sift, Polyphen, Extasy, CADD, ClinVar.

Results: The data was annotated and subjected to quality control. Our results showed 49,454 variants and 19450 rarer SNPs with MAF less than 0.1. Following data filtering and gene prioritization, nine variants were candidate for structural analysis. By further processing, two RAG genes mutations were found as disease causing variants. Functional studies will describe the impact of these variants on protein expression and activation.

Conclusions: More than 150 genes have been implicated in PIDs including the genes that follow the mendelian inheritance. An efficient tool for mutation analysis in these pedigrees enhances genetic counselling and prenatal diagnosis. Our findings draw attention to the significance of whole exome sequencing as powerful and cost-effective strategy in the diagnosis of heterogeneous disorders including PIDs.

P25 De novo assembly and preliminary analysis of the spider crab

Maja brachydactyla transcriptome Rosa M. Garcia-Junco, Grupo de Investigación en Bioloxía Evolutiva (GIBE), Departamento de Bioloxía Celular e Molecular, Facultade de Ciencias and Centro de Investigaciones Científicas Avanzadas (CICA), Universidade da Coruña ES

Andrés Martínez-Lage, Grupo de Investigación en Bioloxía Evolutiva (GIBE), Departamento de Bioloxía Celular e Molecular, Facultade de Ciencias and Centro de Investigaciones Científicas Avanzadas (CICA), Universidade da Coruña ES

Ana M. González-Tizón, Grupo de Investigación en Bioloxía Evolutiva (GIBE), Departamento de Bioloxía Celular e Molecular, Facultade de Ciencias and Centro de Investigaciones Científicas Avanzadas (CICA), Universidade da Coruña ES

Lars Kaderali, Universitätsmedizin Greifswald, Institut für Bioinformatik, Greifswald DE

The spider crab Maja brachydactyla is a brachyuran decapod crustacean occurring in the north-east Atlantic. This species is subject of a heavy exploitation pressure in the coastal area comprehended between Morocco and the north of France and ecological data suggest that there might be overexploitation problems in this region. Several studies about the spider crab have been carried out from an ecological and biochemical perspective but there is very little information on the molecular genetics aspects.

High-throughput sequencing technologies have notably changed our approach to genome and transcriptome analysis. Nowadays it is possible to generate large data sets of sequences at a relatively low cost in price and time. When working with a non-model organism the available genetic information is usually scarce. RNA sequencing (RNA-seq) is a very useful tool for gene expression profiling, gene identification and mining of molecular markers. Even lacking a reference genome, it is possible to assembly a transcriptome de novo as a starting point for gene expression studies. In the last few years RNA-seq technologies have made possible the transcriptome analysis of several crustacean non-model species, widening our knowledge about different biological aspects as signaling, reproduction and development.

In the present study we generate a dataset for the transcriptome of an adult female Maja brachydactyla. We performed Illumina paired-end sequencing of gut, gill, hepatopancreas and heart. The resulting sequences were assembled to contigs and analysed. We obtained 120698 unigenes, of which 29849 were successfully annotated. It has been possible to identify several growth-related genes and a new panel of SSR candidates is being developed, providing a starting point for further investigation in this species.

This research has been supported by a grant from Ministerio de Economía y Competitividad (MINECO), reference CTM2014-53838-R.

P26 RG: Annotation of Genomic Regions with High/Low Variant Calling Concordance

Niko Popitsch, University of Oxford UK

The increasing adoption of whole-genome resequencing (WGS) in clinical and research environments demands for highly-accurate and reproducible variant calling (VC) pipelines. The

observed amount of discordant SNV and INDEL calls between different state-of-the-art VC pipelines, however, indicates non-negligible numbers of false positives and negatives that were shown to be strongly enriched among discordant calls but also in genomic regions with low sequence complexity.

Here, we describe a novel method (ReliableGenome/RG) for partitioning genomes into high and low concordance regions with respect to a set of surveyed VC pipelines. Our method combines call sets derived by multiple pipelines from arbitrary numbers of input datasets and interpolates expected concordance for genomic regions without data.

By applying RG to 219 deep human WGS datasets, we demonstrate that VC concordance depends predominantly on genomic context rather than the actual sequencing data which manifests in high recurrence of regions that can/cannot be reliably genotyped by a single method. This enables the application of pre-computed regions to other data created with comparable sequencing technology and software.

RG mainly differs from comparable previous works that were derived from single datasets (e.g., Genome-In-A Bottle reliable regions) in that it integrates data from entire cohorts, thereby capturing more data variance. It clearly outperforms previous efforts when treated as a binary classifiers for predicting VC concordance and our evaluation confirms that predicted low-concordance regions harbour predominant fractions of presumably false positives which strongly underlines RG's usefulness for variant filtering, annotation and prioritization.

RG allows focusing resource-intensive algorithms (e.g., consensus calling methods) on the smaller, discordant share of the genome (~20-30%) which might result in increased overall accuracy at reasonable costs. Our method and analysis of discordant calls may further be useful for development, benchmarking and optimization of VC algorithms and for the relative comparison of call sets between different studies/pipelines.

P27 ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions

Jie Tan, Geisel School of Medicine at Dartmouth US
John Hammond, Geisel School of Medicine at Dartmouth US
Deborah Hogan, Geisel School of Medicine at Dartmouth US
Casey Greene, Perelman School of Medicine, University of Pennsylvania US

The increasing number of genome-wide assays of gene expression available from public databases presents opportunities for computational methods that facilitate hypothesis generation and biological interpretation of these data. We present an unsupervised machine learning approach, ADAGE (analysis using denoising autoencoders of gene expression), and apply it to the publicly available gene expression data compendium for Pseudomonas

aeruginosa. In this approach, the machine-learned ADAGE model contained 50 nodes which we predicted would correspond to gene expression patterns across the gene expression compendium. While no biological knowledge was used during model construction, cooperonic genes had similar weights across nodes, and genes with similar weights across nodes were significantly more likely to share KEGG pathways. By analyzing newly generated and previously published microarray and transcriptome sequencing data, the ADAGE model identified differences between strains, modeled the cellular response to low oxygen, and predicted the involvement of biological processes based on low-level gene expression differences. ADAGE compared favorably with traditional principal component analysis and indepen dent component analysis approaches in its ability to extract validated patterns, and based on our analyses, we propose that these approaches differ in the types of patterns they preferentially identify. We provide the ADAGE model with analysis of all publicly available P. aeruginosa GeneChip experiments and open source code for use with other species and settings. Extraction of consistent patterns across large-scale collections of genomic data using methods like ADAGE provides the opportunity to identify general principles and biologically important patterns in microbial biology. This approach will be particularly useful in less-well-studied microbial species.

P28 Cancer genomic medicine at University of Tokyo

Satoru Miyano, The University of Tokyo JP

Importance of integrative systems understanding of cancer based on personal omics data is getting more important while the cost for human genome sequence cost is drastically decreasing. By jointly interpreting transcriptome data, the future of clinical sequence of cancer and germline genome sequence are being designed and implemented. At the Institute of Medical Science (The University of Tokyo), a team comprising of Human Genome Center, Advanced Clinical Research Center, Research Hospital has implemented and started personalized cancer clinical sequence as research based on multi-regional whole genome sequencing and integrative omics analysis. Supercomputer System of Human Genome Center and the system of next-generation sequencers are systematically controlled by Clarity LIMS system (a laboratory information management system) to provide secure and traceable data analysis pipelines, computational systems biology tools, and a documentation system. As a perspective view of the future low cost sequencers, we have been collaborating to create clinical applications using the quantum sequencing technology with Quantum Biosystems Inc. We also recognize the important of fostering medical informatics people who will play a central role in clinically interpreting and translating whole genome sequence data and omics data. In USA, IBM Watson Genomic Analytics trained at New York Genome Center are running at 17 institutions/centers including Memorial Sloan Kettering Cancer Center and Mount Sinai. It may figure a future of personalized cancer genomic medicine. In July 2015, IBM Watson Genomic

Analytics was introduced in our institute as research. We present a perspective on cancer genomic medicine based on these issues.

P29 Integrated Platform for Detection of Copy Number Variations

Nita Parekh, International Institute of Information Technology IN
Sriharsha Vogeti, International Institute of Information Technology IN
Shanta Pendkar, International Institute of Information Technology IN

Copy-number variations (CNVs) are a form of structural variation that leads to abnormal copies of large genomic regions in a cell. CNVs may encompass genes and lead to variation in gene expression in the cell. Thus, detection of CNVs plays an important role in understanding the molecular mechanisms leading to pathogenesis and in drug response. Depth of coverage based methods are known to accurately predict exact copy number and can even detect very large insertions in whole genome NGS data. Here we implemented three segmentation algorithms (piecewise constant function with total variation penalized least squares model (Duan et al, 2013), mean-shift (Wand et al, 2009) and circular binary segmentation (Venkatraman et al, 2007)) to fit read depth signal on depth of coverage approach, and also paired-end mapping (PEM) approach. For single samples correction for GC bias and mappability of reads is implemented in the preprocessing step.

Analysis is performed on simulated data, population data (NA12891), and tumor data (SRR1236468) for the three approaches. Features considered for comparison include CNV breakpoints, size, copy number and sequencing depth of the sample (for simulated data). We observe that shorter CNVs (~ 1000) with 1 gain/loss require a higher coverage (~30×), while longer CNVs (~ 4000) are detected even at 10× coverage. The total variation approach is able to detect smaller CNVs (~ 500) in real data while PEM approach improved the initial alignment of reads and accurate breakpoint prediction. In our analysis of tumor data, we identified number of CNVs, some of which are reported to play a role in tumorigenesis. A significant variation is observed in the number and size of CNVs detected by the different methods. Hence we use a conservative approach of consensus of CNVs detected by two or more methods for the analysis of population and tumor data.

P30 Active trancription without canonical histone marks

Sílvia Pérez-Lluch, Centre for Genomic Regulation, Barcelona ES
Enrique Blanco, Centre for Genomic Regulation, Barcelona ES
Hagen Tilgner, Centre for Genomic Regulation, Barcelona ES
Joao Curado, Centre for Genomic Regulation, Barcelona ES
Marina Ruiz-Romero, Universitat de Barcelona ES

Montserrat Corominas, Universitat de Barcelona ES
Roderic Guigó, Centre for Genomic Regulation, Barcelona ES

Transcriptional regulation depends on many factors such as chromatin structure, DNA methylation, transcription factors and histone modifications. The interplay of activating and repressing histone modifications is assumed to play a key role in the regulation of gene expression. H3K4me3, for instance, has been associated to gene activation whereas H3K27me3 has been related to gene silencing. Challenging this generally accepted model, here we show that activation of genes that are temporally and spatially regulated during metazoan development occurs in the absence of canonically activating histone modifications, such as H3K4me3 and H3K9ac. We have seen that genes void of histone modifications are actively transcribed in isolated cells and that the perturbation of the methylation of H3K4 does not affect their expression. We have finally shown that genes lacking chromatin marking are usually located in low binding regions according to ChromHMM segmentations and show a different pattern of transcription factors binding from genes expressed throughout development. Our results support a dual model of chromatin associated transcription regulation, in which chromatin marking is associated to stable, tightly controlled production of RNA, while a more flexible, unmarked chromatin state would permit rapid gene activation and deactivation during development. In these genes, transcription factors binding to chromatin would play the predominant regulatory role.

P31 SeqPurge: highly-sensitive adapter trimming for paired-end NGS data

Marc Sturm, Institute of Medical Genetics and Applied Genomics, University Hospital Tübingen DE

Christopher Schroeder, Institute of Medical Genetics and Applied Genomics, University Hospital Tübingen DE

Peter Bauer, Institute of Medical Genetics and Applied Genomics, University Hospital Tübingen DE

Trimming adapter sequences from short read data is a common preprocessing step in most DNA/RNA sequence analysis pipelines. For amplicon-based approaches, which are mostly used in clinical diagnostics, sensitive adapter trimming is of special importance. Untrimmed adapters can be located at the same genomic position and can lead to spurious variant calls. Shotgun approaches are more robust towards adapter contamination, because untrimmed adapters are randomly distributed over the target region. This reduces the probability of spurious variant calls.

When performing paired-end sequencing, the overlap between forward and reverse read can be used to identify excess adapter sequences. This is exploited by several published adapter

trimming tools. However, in our evaluations on amplicon-based paired-end data we found that these tools fail to remove all adapter sequences and that adapter contamination leads to spurious variant calls.

Here we present SeqPurge, a highly-sensitive adapter trimmer that uses a probabilistic approach to detect the overlap between forward and reverse reads of paired-end Illumina sequencing data. The overlap information is then used to remove adapter sequences, even if only one base long. Compared to other adapter trimmers specifically designed for paired-end data, we found that SeqPurge achieves a higher sensitivity. The number of remaining adapters after trimming is reduced by 40-90%, depending on the compared tool. The specificity of SeqPurge is comparable to that of other state-of-the-art tools. In addition to adapter trimming, SeqPurge can also perform quality-based trimming, trimming of no-call (N) stretches, merging of FASTQ files and FASTQ quality metrics calculation.

SeqPurge is implemented in C++ and runs both under Linux and Windows. It is available under the 'GNU General Public License version 2' as part of the ngs-bits project: https://github.com/marc-sturm/ngs-bits

---

P32 Role of G9a in different adult stem cell populations: impact on homeostasis and cancer

Aikaterini Symeonidi, Institut for Research in Biomedicine (IRB-Barcelona) ES
Alexandra Avgustinova, Institut for Research in Biomedicine (IRB-Barcelona) ES
Salvador Aznar Benitah, Institut for Research in Biomedicine (IRB-Barcelona) ES

To replenish damaged cells, adult stem cells divide to self-renew and fuel cellular differentiation. Failure to balance these events predisposes the tissue to ageing, loss of regenerative capacity, or cancer. Here we investigate the role of the histone methyltransferase G9a in stem cell function using a transgenic mouse model deficient for G9a in the epidermis and mammary epithelium. On a whole tissue level, G9a is largely dispensable for epidermal development and homeostasis, while mammary development is severely impaired. G9a deposits mono- and dimethyl marks on lysine 9 of histone 3, thereby triggering chromatin compaction and concomitant transcriptional repression. Indeed, we observed a global transcriptional activation in G9a deficient cells derived from both epidermis and mammary epithelium, with a significant overlap of the activated genes and pathways in the two tissues. Interestingly, the de-repressed targets included embryonic and fetal proteins, suggesting that G9a may be involved in shutting off the embryonic gene expression program during development. De-repressed genes clustered at specific genomic regions, consistent with the mechanism of de-repression being the loss of H3K9me2, which is often found in large organized chromatin K9-modification (LOCK) domains. H3K9me2 and G9a ChIP-sequencing revealed that, as predicted, H3K9me2 levels are severely reduced in G9a deficient cells. Further, both

H3K9me2- and G9a-enriched regions predominantly occur in intergenic regulatory regions, with only few peaks at gene TSSs.

H3K9 di- and trimethylated chromatin has been suggested to positively correlate with the mutation rate of human cancers. To experimentally demonstrate this association, we are analyzing exome sequencing data derived from chemically induced cutaneous squamous cell carcinomas of control or G9a-deficeint mice. In summary, by collating different types of data sets we are dissecting the molecular role of G9a in adult stem cell function across different tissues.

P33 MOLAS: Multi-Omics onLine Analysis System for Gene Expression Profiling

Shu-Hwa Chen, Institute of Information Science, Academia Sinica TW
Sheng-Yao Su, Institute of Information Science, Academia Sinica TW
Chung-Yen Lin, Institute of Information Science, Academia Sinica TW
I-Hsuan Lu, Institute of Information Science, Academia Sinica TW

Next generation sequencing technologies bring the gene profiling study into big-data science era. However, the increasing amount of data made itself a problem for viewing data and deciphering the biological implication from it.

Here we present MOLAS, Multi-Omics onLine Analysis System, a robust web application which can take gene expression data (FPKM/RPKM) from different libraries as inputs, map these expressed genes with annotations for further analyses and reveal biological meaning of the complex data with build-in analysis tools in the intuitive interface.

Via an intuitive data loading process in MOLAS web portal, the submitted project will be created and connected to the build-in annotations and data analysis pipeline, then turned the whole set into a website in few minutes. In each project, the infrastructure of MOLAS can provide data accessing functions including full-text search, KEGG pathways and module hierarchy view, pairwise libraries comparison with tailor-made choices, clustering by user-defined scheme, enrichment analysis in KEGG pathway and GO terms of differentially expressed genes identified by pairwise comparison or by clustering analysis. The novel functions of this system include gene list import, management of gene lists generated from above functions, illustrate Venn -diagram with up to four gene lists, select new lists from intersections of Venn diagram and generate heatmaps for specific gene list. In such approach, user can have gene lists with function enrichment in different dimensions to their own core problem. Currently, MOLAS accepts gene expression data table in FPKM value, derived from Cufflinks or other tools that map reads to human reference (GRCh38/ GRCh37, hg19) or mouse reference (mm10) in Ensembl transcript view.

Website URL (Web portal of Data submission): http://molas.iis.sinica.edu.tw/

P34 Core Vertebrate Genes (CVG): phylogeny-aware completeness assessment of genome and transcriptome assemblies

Yuichiro Hara, RIKEN CLST JP
Shigehiro Kuraku, RIKEN CLST JP

Coverage of protein-coding genes has been employed as an essential metric for assessing 'completeness' of de novo genome and transcriptome assemblies. CEGMA, an existing ortholog identification pipeline, has been used for this measurement referring to the Core Eukaryotic Genes (CEG) consisting of 248 orthologs. For this purpose, the CEG was designed so that the orthologs were widely conserved in eukaryotic taxa with no or minimal gene duplicates. However, our examination revealed that a CEGMA execution potentially misidentifies paralogs of the CEG as orthologs, which leads to overestimation of completeness and then resulting in an unreliable assessment. For more accurate assessment, we introduce Core Vertebrate Genes (CVG), a new reference gene set of 233 ortholog groups. All the ortholog groups are aimed to contain no paralogs among 29 vertebrate genomes that cover all the extant major vertebrate taxa including cyclostomes and chondrichthyans. Using the CVG, we assessed completeness of vertebrate genome assemblies and embryonic transcriptome assemblies of Madagascar ground gecko (Paroedura picta). The result demonstrated that the evaluation referring to the CVG achieved higher accuracy and resolution than that with the CEG. The performance of the CVG was also confirmed with BUSCO, a pipeline that is supposed to take over the function of CEGMA. For wide uses in assessing emerging sequence resources, the CVG data set is available online (http://www2.clst.riken.jp/phylo/reptiliomix.html; see Hara et al., 2015. BMC Genomics 16:977). Our approach with a custom choice of reference gene sets for more accurate complete assessment should also be useful in other taxa.

P35 Benchmarking 16S rRNA gene sequencing and bioinformatics tools for identification microbial abundances

Luca Cozzuto, Centre for Genomic Regulation, Barcelona ES
Carlos Company, Centre for Genomic Regulation, Barcelona ES
Nuria Andreu, Centre for Genomic Regulation, Barcelona ES
Jochen Hecht, Centre for Genomic Regulation, Barcelona ES
Julia Ponomarenko, Centre for Genomic Regulation, Barcelona ES

Genomic DNA from two microbial Mock communities, provided by the BEI resources of the Human Microbiome Project, was sequenced using both shotgun and 16S rRNA (amplifying the V3-V4 regions) sequencing on the HiSeq and MiSeq instruments. For the 16S rRNA and whole

DNA, nine and three independent sequencing runs were carried out, respectively. We set up and tested three bioinformatics pipelines for the 16S rRNA analysis: QIIME, mothur, and the in-house built pipeline based on the skewer, pear and ribopicker algorithms. The SILVA database was used for aligning against and identifying bacteria at the taxon levels of genera and species. The distributions of relative abundances of bacteria genera and species were estimated using three methods and were compared with the rRNA operon counts, provided by the BEI resources and obtained from the shotgun sequencing. In this work, we will show the results of this benchmarking experiment and discuss applicability of different evaluation criteria, including both parametric and non-parametric test statistics.

P36 Understanding cell fate decisions through integrative analyses of multi-dimensional genome-wide sequencing data

Ms Vijayabaskar, University of Leeds UK
Debbie Goode, Cambridge Institute for Medical Research and Wellcome Trust and MRC Cambridge Stem Cell Institute UK
Nadine Obier, University of Birmingham UK
Michael Lie-A-Ling, University of Manchester UK
Andrew Lilly, University of Manchester UK
Rebecca Hannah, Cambridge Institute for Medical Research and Wellcome Trust and MRC Cambridge Stem Cell Institute UK
Monika Lichtinger, University of Birmingham UK
Kiran Batta, CRUK Manchester Institute, University of Manchester UK
Magdalena Florowska, CRUK Manchester Institute, University of Manchester UK
Rahima Patel, CRUK Manchester Institute, University of Manchester UK
Mairi Challinor, CRUK Manchester Institute, University of Manchester UK
Kirstie Wallace, CRUK Manchester Institute, University of Manchester UK
Jane Gilmour, University of Birmingham UK
Salam Assi, University of Birmingham UK
Pierre Cauchy, University of Birmingham UK
Maarten Hoogenkamp, University of Birmingham UK
David Westhead, University of Leeds UK
Georges Lacaud, CRUK Manchester Institute, University of Manchester UK
Valarie Kouskoff, CRUK Manchester Institute, University of Manchester UK
Berthold Gottgens,Cambridge Institute for Medical Research and Wellcome Trust and MRC Cambridge Stem Cell Institute UK
Constanze Bonifer, University of Birmingham UK

Cellular differentiation is a tightly controlled process that is essential for the development of a single celled embryo into a multicellular organism. This process is regulated primarily through changes in the dynamics of gene differentiation and is orchestrated by the interplay of various

biological mechanisms such as chromatin accessibility, modifications and most importantly, binding of transcription factors. In this study we have generated, integrated and analysed a large multi-dimensional dataset for the in-vitro differentiation of mouse embryonic stem cells to macrophages. At the first step we successfully associate gene expression changes with important cell fate decisions that determine lineage specification. By interrogating DNaseI accessibility, active histone marks such as H3K27ac, H3K9ac and H3K4me3, repressive H3K27me3 histone mark and 16 transcription factors, in a unified manner, we show that we can map this four dimensional data to gene expression events, vis-a-vis cellular differentiation. Here, we also show how we systematically build up the complexity of the analysis and derive a simple yet highly informative core gene regulatory network for the transcriptions factors important in our developmental pathway. Follow-up experiments driven by the new hypotheses derived from our analyses shows that this dataset can be powerful in implicating new biological gene networks and pathways in definitive hematopoiesis. We have also provided these analyses as a user friendly website for the access of general scientific community.

P37 Creating a Multi Genome Graph by Minimizing Shannon Information

Seyedeh Leily Rabbani, Max Planck Institute for Developmental Biology DE
Jonas Müller, Max Planck Institute for Developmental Biology DE
Detlef Weigel, Max Planck Institute for Developmental Biology DE

Many large-scale genome projects now aim to analyze thousands of genomes. Clearly, comparing these only against a single reference genome sequence is no longer adequate. To overcome the limits imposed by using a single-genome reference, we developed methods that allow for simultaneous comparison against multiple high-quality reference genomes. To reduce the size and complexity of the resulting graph, highly similar orthologous and paralogous regions are collapsed while more substantial differences are retained. To evaluate the performance of our model, we created a genome compression tool that can be also used for global genome comparison.

The major design problem in creating the graph is defining criteria for what constitutes a single node of related sequences. Minimizing Shannon information provides a way to solve this problem non-parametrically. Genome specific Markov chain models with variable orders are trained on genomic sequences and pairwise alignments. This allows for computing the information cost of creating a DNA sequence de novo as well as of creating it by modifying a template sequence. Using a model where DNA sequences derive from their corresponding cluster center sequence, we create a clustering which minimizes information cost. This yields a graph through which each input contig corresponds to a path and which collapses repetitive sequences.

As the graph structure is the result of algorithms which minimize Shannon information, it can be used for lossless DNA compression. Superior compression performance shows that our model fits sequence properties better than existing DNA compression programs. We create a simplified multi-genome reference which can be used for read mapping and genotyping as well as comparative genomics. As each piece of input sequence is contained in one sequence cluster, the graph structure can also be used for classification of repetitive elements. In addition, it allows for global genome comparison by computing the common information content.

P38 Untangling the gene networks for motor neuron degeneration: from disease model transcriptomes to cellular systems

Hugo A F Santos, BioISI - Biosystems & Integrative Sciences Institute - Faculty of Sciences, University of Lisboa, Gene Expression and Regulation Unit, Lisbon PT
Andreia Amara, BioISI - Biosystems & Integrative Sciences Institute, Gene Expression and Regulation Unit, Faculty of Sciences, University of Lisboa / Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa PT
Takakazu Yokokura OIST - Okinawa Institute of Science and Technology Graduate University, Formation and Regulation of Neuronal Connectivity Research Unit, Okinawa, JP
David Van Vactor, Harvard Medical School, Department of Cell Biology, Boston, US/ Okinawa Institute of Science and Technology Graduate University, Formation and Regulation of Neuronal Connectivity Research Unit, Okinawa JP
Margarida Gama-Carvalho, BioISI - Biosystems & Integrative Sciences Institute - Faculty of Sciences, University of Lisboa, Gene Expression and Regulation Unit, Lisbon PT

Spinal Muscular Atrophy (SMA), a lethal inherited neurodegenerative disorder, is characterized by low levels of the Survival of Motor Neuron (SMN) protein, which is essential for the assembly of spliceosomal small nuclear ribonucleoproteins (snRNPs). Strikingly, low levels of this ubiquitous protein mainly affect motor neurons (MNs), disrupting neuromuscular junctions (NMJs) and leading to MNs degeneration. Despite robust knowledge of SMA's genetics, the exact molecular mechanisms underlying the disease's phenotype remain largely elusive, preventing the development of rational therapeutics.

We performed RNA-Seq profiling of the central nervous system (CNS) transcriptome of a Drosophila melanogaster SMA disease model, in parallel with a similar analysis focused on human motor neuron cultures derived from patient induced pluripotent stem cells (iPSCs).

Upon SMN down-regulation we observe changes in exon usage in a subset of genes crucial for neuronal development, viability and NMJ function in both disease models. This strongly suggests that SMN-dependent changes in the splicing machinery do not have widespread effects, affecting specific genes possibly due to the existence of certain features in their

sequence or structure, hence corroborating the SMN-dependent splicing modulation as a key hallmark of SMA pathomechanisms. Interestingly a large proportion of identified genes with altered splicing are known genetic modifiers of the SMN loss-of-function phenotype in SMA fly models, thereby supporting the search for innovative therapeutic approaches to SMA.

In addition, we have found significant subsets of genes sharing coordinated responses with SMN down-regulation - those positively correlated were intimately related to mRNA processing terms while genes that were negatively correlated showed enrichment in motor neuron key processes (axon guidance/ neuron differentiation) and inclusively amyotrophic lateral sclerosis, a neurodegenerative disease previously linked to SMA. An assessment of the networks established by these genes coupled with pathway-analysis identified essential pathways (e.g. SLIT/Robo) as being promising molecular targets for future therapeutics.

P39 FASTAIR, Functional Analysis of Sequenced Transcripts At the Isoform Resolution

Lorena de La Fuente, Centro de Investigación Príncipe Felipe ES
Manuel Tardaguila, University of Florida US
Cristina Martí, Centro de Investigación Príncipe Felipe ES
Hecor Del Risco, University of Florida US
Maravillas Mellado, Centro de Investigación Príncipe Felipe ES
Marissa Macchietto, University of California Irvine US
Ali Mortazavi, University of California Irvine US
Susana Rodriguez, Centro de Investigación Príncipe Felipe ES
Victoria Moreno, Centro de Investigación Príncipe Felipe ES
Ana Conesa, Centro de Investigación Príncipe Felipe ES

Transcriptomes of higher eukaryotes are characterized by the presence of multiple isoforms coded by the same gene. Although a dense structural catalog of isoforms and splice junctions has been created for many organisms, the study of the functional implications of alternative isoform expression (AIE) has not been yet addressed at a whole-genome level. Therefore, we have developed a new methodology called FASTAIR which allows the study of the functional consequences of AIE at the genome-wide level. We have applied this approach in a mouse cell differentiation system from Neural Stem Cells to Oligodendrocytes. FASTAIR methodology combines both PacBio (long reads) and Illumina (short reads) sequencing. PacBio sequencing resolves whole transcripts up to 10kb and is ideal to elaborate precise transcriptomes and perform isoform discovery, whereas Illumina allows for accurate quantification of expression due to its high depth sequencing. In the first part of the approach, FASTAIR addresses the curation, classification and functional annotation of the transcriptome obtained from PacBio reads. Thus, FASTAIR allows to reduce the sequence error rate of PacBio isoforms, annotate the transcriptome according to RefSeq and point out novel isoforms and splice events. The functional annotation of each isoform involve the prediction of ORFs and a rich diversity of

functional layers, which includes, among others, miRNA binding sites, protein domains, post-translational modifications, UTR motifs and binding elements. Finally, as FASTAIR aim is to figure out the relevance of AIE in the functionality of the system, it applies different statistical methods which combine both expression data and functional annotation. Preliminary FASTAIR analysis of our neural differentiation model shows a 30% of novel isoforms uncovered by PacBio, a high functional diversity between isoforms coded by the same gene, and several processes specifically regulated by the alternative expression of isoforms.

P40 ChroGPS, visualization, functional analysis and comparison of epigenomes

Oscar Reina Garcia, Institute for Research in Biomedicine, IRB Barcelona ES
Fernando Azorin Marin, Institute for Research in Biomedicine, IRB Barcelona ES. Institute of Molecular Biology of Barcelona, IBMB, CSIC Barcelona ES

In recent years consortiums such as modENCODE, ENCODE, Roadmap Epigenomics or the Blueprint project have generated and published an unprecedented amount of epigenomics data comprising several organisms, tissues, cell lines and diseases. However, development of tools to perform intuitive visualization, functional analysis and comparison of epigenomes between different biological backgrounds or conditions remains a challenge.

ChroGPS is a Bioconductor package that addresses this question using multidimensional scaling techniques to represent similarity between epigenetic factors, genomic features or specific regions on the basis of their epigenetic state, in 2D/3D reference maps, integrating gene expression and other functional data. Emphasis is placed on interpretability, computational feasibility and statistical considerations to guarantee reliable representation, integration and comparison of data from multiple sources (studies, technologies, genetic backgrounds, etc.).

To illustrate this functionality we focus on data from the modENCODE project on the genomic distribution of a large collection of epigenetic factors in Drosophila melanogaster, as well as ENCODE and Roadmap Epigenomics data in human embryonic stem cells (hESCs) and cancer cell lines.

Our results show that the maps allow straightforward visualization of relationships between factors and elements, capturing relevant information about their functional properties, as well as intuitive comparison of epigenomic information from different cellular backgrounds or conditions, that helps to interpret epigenetic information in a functional context and derive testable hypotheses.

P41 A computational approach for functional classification of the epigenome

Francesco Gandolfi, Universita Sapienza di Roma IT
Anna Tramontano, Universita Sapienza di Roma IT

In the last decade, advanced functional genomics approaches and deep sequencing have allowed large-scale mapping of histone modifications and other epigenetic marks, highlighting functional relationships between chromatin organization and the genome function. However, it has been shown that many epigenetic modifications do not act as isolated signals along the DNA but co-occur in a range of combinatorial patterns, which demarcate the presence of distinct functional elements in the genome.

With the expanding amount of chromatin data now available in public domains, the need for computational methods able to integrate different types of epigenetic signals and identify biologically meaningful combinations of chromatin marks is emerging.

Here, we propose a novel approach to explore functional interactions between different epigenetic modifications and extract combinatorial patterns that can be used to annotate the chromatin in a fixed number of functional classes. Our method is based on Non-negative Matrix Factorization (NMF), an unsupervised learning technique originally employed to decompose high-dimensional data in a reduced number of meaningful patterns.

We applied the NMF algorithm on a collection of genomic datasets representing 13 different epigenetic marks, consisting of ChIP-seq assays for multiple histone modifications, Pol-II binding and chromatin accessibility assays from human H1-hESCs.

Interestingly, this approach allowed us to identify a number of chromatin patterns that contain functional information and are biologically interpretable. Moreover, we observed that single epigenetic patterns are characterized by distinct genomic profiles, which correlate well with specific features of the genome. These preliminary results highlight the utility of NMF in studying functional relationships between different epigenetic modifications and may provide new biological insights in the interpretation of the chromatin dynamics.

P42 Integrating multi-omic NGS data in regression models to understand gene expression regulation

Sonia Tarazona, Centro de Investigacion Principe Felipe, Madrid ES
Mónica Clemente-Císcar, CIPF, Madrid ES
Ana Conesa, Genomics of Gene Expression Lab ES

The continuous evolution of NGS technologies has made it easier to obtain genome-wide multi-omic data from the same biological system that can be integrated to study the regulatory mechanisms driving gene expression. However, it is essential to have efficient statistical integration models to extract the most significant regulators of gene expression from given experimental data.

We propose a novel algorithm to compute generalized linear regression models (GLM) for explaining gene expression as a function of its regulators and the experimental conditions. The potential regulators may be miRNAs, transcription factors, or genomic regions with chromatin accesibility, proteing binging or methylation, etc.

The algorithm, that will be part of an R package, provides different functionalities. Apart from the flexibility to define the GLM models (linear or non-linear, interactions, family of probability distribution, stepwise variable selection procedure, etc.), it allows for filtering regulators with low variability or aggregating them in a multi-collinearity situation. Moreover, it returns a global summary for the results, and also a function for plotting either all the significant regulators of a given gene or all the genes significantly regulated by a given regulator, as well as partial or global regulatory networks.

We show the ability of the algorithm to generate useful biological knowledge by applying it to the huge collection of NGS data generated within the European STATegra project: time course data for control and treatment conditions in mouse from RNA-seq, miRNA-seq, DNase-seq and RRBS-seq technologies.

As far as we know, this is the first bioinformatic tool that deals with the integration of different omic data types from time-course experiments in order to identify the genomic elements that regulate gene expression (also valid for transcripts, proteins, etc.). Identifying the regulatory mechanisms of gene expression will help us to understand diseases, define therapeutic targets, improve treatments, etc.

P43 An integrated gene annotation pipeline for de-novo sequenced organisms

Francisco Câmara Ferreira, Center for Genomic Regulation, Barcelona ES
Jèssica Gómez Garrido, Centre Nacional d'Anàlisi Genòmica, Barcelona ES
Anna Vlasova, Center for Genomic Regulation, Barcelona ES
Didac Santesmasses, Center for Genomic Regulation, Barcelona ES
Roderic Guigó, Center for Genomic Regulation, Barcelona ES
Tyler Alioto, Centre Nacional d'Anàlisi Genòmica, Barcelona ES

Accurate gene model prediction and good quality genome assemblies are both crucial for addressing many interesting biological questions. For protein-coding gene (PCG) annotation, we have been employing a semi-automatic pipeline, based on the well-established programs PASA and EvidenceModeler, that combines protein, transcript and gene prediction sources of evidence in order to generate protein-coding gene annotations. The full pipeline includes estimation of completeness of the assembly; generation of species-specific repeat libraries; utilization of a number of gene prediction tools (including their training and evaluation); mapping of available transcripts and proteins; prediction of particular classes of proteins; and assignment of possible functions and putative names to the gene models. A number of tools

used in the pipeline have been developed by the Guigó and Alioto groups. These programs include the ab-initio gene prediction program geneid, and its training tool (geneidtrainer), the homology evidence-based SGP2, GEMTOOLS (a pipeline to align RNAseq reads to the genome using the GEM mapper) and the program Selenoprofiles. Using this pipeline we have annotated de-novo sequenced eukaryotic organisms from many different taxa - vertebrates, plants, arthropods and fungi. We show how the quality of the PCG annotation is dependent on the state of the assembly and range of the available transcript/protein data - the more comprehensive the data sets the better the quality of the assembly and of the gene models as well as the resolution of their function. The degree of completeness/quality of the annotation can be up to 96% of complete gene models with ~70% of these having annotated UTRs and ~25% alternative splicing forms.

P44 Protein coding genes have a single dominant isoform

Jose Manuel Rodriguez, Spanish National Bioinformatics Institute (INB-CNIO), Madrid ES
Alfonso Valencia, Spanish National Bioinformatics Institute (INB-CNIO), Madrid ES
Michael Tress, Spanish National Cancer Research Centre, Madrid ES

The question of whether or not genes have a dominant variant has become increasingly important as the numbers of annotated alternative transcripts has grown. While the longest isoform is chosen as the reference isoform in practically all studies and databases, this has no biological basis. Large-scale transcriptomics studies have suggested that genes may have dominant transcripts, although there is disagreement as to whether this dominant transcript is tissue-specific or across all cell lines.

Recent results from large-scale proteomics studies have demonstrated that the vast majority of protein-coding genes have a single main protein isoform. Main protein isoforms from the proteomics experiments were in agreement with two other means of determining main variants, the unique CCDS isoforms and the APPRIS principal isoforms, in over 97% of comparable genes. The agreement between three entirely orthogonal sources significantly reinforces the likelihood that the main experimental proteomics isoform – and by extension the unique CCDS and APPRIS principal isoforms - is the dominant cellular isoform.

The importance of APPRIS principal isoforms has been underscored by a recent analysis of data from the 1,000 genomes project, results that we have since confirmed in our laboratory. The analyses demonstrated that exons from APPRIS principal isoforms had proportionally many fewer high impact variants than alternative exons, indicating that exons from principal isoforms are under stronger purifying selection than alternative exons.

Finally we have been able to predict main transcripts from large-scale RNAseq experiments using two aligners, STAR and hisat, and we find that the agreement between APPRIS principal isoforms and the transcriptomics data is still high, rising to 93.5% where both aligners agree.

The APPRIS database of principal isoforms (http://appris.bioinfo.cnio.es), which was developed as part of the GENCODE human genome consortium, is now available for eight organisms and can be extended via web services to many others.

P45 Altered oncomodules underlie chromatin regulatory factors driver mutations

Joan Frigola, Research Program on Biomedical Informatics, IMIM Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Barcelona ES
Ane Iturbide, Programa de Recerca en Càncer, Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona ES
Nuria Lopez-Bigas, Research Program on Biomedical Informatics, IMIM Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Barcelona ES
Sandra Peiro, Programa de Recerca en Càncer, Institut Hospital del Mar d'Investigacions Mèdiques, Barcelona ES
Abel Gonzalez-Perez, Research Program on Biomedical Informatics, IMIM Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Barcelona ES

Chromatin remodeling is crucial for gene expression regulation. It is carried out by a set of proteins generically referred to as chromatin regulatory factors (CRFs), known to be involved in tumorigenesis. Nevertheless, the molecular mechanisms through which driver alterations of CRFs cause tumorigenesis remain unknown. We identified gene modules miss-regulated upon driver mutation of CRFs. To this end, we developed a CRFs Oncomodules Discovery approach, which comprises a scoring system that mines several sources of cancer genomics and perturbomics data to prioritize potential oncogenic modules. The approach highlights possible therapeutic strategies to indirectly target driver mutations in CRFs. We were able to uncover the potential mechanisms of tumorigenesis unleashed by driver mutations of 5 driver CRFs in 3 TCGA tumor cohorts. In particular, we clearly pinpointed the involvement of the mTOR pathway oncogenic module in relation with loss-of-function mutations of MLL2 in head and neck squamous cell carcinomas. Furthermore, we proposed, and experimentally demonstrated the suitability of indirectly targeting these mutations through the uncovered oncogenic modules. The potential oncogenic modules detected by our approach may guide experiments proposing ways to indirectly target driver mutations of CRFs. To this end we have made all the details on detected CRFs oncomodules available.

P46 Improving the annotation of disease-related variations: a large-scale study at the chromosome level

Giulia Babbi, University of Bologna IT
Pier Luigi Martelli, University of Bologna IT

Giuseppe Profiti, University of Bologna IT
Rita Casadio Italy, University of Bologna IT

Modern genome investigation techniques like NGS and GWAS are becoming rapid and inexpensive, generating a large amount of data allowing variation calling. However, variations are informative when they are annotated for their different features[1]. For any personal genome, there will be numerous variants initially classified as variants of unknown significance because of uncertainty regarding how they will impact the physiology of the patient. When variations are in protein coding regions, they may affect protein structure, function and/or interactions with specific proteins and RNA/DNA molecules.

As a result, molecular mechanisms explaining the relationship among genotype and phenotype are under debate with a shift of interest towards precision medicine (http://www.nih.gov/precisionmedicine/).

Our aim was to analyse the association between protein variants and human diseases. Starting from ClinVar (www.ncbi.nlm.nih.gov/clinvar/), we selected a dataset of 89,991 variations in protein coding regions. Using the Variant Effect Predictor (www.ensembl.org/info/docs/tools/vep) and the mapping tables of UniProt, we mapped 98% of the variations into 11,121 SwissProt identifiers and we endowed them with cross-references to PDB, PFAM, KEGG, STRING, with GO terms and OMIM identifiers, improving their automatic annotation.

The curated dataset was used to build networks connecting chromosomes through genes associated to the same disease; this analysis highlights that polygenic maladies point to genes clustering in the same chromosome.

Investigating subnetworks in details, the Y chromosome does not show significant links to other chromosomes, while the mitochondrial DNA network mainly presents self-edges (corresponding to 32 OMIM IDs) but also connections to human chromosomes (10 shared OMIM IDs). These associations reflect the protein-protein interactions in protein complexes involved in the mitochondrial respiration process.

In a personalized medicine perspective, this is a new approach for studying

the basis of the molecular relations between pathogenic variations and their physiological consequences.

References

1. Vihinen M.: No more hidden solutions in bioinformatics. *Nature*,521,261(2015)

P47 Mikado: leveraging multiple transcriptome assembly methods for improved gene structure annotation

Luca Venturini, The Genome Analysis Centre UK
Shabhonam Caim, The Genome Analysis Centre UK
Gemy Kaithakottil, The Genome Analysis Centre UK
Daniel Mapleson, The Genome Analysis Centre UK
David Swarbreck, The Genome Analysis Centre UK

The reconstruction of transcript sequences from RNA-Seq reads is a powerful approach to study alternative splicing and improve gene structure annotation. However, while numerous methods have been proposed to infer the original transcript structure from sequencing data, all of them exhibit idiosyncratic strengths and weaknesses with substantial variation in accuracy across methods. Pooling together the assemblies of multiple methods provides a potential strategy to improve sensitivity, but creates the problem of determining which of the alternative assemblies best represents the correct structure at each locus. To solve this conundrum, we have developed a novel algorithm that leverages transcript assemblies generated by multiple methods to define expressed loci, select a representative transcript and provide a more accurate and comprehensive set of gene models than existing approaches. We tested our method on human, D. melanogaster and A. thaliana, recovering several thousand gene structures that were either missed entirely, incorrectly fused or with erroneous or incomplete gene models in the best of the assembly methods tested, while also removing the majority of incorrect/undesirable gene structures. For initial testing we evaluated our approach on species with well annotated manually curated genes, the method however can be configured for different species with only minor modifications to the scoring scheme. The algorithm is implemented in a novel software tool, Mikado, currently under active development.

P48 Identification and profiling of the non-coding transcriptome expressed in the porcine gluteus medius muscle

Tainã Figueiredo Cardoso, Center for Research in Agricultural Genomics (CSIC-IRTA-UAB-UB) ES
Angela Cánovas, Center for Research in Agricultural Genomics (CSIC-IRTA-UAB-UB) ES
Marcel Amills, Center for Research in Agricultural Genomics (CSIC-IRTA-UAB-UB) ES
Oriol Canela, IRTA, Genètica i Millora Animal, Barcelona ES
Rayner González-Prendes, Center for Research in Agricultural Genomics (CSIC-IRTA-UAB-UB) ES
Raquel Quintanilla, IRTA, Genètica i Millora Animal ES

Non-coding RNA (ncRNA) are fundamental regulators of gene expression, but so far they are poorly characterized and annotated in pigs. In the current study, we have examined the total ncRNA transcriptome of the porcine gluteus medius (GM) muscle of 52 individuals by using a RNA-seq approach. Sequencing data were produced in a HiSeq2000 platform and the resulting reads were mapped, assembled and annotated according to the latest pig reference genome (Sscrofa10.2). We performed the bioinformatic analyses of RNA-seq data with the CLCBio

software and BLASTN 2.3.0+. In this way, we identified 1,558 ncRNA transcripts with sizes between 35 and 9032 bp. A total of 1,257 transcripts were classified as small ncRNA, while 301 transcripts corresponded to long ncRNA. There was a remarkable variability in ncRNAs (we detected a total of 6,663 putative SNPs), a feature that could have functional consequences since variation at ncRNAs has been related with the progression of cancer and neurodevelopmental diseases. Moreover, we analysed the conservation of the 100 most expressed ncRNAs, and we found that 59 of them present more than 75% sequence homology with those of other mammalian species (p.e. Homo sapiens, Macaca sp, Ovis sp). These results suggest that there is a medium level of conservation of the muscle ncRNA transcriptome amongst species that are phylogenetically related.

P49 mint: A pipeline for analysis, integration, classification, and annotation of genome-wide DNA methylation and hydroxymethylation data

Raymond G. Cavalcante, University of Michigan US
Maureen A. Sartor, University of Michigan US

The two most common forms of DNA methylation in metazoan genomes are 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC). Recent studies demonstrate different biological roles for the two marks. However, the most commonly used sequencing experiments for detecting DNA methylation (e.g., bisulfite sequencing and reduced representation bisulfite sequencing) cannot distinguish between the two marks because both 5mC and 5hmC are resistant to bisulfite conversion. Discriminating between 5mC and 5hmC marks requires pairing bisulfite-conversion methods measuring both 5mC and 5hmC, with another sequencing experiment measuring only 5mC or 5hmC (e.g., MeDIP-seq, hMeDIP-seq, oxBS-seq, TAB-seq).

To facilitate interpretation of how 5mC and 5hmC cooperate to produce functional effects, we have developed a methylation integration (mint) analysis pipeline that transforms raw sequencing reads into 5mC and/or 5hmC CpG- or region-level classifications and genomic annotation summaries using existing and newly developed software. Supported experimental setups include pure pulldown experiments (e.g. MeDIP-seq for 5mC; hMeDIP-seq for 5hmC) or hybrid experiments (e.g. RRBS for 5mC + 5hmC; hMeDIP-seq for 5hmC) either comparing two groups for differential methylation or performing a sample-wise analysis.

The mint pipeline constructs a customized UCSC Genome Browser track hub that collates the data tracks created by the pipeline, enabling easy and fast visual exploration. To facilitate interpretation, we developed a fast and flexible R package, annotatr, to annotate methylation and classification tracks to genomic annotations. Data tracks can be annotated to custom or pre-defined genomic annotations, including detailed gene models, CpG island features, and enhancers for human and mouse genomes. Our approach maps one-region to many-annotations, providing a richer understanding of the genomic context. Finally, annotatr enables

easy visualization of categorical or numerical data associated with the tracks with respect to the genomic annotations. Our mint and annotatr software are available on GitHub at https://github.com/sartorlab/mint/ and https://github.com/rcavalcante/annotatr.

P50 RiboDB : A dedicated database of prokaryotic ribosomal proteins

Frédéric Jauffrit, UMR 5558, LBBE; Technology Research Department, bioMérieux FR
Simon Penel, UMR 5558, LBBE FR
Jean-Pierre Flandrois, UMR 5558, LBBE FR
Carine Rey, UMR 5558, LBBE FR
Stephane Delmotte, UMR 5558, LBBE FR
Manolo Gouy, UMR 5558, LBBE FR
Jean-Philippe Charrier, Technology Research Department, bioMérieux FR
Céline Brochier-Armanet, UMR 5558, LBBE FR

Since the end of the 70's, phylogenies of prokaryotes have been mainly relying on the analysis of the RNA component of the small subunit of the ribosome or a small set of housekeeping genes. The resulting phylogenies have provided interesting but partial information on the evolutionary history of these organisms because the corresponding genes do not contain enough phylogenetic signal to resolve all nodes of the Bacteria and Archaea domains. Thus, many relationships, and especially the most ancient and the most recent ones, remained elusive.

The recent burst of complete genome sequencing projects have made a lot of protein markers available as an alternative to SSU rRNA. Among protein markers, ribosomal proteins have been shown to be a valuable alternative of 16S. In addition, mass spectrometry (MALDI-TOF) studies showed that ribosomal proteins can be used to discriminate bacterial species. It is worth noting that the phylogenetic signal contained in r-proteins is a good proxy of the phylogenetic signal contained in larger sets of conserved core genes, while allowing applying ML and Bayesian approaches in acceptable computational time. Despite these advantages, ribosomal proteins remain difficult to use in routine in the absence of a dedicated resource. Furthermore, they are often misannotated in public databases.

Here we present RiboDB, a database of prokaryotic ribosomal proteins. RiboDB is built from the automated reannotation of publicly available genomes using a specialized annotation engine. The engine uses a combination of sequence similarity-based (using BLAST) and profile-based (using HMMer) approaches to ensure a high level of sensitivity without compromising specificity. The RiboDB database is available at http://ribodb.univ-lyon1.fr.

P51 Transposable element annotation in eukaryote genomes using REPET: pipelines and use cases

Joelle Amselem, INRA FR
Veronique Jamilloux, INRA FR
Nathalie Choisne, INRA FR
Isabelle Luyten, INRA FR
Florian Maumus, INRA FR
Olivier Inizan, INRA FR
Francoise Alfama-Depauw, INRA FR
Tina Alaeitabar, INRA FR
Nicolas Francillonne, INRA FR
Nicolas Lapalu, INRA FR
Mikael Loaec, INRA FR
Claire Viseux-Guerche, INRA FR
Hadi Quesneville France, INRA FR

With the development of new generation sequencing technologies, a large number of genomes were sequenced, producing very large amount of data. This increasing amount of sequences has to be analyzed, stored and searched. To face this challenge, the URGI platform (http://urgi.versailles.inra.fr) provides tools to annotate sequenced eukaryote genomes, such as structural and functional gene and repeat annotation pipelines as well as databases accessible through user-friendly interfaces to browse and query the data.

In the frame of several international plant and crop parasites (or symbiotic) whole genome sequencing projects, we were involved in genome annotation and more specifically annotation of Transposable Elements (TEs). TEs abundance is usually correlated with genome size and organization. TEs shape genomes and are a source of genetic variations and evolution. The first step of a genome annotation should be a good characterisation of TE content regions before any other genome annotation such as gene prediction. URGI develops efficient strategy relying on the REPET package (http://urgi.versailles.inra.fr/Tools/REPET) to efficiently detect, classify and annotate TEs including nested and degenerated copies.

We will present here the different pipelines included in the REPET package used to successfully annotate about 50 genomes such as wheat, grape, brassicaceae, trees, fungi and insects. We will also present results of analyses based on TEs annotation: (i) a strategy based on REPET pipeline to analyse the ancestral structure of 7 brassicaceae genomes that reveals that an important fraction of these genomes was constituted of TE relics, (ii) a whole genome comparative analysis of TE in fungi provided new insights on silencing mechanisms in fungal genomes, (iii) the role of TE in the evolution of powdery mildew avirulence-effector genes, and (iv) how they were associated across the genome of Microbotryum with gene clusters of small secreted proteins, which may mediate host interactions.

P52 Robustness of RNA-seq Gene Expression and implications in therapeutic targeting

Alexey Stupnikov, Queen's University Belfast UK
Shailesh Tripathi, Queen's University Belfast UK
Manuel Salto-Tellez, Queen's University Belfast UK
Frank Emmert-Streib, Tampere University of Technology FI
Darragh McArt, Queen's University Belfast UK

RNA-seq is an NGS-based technology that is both sensitive and dynamic in valorising the exploration of the transcriptional landscape. During recent years RNA-seq has been widely adopted, amidst many potential applications, for Differential Gene Expression (DGE) analysis. In turn, community led efforts have elucidated a number of statistical models to support and evaluate RNA-seq data for DGE.

Robustness, i.e. the characteristics of an analysis outcome caused by data shifts or perturbations is one of the key characteristics of any computational method. By specifying the data alterations type, different types of robustness can be defined.

Introducing robustness as a measure of a methods performance allows for the comparison and ranking of existing models, thus, potentially bringing more standardisation to the RNA-seq field.

We have compared the performance of several popular methods for DGE on RNA-seq data from a cancer specific context, and explored their robustness to perturbation by simulated altering within of RNA-seq experiment parameters.

In addition, we have tested the impact of the perturbation factor on the performance of candidate therapeutic discovery and its clinical context.

We found that the pattern is very conservative for large sample size and dataset-dependent for a smaller sample size.

P53 Transcriptome profile analysis of ovine ovarian tissues highlights steroids biosynthetic genes in multiparous ewes

Hamed Ghaderi, Tarbiat Modares University IR
Kobra Mostafaee, Tarbiat Modares University IR
Ali Akbar Masoudi, Tarbiat Modares University IR
Babak Arefnejad, OMICS Research Group IR

Recently RNA-Seq has been used to investigate the differentially expressed genes in different species. Prolificacy is one of the important traits in sheep production system. Study of the genetic differences between uniparous and multiparous animals may help to improve breeding

programs in favor of increasing prolificacy in sheep. Shal sheep is an indigenous Iranian breed, which is capable to have a high prolificacy rate. The aim of this study was to analyze the transcriptomes of ovarian tissues in the groups of the uniparous and multiparous Shal sheep. Using Illumina HiseqTM 2000 deep sequencing, a total of 46956986 and 51747856 reads from the samples of uniparous and multiparous Shal sheep were obtained, respectively. Pair-end reads of the sequences were obtained by filtration, and they aligned to the ovis aries reference genome (Ova Version 1.3) using TOPHAT (v2.0.13). After aligning, 26853 genes were recognized to be expressed in ovarian tissue. To identify significantly differential expression profiles between two groups of the animals, Cufflinks (v 2.1.1), Cuffmerge and Cuffdiff (v 2.1.1) softwares were used. The results shown that 619 genes had significant differences ($p<0.05$) between uniparous and multiparous animals. Some studies indicated that GDF9 gene has an important effect on prolificacy in sheep. Therefore, this study focused on chromosome 5, which is the host for this gene. In the current study, 1356 genes were located on chromosome 5. Of these genes, the expression of 28 genes were significantly different between the uniparous and multiparous animals. From these 28 genes, 18 down-expressed and 10 over-expressed were identified in the multiparous animals. Gene ontology terms of KEGG pathway analysis in observed genes showed that the biosynthesis of the steroids pathways is enriched in the ovarian tissue.

P54 DNA Barcodes Adapted to the Illumina Sequencing Platform

Tilo Buschmann, Fraunhofer IZI DE
Leonid Bystrykh, ERIBA, UMCG, University of Groningen NL

The successful completion of multiplexed high-throughput sequencing experiments depends heavily on the proper design of the DNA barcodes. Mutations during barcode synthesis, PCR amplification, and sequencing make decoding of DNA barcodes and their assignment to the correct samples difficult. Previously, we introduced a generalised barcode design for the correction of insertions, deletions, and substitutions which we called the Sequence-Levenshtein distance.

However, generalised barcode designs may be wasteful when applied to specific technologies. The Illumina Sequencing by Synthesis platform shows a very large number of substitution errors as well as a very specific shift of the read that results in inserted and deleted bases at the 5'-end and the 3'-end (which we call phaseshifts). As a solution, we propose the Phaseshift distance that exclusively supports the correction of substitutions and phaseshifts.

Additionally, we enable the correction of arbitrary combinations of substitution and phaseshift errors. Thus, we address the lopsided number of substitutions compared to phaseshifts on the Illumina platform.

To compare codes based on the Phaseshift distance to Hamming Codes (correction of substitution errors) as well as codes based on the Sequence-Levenshtein distance (correction of indels and substitution errors), we simulated experimental scenarios based on the error pattern we identified on the Illumina platform. Furthermore, we generated a large number of different sets of DNA barcodes using the Phaseshift distance and compared codes of different lengths and error correction capabilities. We found that codes based on the Phaseshift distance can correct a number of errors comparable to codes based

on the Sequence-Levenshtein distance while offering the number of DNA barcodes comparable to Hamming codes. Thus, codes based on the Phaseshift distance show a higher efficiency in the targeted scenario.

P55 Computational methods for prediction of selenoprotein genes

Didac Santesmasses, Centre for Genomic Regulation, Barcelona ES
Marco Mariotti, Centre for Genomic Regulation, Barcelona ES
Roderic Guigó, Centre for Genomic Regulation, Barcelona ES

Selenoproteins contain the 21st amino acid selenocysteine (Sec), a selenium-containing cysteine analogue. Sec is co-translationally inserted in response to specific in-frame UGA codons, normally a stop, through a dedicated machinery. The main signal for UGA recoding is a RNA hairpin loop, known as the Sec insertion sequence (SECIS), present in all selenoprotein mRNAs. Although they constitute a very small fraction of the proteome, selenoproteins cover important roles in antioxidant defense, redox regulation, thyroid hormone activation and several others. Since the UGA codon can serve as both stop or Sec signal, standard gene annotation tools do not correctly predict selenoprotein genes, and hence they are usually misannotated in genome projects and protein databases. For this reason, we developed different computational methods to identify UGA-Sec codons in genomic sequences. Our toolkit consists of: Selenoprofiles, a profile-based gene predictor for the annotation of known selenoproteins; Seblastian, a pipeline based on the search of SECIS elements as first step, that can predict new selenoproteins; and Secmarker, for the identification of selenocysteine tRNA gene (tRNA-Sec), a marker for the Sec utilization trait.

Selenoproteins are present in the three domains of life, but not in all species. We characterized the set of selenoproteins present across sequenced genomes, revealing a detailed map of the use of Sec across the tree of life. Using genome sequencing, we traced with precision the path of genomic events that lead to recent independent selenoprotein extinctions in several Drosophila species.

P56 Cancer type-specific pathogenicity estimates reveal somatic and germline variants affecting enhancer function

Aliaksei Holik, Centre for Genomic Regulation, Barcelona ES
Shalu Jhanwar, Centre for Genomic Regulation, Barcelona ES
Stephan Ossowski, Centre for Genomic Regulation, Barcelona ES

A number of recent studies have proposed resources and methods for estimating the effect of non-coding regulatory variants, broadly based on epigenetic marks and evolutionary conservation (e.g. RegulomeDB, CADD, GWAVA). A common drawback of these methods is the lack of consideration for tissue-specificity of the variants analysed, which limits the extent to which the variant annotation can be generalised to different cancer types.

In order to overcome this limitation, we developed approaches that enable us to estimate the damage potential of the regulatory variants in the context of a specific cancer by integrating epigenome information from the relevant tissue types. We make use of the projects generating context-specific epigenome and transcriptome data (e.g. ENCODE and Roadmap Epigenomics) and apply this information in a tissue-specific manner to the 2700 whole cancer genomes spanning 22 cancer types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project.

First, we use tissue-specific epigenome information to train a support vector machine based classifier to identify genomic regions consistent with enhancer activity. The classifier is trained on a small number of tissues, for which a validated set of enhancers is available and using a limited number of features that are consistently available across many tissues (i.e. core histone marks and chromatin accessibility). In order to estimate the damage potential of the variants overlapping these regions, we further train a random forest based classifier using a set of commonly exploited features (histone marks, DNAse hyper-sensitivity, evolutionary conservation, etc), pre-computed scores from existing methods (e.g. CADD), as well as motif-based estimates of variant pathogenicity. We train this classifier using tissue-specific expression quantitative trait loci from GTEx project, as a positive training set. Finally, we apply the classifiers trained on the tissue-specific training sets to the somatic and germline variants from the relevant cancer types from the PCAWG project.

P57 eDIVA: Computational analysis of exome-seq for disease gene identification

Mattia Bosio, Centre for Genomic Regulation, Barcelona ES
Oliver Drechsel, Institute of Molecular Biology (IMB) DE
Rubayte Rahman Netherlands Netherlands Cancer Institute NKI
Stephan Ossowski, Centre for Genomic Regulation, Barcelona ES

Novel or inherited genetic variations can lead to drastic phenotypes including rare and common diseases. Human exome analysis using next generation sequencing (Exome-seq) has recently been established as a key approach to identify genetic variations in protein coding genes. Several tools predict the impact of variants on the mutated gene products, and prioritize

variants to highlight the disease causing ones. Simple and intuitive workflows leading the researcher from sequencing results towards causal variant prediction reducing the false positive rate are needed when selecting potentially disease-causing variants.

We developed eDiVA, www.ediva.crg.eu, an integrated pipeline that massively facilitates and accelerates the analysis of sequencing data and disease causing variant identification. EDiVA goes from exome-seq alignment and recalibration, to SNP prediction as well as insertion and deletion (InDel) detection using multiple tools to decrease false positive rate. Variants are annotated and enriched with functional information, e.g. damage predictions from SIFT and PolyPhen2, OMICs information from dbSNP, 1000genomes, EVS, ExAC, UCSC Genome Browser, KEGG, and OMIM. We developed a rank product based algorithm to prioritize candidate SNPs and InDels using all integrated information. EDiVA supports various disease models, (autosomal dominant, recessive, de novo, X-linked, and compound heterozygous) to identify correctly segregating mutations in small families and parent-child trios.

EDiVA proved its validity in clinical setting, finding causal variant for mendelian diseases such as familial hyperkalemia, mitral valve prolapse, congenital ataxia, myasthenia, cystic fibrosis and phenylketonuria.

We also developed a benchmark algorithm comparing eDiVA against state-of-the-art tools like Pheno-db and PhenGen measuring precision, recall, and ease to find the causal variant. It is fully reproducible, publicly available and based on real data from ClinVar. Benchmarking demonstrates that eDiVA provides superior variant prioritization compared to similar fast-setup algorithms, and achieves equally good results compared to algorithms requiring human fine-tuning and comprehensive phenotype definitions to work properly.

P58 Gene capture approach for a genome draft in Solea senegalensis for sexual genes characterization

Pedro Seoane, Universidad de Málaga ES
Alvaro Lorenzo, IFAPA Centro El Toruño Málaga ES
Manuel Manchado, IFAPA Centro El Toruño, Málaga ES
M. Gonzalo Claros, Universidad de Málaga ES

Nowadays, many medical and biological research is facilitated with the knowledge of the genomic structure of genes. NGS technologies straightens this using two approaches: sequencing the whole genome (this is a cumbersome task requiring supercomputation facilities and assembling skills) or performing gene capture experiments (requiring previous genomic or transcriptomic information). The gene capture method is widely employed in medicine and in non-model organisms. With this technique, researchers used to obtain information about polymorphisms, but several reports indicate that this technology can also serve to obtain the gene structure of the pool of genes of interest. We have developed specific software (called

GeneAssembler) for a gene capture experiment in Pinus pinaster that served to provide the genomic structure of more than 800 pine genes (Seoane et al, in evaluation). It is our purpose to demonstrate that the GeneAssembler algorithm can be applied with standard NGS data of one non-model organism to recover the gene structure of genes of interest starting from the closest protein from another species. In fact, since we are interested in the elucidation of the genetic mechanism involved in S. senegalensis sexual determination, we took all proteins coded in chromosomes Z and W from a related flatfish, Cynoglossus semilaevis. A total of 1498 Z-W proteins were loaded on GeneAssembler together with a male S. senegalensis draft assembly. We obtained a gene recovery percentage average of 24,83% and only 64 of them were nearly completed (protein sequence recovery >90%). In fact, 38 reconstructed genes have the same exon distribution than their C. semilaevis orthologues. We are currently applying this methodology to other S. senegalensis females with the aim of finding sex specific regions.

This research was funded by AQUAGENET project INTERREG IVB SUDOE (ERDF), INIA and EU through the ERDF 2014-2020 (RTA2013-00023-C02-01), and Junta de Andalucía and ERDF (P10-CVI-6075).

P59 Using NGS technologies and workflow management tools for geminivirus analysis in plants

Pedro Seoane, Universidad de Málaga ES
Luis Diaz, Universidad de Málaga ES
Enrique Viguer, Universidad de Málaga ES
Ana Grande, Universidad de Málaga ES
M. Gonzalo Claro, Universidad de Málaga ES

Studies of quasiespecies in viruses are emerging last years. To facilitate them, we have developed QuasiFlow, a workflow designed in AutoFlow that takes advantage of NGS technologies to reconstruct quasiespecies. QuasiFlow firstly separates virus reads from host and other contaminant reads. Then it characterises and computes several key parameters of a virus population, such as recombination events, SNPs, transitions, transversions, indels, quasispecies reconstruction, normalized Shannon index, nucleotide diversity and mutation networks. In the next step, a comparative study of the samples is performed, comprising correlation, ANOVA and PCA analyses of the previously obtained virus population parameters. The results allow to determine which parameters are affected by the experiment and how the samples behave according to their biological origin. We have applied QuasiFlow to two Arabidopsis thaliana plants inoculated with the infectious clone of the begomovirus TYLCV, using HiSeq or MiSeq reads. The results show that the MiSeq reads allows better haplotype reconstruction of TYLCV. Analysis of both HiSeq and MiSeq reads allowed detection of minor quasispecies variants with a frequency of 10-4 to 10-5. In addition, QuasiFlow was used to discover variants and recombinants in mixed infections of tomato plants after 15 and 30 days

post inoculation(dpi). A consensus sequence was generated, showing to be a recombination between begomoviruses TYLCV and TYLVMaV. Haplotype reconstruction shown mutant clouds surrounding haplotypes belonging to the two different begomovirus species and mutant clouds of recombinant nature derived from them. Interestingly, the recombinant haplotypes were the most representative sequences in the mutant spectra at 30 dpi. These results show the fast generation of recombinant genomes in geminivirus mixed infections and demonstrate the potential of Quasiflow for analysis of mutant spectra using Illumina MiSeq sequencing data.

This research was funded by Junta de Andalucía and EU through the ERDF 2014-2020, project P10-CVI-6075.

P60 Exome sequencing in 111 Czech families with inherited cardiovascular diseases

Lenka Piherová, Institute of Inherited Metabolic Disorders, 1st Faculty of Medicine, Charles University, Prague CZ
Viktor Stránecký, Institute of Inherited Metabolic Disorders, 1st Faculty of Medicine, Charles University, Prague CZ
Anna Přistoupilov, Institute of Inherited Metabolic Disorders, 1st Faculty of Medicine, Charles University, Prague CZ
Hana Hartmannová, Institute of Inherited Metabolic Disorders, 1st Faculty of Medicine, Charles University, Prague CZ
Jana Paděrová, Department of Biology and Medical Genetics, 2nd Faculty of Medicine, Charles University, Prague CZ
Alice Krebsova, Department of Cardiology, Institute for Clinical and Experimental Medicine-IKEM, Prague CZ
Milos Kubánek, Department of Cardiology, Institute for Clinical and Experimental Medicine-IKEM, Prague CZ
Vojtech Melenovský, Department of Cardiology, Institute for Clinical and Experimental Medicine-IKEM, Prague CZ
Tomáš Paleček, Department of Cardiovascular Medicine, 1st Faculty of Medicine, Charles University, Prague CZ
Milan Macek, Department of Biology and Medical Genetics, 2nd Faculty of Medicine, Charles University, Prague CZ
Stanislav Kmoch, Institute of Inherited Metabolic Disorders, 1st Faculty of Medicine, Charles University, Prague CZ

Exome sequencing (ES) facilitates genetic diagnostics of inherited cardiovascular diseases, enables genetic stratification and may eventually foster individualized therapies. Due to integrated cardio-genetic care, altogether 111 families with ≥ 2 affected individuals were characterized and ES was carried out in all families (TruSightOne Exome, Illumina, USA). 27 families suffered from dilated cardiomyopathy (DCM), 23 hypertrophic cardiomyopathy (HCM),

35 arrhythmogenic cardiomyopathy (ACM), 4 restrictive cardiomyopathy (RCM), 19 ion channelopathies and 3 had unexplained cardiac arrest (UCA). Detected variants were confirmed by Sanger DNA sequencing and by segregation analysis. ES revealed a putative molecular genetics mechanism in 62/111 (56%) families. The causal mutations were identified in 16/27 (59%) DCM, 18/23 (78%) HCM, 17/35 (49%) ACM, 3/4 (75%) RCM, 6/19 (32%) ion channelopathies and 2/3 families with UCA. The most frequently mutated gene in DCM was TTN – truncated mutations (25%), in HCM the MYH 7 – missense mutation (28%) and in ACM the PKP2 – missense, deletions and splicing mutations (40%). Although disease genes were already identified, most mutations were novel. Our results and mutation distribution are in accordance with other studies. Our model of integrated genetic care of patients with inherited cardiovascular diseases is the first one in Czech Republic.

P61 Annotation of selenoprotein genes in vertebrates and their human polymorphisms

Frederic Romagne, MPI-EVA DE
Elias Mueller, MPI-EVA DE
Didac Santesmasses, Centre for Genomic Regulation, Barcelona ES
Marco Mariotti, Centre for Genomic Regulation, Barcelona ES
Louise White, MPI-EVA DE
Vadim N. Gladyshev, HMS US
Aida Andres, MPI-EVA DE
Roderic Guigo, Centre for Genomic Regulation, Barcelona ES
Sergi Castellano Germany MPI-EVA

Selenoproteins are proteins which contain the amino acid selenocysteine (Sec) as one of their constituent residues. Sec, is the 21st amino acid in the genetic code, and it is analogous to cysteine (Cys) but with a selenium atom in place of sulfur. It is specified in the mRNA by an in-frame UGA (STOP) codon which, in conjugation with an RNA structure in the mRNA (SECIS), is recoded to incorporate the Sec residue instead of terminating protein synthesis. Because of the dual meaning of the UGA codon in genomes, gene prediction algorithms usually fail to identify selenoproteins and either predict shorter or wrong (incorrect coding exons) protein sequences. Thus, genome annotations for selenoproteins are usually wrong. To correct this, we have used a gene prediction tool (Selenoprofiles), specifically designed to handle the uncommon features of selenoproteins, and annotated the selenoprotein genes, proteins and SECIS elements in 58 vertebrate genomes, including humans. Because selenium is an essential micronutrient in the human diet, and humans have encountered environments with selenium deficiency or toxicity as they settled the world, we have surveyed the patterns of single nucleotide polymorphism (SNP) in all selenoprotein genes and genes involved in selenium metabolism in 50 human populations. We find that humans have genetically adapted to the environments that do not provide adequate levels of selenium and are associated with selenium-related diseases.

Selenoprotein genes and their SNPs are available at SelenoDB2.0 (http://www.selenodb.org)

P62 Small-RNA-seq of HIV infected human T cells reveals the existence of HIV-encoded miRNA-like molecules and the generation of host antisense small RNAs

Andreia Amaral, University of Lisboa, Faculty of Sciences, BioISI – Biosystems & Integrative Sciences Institute, Campo Grande, Lisboa PT
Russel Foxall, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa PT
Francisco Pinto, University of Lisboa, Faculty of Sciences, BioISI – Biosystems & Integrative Sciences Institute, Campo Grande, Lisboa, PT
Paula Matoso, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa PT
Rui Soares, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa PT
Rita Tendeiro, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa PT
Ana Sousa, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa PT
Margarida Gama-Carvalho, University of Lisboa, Faculty of Sciences, BioISI – Biosystems & Integrative Sciences Institute, Campo Grande, Lisboa PT

The Human immunodeficiency virus (HIV-1 and HIV-2) has been the subject of intense investigation, and a great deal has been learned about how the retrovirus infects cells, replicates, and causes disease. However there is still a lot to uncover in relation to the RNA machinery involved.

Small interfering RNAs (siRNAs) and microRNAs (miRNAs) are small noncoding RNAs (sncRNAs) involved in the regulation of gene expression. MiRNAs, have been identified in animals, plants and in DNA viruses. These display size range of eukaryotic miRNAs but display interspecies functional conservation and not sequence conservation. Regarding RNA viruses, it has been generally assumed that these do not encode miRNAs, with the argument that the production of canonical miRNAs from a pri-miRNA hairpin would result in unproductive cleavage of the viral genome and its transcripts. As is the case for other viruses, HIV-1 could highly benefit from encoding miRNAs. In contrast to viral proteins, viral miRNAs are non-immunogenic, require less coding capacity and can evolve rapidly to target new transcripts. However, the existence of HIV-1 encoded miRNAs remains controversial, and regarding HIV-2, this hypothesis has never been investigated.

CD4+ T cells are the main targets of HIV and the central orchestrators of immune responses. To identify novel regulators of this process, we used next generation sequencing to profile changes in the expression of sncRNAs in response to HIV infection (HIV-1 and HIV-2) of in vitro stimulated human naive CD4 T cells.

Our results revealed the existence of miRNA-like molecules encoded by HIV-1 and HIV-2, which display interspecies functional conservation and can potentially regulate the expression of genes involved in T cell activation, signaling and activation of immune response. Moreover, we find evidence for the activation of virus-specific antisense small RNAs derived from host repeat sequences, which may represent endogenous siRNA molecules.

P63 BenchCT & SimCT, a multi-purpose benchmarking workflow for RNA-Seq analysis

Jérôme Audoux, Institute for Regenerative Medicine & Biotherapy, Computational Biology Institute FR
Mickaël Salson, Laboratoire d'Informatique Fondamentale de Lille FR
Anthony Boureux, Institute for Regenerative Medicine & Biotherapy, Computational Biology Institute FR
Thérèse Commes, Institute for Regenerative Medicine & Biotherapy, Computational Biology Institute FR
Nicolas Philippe, Institute for Regenerative Medicine & Biotherapy, Computational Biology Institute FR

Throughout the past years, the unprecedented evolution of next-generation sequencing (NGS) technologies has shaped computational biology and facilitated personalized medicine. Transcriptomics studies have increasingly come to rely on the use of NGS to directly sequence libraries of short sequences at nucleotide scale arising from the transcriptome (RNA-seq). Many tools have been developed for RNA-seq analysis but good choices will specifically depend on the biological question. Indeed the underlying algorithmic machinery of each method, stacking a plurality of tools to create a full analysis pipeline is tricky and requires a particular attention. Several global studies benchmarking RNA-Seq analysis software have been published since the advent of RNA-Seq but each of them focuses on a specific aspect whether technological or methodological. Here, we propose to fill this gap by introducing a complete benchmarking workflow from the generation of synthetic dataset to the assessment of analysis pipeline results.

For that purpose, we have developed SimCT and BenchCT. The former is an integrated pipeline that includes FluxSimulator with a layer of genomic mutations (SNP, Indels, chimeras). The later is a software able to assess RNA-Seq analysis results and it handles various types of file formats produced by the analysis pipelines from read alignment to mutation discovery.

By introducing these softwares, we aim to make benchmarking a regular bioinformatic practice. Such a standard could be used: i/ by bioinformaticians for finding and tuning a pipeline in a given biological and technological context ; ii/ by companies or academic platforms for quality control in order to evaluate the impact of a software or genomic reference updates ; ii/ by developers for debugging and optimization purposes. In order to propose a showcase for our

modular benchmarking framework, we have illustrated three restricted comparative studies on the discovery of splicing events, small mutations and chimeric RNAs.

P64 Identification of global regulators of T-helper cell lineage specification

Kartiek Kanduri, Turku Centre for Biotechnology, University of Turku and Åbo Akademi University FI
Subhash Tripathi, Turku Centre for Biotechnology, University of Turku and Åbo Akademi University FI
Antti Larjo, Aalto university FI
Henrik Mannerström, Aalto university FI
Ubaid Ullah, Turku Centre for Biotechnology, University of Turku and Åbo Akademi University FI
Riikka Lund, Turku Centre for Biotechnology, University of Turku and Åbo Akademi University FI
R. David Hawkins, University of Washington US
Bing Ren, University of California, San Diego US
Harri Lähdesmäki, Aalto university FI
Riitta Lahesmaa, Turku Centre for Biotechnology, University of Turku and Åbo Akademi University FI

Activation and differentiation of T-helper (Th) cells into Th1 and Th2 types is a complex process orchestrated by distinct gene activation programs engaging a number of genes. This process is crucial for a robust immune response and an imbalance might lead to disease states such as autoimmune diseases or allergy. Therefore, identification of genes involved in this process is paramount to further understand the pathogenesis of, and design interventions for, immune-mediated diseases. We aimed at identifying protein-coding genes and long non-coding RNAs (lncRNAs) involved in early differentiation of T-helper cells by transcriptome analysis of cord blood-derived naïve precursor, primary and polarized cells. Here, we identified lineage-specific genes involved in early differentiation of Th1 and Th2 subsets by integrating transcriptional profiling data from multiple platforms. We have obtained a high confidence list of genes as well as a list of novel genes by employing more than one profiling platform. We show that the density of lineage-specific epigenetic marks is higher around lineage-specific genes than anywhere else in the genome. Based on next-generation sequencing data we identified lineage-specific lncRNAs involved in early Th1 and Th2 differentiation and predicted their expected functions through Gene Ontology analysis. We show that there is a positive trend in the expression of the closest lineage-specific lncRNA and gene pairs. We also found out that there is an enrichment of disease SNPs around a number of lncRNAs identified, suggesting that these lncRNAs might play a role in the etiology of autoimmune diseases. The results presented here show the involvement of several new actors in the early differentiation of T-helper cells and will be a valuable resource for better understanding of autoimmune processes.

P65 An extensible genome annotation workbench based on the Galaxy platform

Peter van Heusden, South African National Bioinformatics Institute (SANBI) UWC ZA
Shubha Vij, Temasek Life Sciences Laboratory SG
Laszlo Orban, Temasek Life Sciences Laboratory SG
Alan Christoffels, South African National Bioinformatics Institute (SANBI) UWC ZA

The development of low cost genome sequencing using next generation technology has enabled the proliferation of genome annotation projects. The data generated by these projects requires gene prediction, annotation and subsequent analysis. While there are toolkits (MAKER2, Ensembl Genebuild, Sma3s, annot8r) that automate the annotation of novel genomes, these operate as black boxes that hide the details of their operation from the user and are difficult to extend and customise.To address this limitation, we have created a genome annotation workbench based on the Galaxy platform. This platform makes our platform extensible and allows for reproducible annotations through capture of all details relevant to data provenance. We demonstrate its use for the annotation of a novel genome (the Asian seabass). We contrast its operation and resulting annotation with MAKER2.

P66 Expanding the repertoire of small secretory peptides in plants

Yao-Cheng Lin, VIB Department of Plant Systems Biology, Ghent University BE

Plant genomes encode numerous small secretory peptides (SSPs) the functions of which remain yet to be explored. Based on structural features that characterize SSP families known to take part in postembryonic development, our comparative genome analysis resulted in the identification of genes coding for oligopeptides potentially involved in cell-to-cell communication. Because genome annotation based on short sequence homology is difficult, the criteria for the de novo identification and aggregation of conserved SSP sequences were first benchmarked across five reference plant species. The resulting gene families were then extended to 32 genome sequences, including major crops. The global phylogenetic pattern common to the functionally characterized SSP families suggests that their apparition and expansion coincide with that of the land plants. The SSP families can be searched online for members, sequences and consensus (http://bioinformatics.psb.ugent.be/webtools/PlantSSP/). Looking for putative regulators of root development, Arabidopsis thaliana SSP genes were further selected through transcriptome meta-analysis based on their expression at specific stages and in specific cell types in the course of the lateral root formation. As an additional indication that formerly uncharacterized SSPs may control development, we showed that root growth and branching were altered by the application of synthetic peptides matching conserved SSP motifs, sometimes in very specific ways. Our strategy combining comparative genomics, transcriptome meta-analysis and peptide functional assays in planta pinpoints factors potentially involved in non-cell-autonomous regulatory mechanisms. A similar approach

can be implemented in different species for the study of a wide range of developmental programs.

Reference: Expanding the Repertoire of Secretory Peptides Controlling Root Development with Comparative Genome Analysis and Functional Assays. J. Exp. Bot. (2015) 66 (17): 5257-5269

P67 Assessment of Gene Contribution by Tissue in Teleost Species

Jorge Langa, University of the Basque Country UPV/EHU ES
Darrell Conklin, University of the Basque Country UPV/EHU, Ikerbasque Foundation ES
Andone Estonba, University of the Basque Country UPV/EHU ES

The class of bony fishes comprises the great majority of fish species on the planet and they are represented only by 10 reference genomes at Ensembl, the most studied being the Zebrafish.

The disruption of Second Generation Sequencing technologies unlocked the exploration of genomes and transcriptomes of non-model species. Small laboratories now have access to cheap, high throughput and cost-effective methods that enable the molecular description of entire organisms via whole genome and transcriptome sequencing.

RNA expression is influenced by tissue type, which can lead to transcripts being fragmented or unexpressed. For the generation of a de novo assembly, one should be able to select those most representatives and generate a reference as completely as possible.

It is known for genome assembly that there is a point at which as sequencing depth increases, the quality of the assembly decreases, which also holds true for transcriptome assembly. Therefore, one has to know which tissues would be optimal to choose and how deep.

Here we present the study a freely available RNA-Seq Illumina dataset from ENA made up of 12 tissues from Danio rerio. To assess the contribution of each tissue, we: 1) subsampled each library from 1M to 20M PE reads in steps of 1M; 2) performed Quality Control with Trimmomatic; 3) mapped each sublibrary to the reference with Hisat2; 4) assembled the transcripts with StringTie; 5) Performed de novo assemblies with Trinity; and 6) Measured the completeness of the libraries reference-based and de novo.

The experiment concludes that the tissues yielding more transcripts are brain, testis and embryos, whilst kidney, unfertilized eggs and ovaries were the ones that worst performed.

We hope this study is a step towards allowing experimentalists to do de novo reference construction of non-model fish transcriptomes as completely as possible and in a cost-effective way.

P68 The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea

Yao-Cheng Lin, VIB Department of Plant Systems Biology, Ghent University BE

Seagrasses colonized the sea on at least three independent occasions to form one of the most productive and widespread coastal ecosystems on the planet. The genome of Zostera marina (L.), the first marine angiosperm to be fully sequenced, reveals unique insights into the genomic losses and gains involved in achieving the structural and physiological adaptations required for its marine lifestyle, arguably the most severe habitat shift ever accomplished by flowering plants. Key angiosperm innovations that were lost include the entire repertoire of stomatal genes, genes involved in the synthesis of terpenoids and ethylene signaling, and genes for UV protection and phytochromes for far-red sensing. Seagrasses have also regained functions enabling them to adjust to full salinity. Their cell walls contain all of the polysaccharides typical of land plants but also polyanionic, low-methylated pectins and sulfated galactans, a feature shared with the cell walls of all macroalgae and important for ion homoeostasis, nutrient uptake and O2/CO2 exchange through leaf epidermal cells. The Z. marina genome resource will significantly advance a wide range of functional ecological studies from adaptation of marine ecosystems under climate warming to unravelling the mechanisms of osmoregulation under high salinities that may further inform our understanding of the evolution of salt-tolerance in crop plants.

Reference: The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. *Nature* (in press)

P69 Detecting novel intergenic structured non-coding RNAs using Capture based sequencing

Christian Anthon, University of Copenhagen DK
Claus Hansen, University of Copenhagen DK
Aashiq Mirza, University of Copenhagen DK
Claus Heiner Bang-Berthelsen, Herlev Hospital DK
Niels Tommerup, University of Copenhagen DK
Flemming Pociot, Herlev Hospital DK
Stefan Seemann, University of Copenhagen DK
Jan Gorodkin, University of Copenhagen DK

Based on a genome wide scan for RNA structure potential in the human genome, we found around 800,000 such regions with RNA structure conservation in vertebrate genomes. Many of these regions are found in known genes such as UTRs of protein coding genes, and structured ncRNAs (both short and long ncRNAs), however, a substantial part of the 800,000 regions are found were genes and transcription have not been reported previously.

Standard methods for high throughput sequencing will not be suited to detect transcripts of very low expression since the resulting read counts will always be dominated by mainly protein coding genes that are several orders of magnitudes more expressed. In a recent development, high throughput sequencing is preceded by a capture step where selected transcripts are targeted using designed capture probes, which allows for detection of low-abundant transcripts. Previously capture based sequencing have mainly been used to investigate known protein coding genes, to for example detect additional exons.

In this study we have used capture based sequencing to investigate a subset of the 800,000 structurally conserved regions we detected in the human genome. We have focused on the intergenic regions where transcription is previously unreported, but we have complemented these intergenic regions with regions overlapping known structured ncRNAs. Early results indicates that 1000s of new transcripts may be detected this way.

P70 Integrating transcript information into pathway analysis

Felix Eichinger, University of Michigan US

In spite of the establishment of technology able to measure transcript abundance, such as modern microarrays and in particular RNAseq, most mRNA expression studies are still using gene level expression. In theory, conducting analyzes on transcript level appears more desirable: both technologies essentially measure the abundance of transcripts or parts thereof. Additionally there is less ambiguity in the relationship between a transcript and the functional unit, the protein, than between a gene and the protein. However, there are several reasons for the slow adaption of transcript level analyzes. For one, it is much easier to assign fragments of mRNA to genes than to individual transcripts and correctly quantify expression. Secondly, most enrichment and pathway analysis tools work on gene level, even if they accept transcript identifiers as input. Consequently, potential advantages of transcript level information do not carry through the entire analysis, further discouraging the generation and use of it. Here we propose to explore ways to integrate transcript level information into downstream analysis tools. As an initial, simple approach, we build on the capabilities of the bioconductor graphite package. This package can download and analyze pathways from 6 different sources. In order to integrate transcript level information, we expand the gene level nodes by transcripts associated with these genes. This approach directly enables scientists to see which transcripts are involved in activation of a pathway and to see if different transcripts are used in different tissues or conditions. In a secondary step, we will try to integrate the transcript information into the package in a way that the internal analysis methods can be used to analyze differential regulation of pathways, identification of key molecules for regulation/deregulation of a pathway and identification of the most relevant signal propagation path.

P71 Composition of the gut microbiome of colorectal cancer patients from Morocco

Imane Allali, Faculty of Sciences, University Mohammed V, Rabat MA?
Noureddine Boukhatem, Faculty of Sciences, University Mohammed Premier, Oujda MA
Leila Bouguenouc, University hospital Hassan II of Fez MA
Hanaa Hardi, Faculty of Sciences, University Mohammed Premier, Oujda MA
M. Belen Cadenas, University of North Carolina, Chapel Hill, NC US
Karim Ouldim, University hospital Hassan II of Fez MA
Saaïd Amzazi, Faculty of Sciences, University Mohammed V, Rabat MA
M. Andrea Azcarate-Peril, University of North Carolina, Chapel Hill, NC US
Hassan Ghazal, Polydisciplinary Faculty of Nador & Faculty of Sciences of Oujda, University Mohammed Premier, Oujda/Nador MA

Colorectal cancer (CRC) is the third most common cancer in the world and the third leading cause of cancer mortality in Morocco. The colonic mucosa is permanently in contact with the microbiota and its metabolic products, which can potentially induce oncogenic transformation. The molecular mechanisms involved in the etiology of CRC are not yet elucidated due in part to the complexity of the human gut microbiota. The aim of this study was to characterize the gut microbiota of CRC Moroccan patients to identify bacterial taxa over or under represented in stools from CRC patients compared to healthy subjects by 16S rRNA amplicon sequencing. Our results showed higher Phylogenetic Diversity (PD) and Species Richness (S) in CRC samples. Principal Coordinates Analysis (PCoA) revealed that CRC samples clustered separately from controls (ANOSIM P=0.008, R=0.2039, and PERMANOVA P=0.005, F=1.8976) suggesting differences in the microbiome associated to CRC. Our findings indicate that CRC samples were enriched in Firmicutes (T=50.5%; N=28.4%; P=0.04) and Fusobacteria (T=0.1%; N=0.0%; P=0.02) while Bacteroidetes were enriched in healthy samples (T=35.1%; N=62.8%; P=0.06). Despite the small number of patients included in the study (11 CRC patients, 12 healthy controls), we observed significantly overrepresented genera in the CRC group compared to controls. Porphyromonas (T=0.6%; N=0.0%; P=0.04), Clostridium (T=0.2%; N=0.1%; P=0.02), Ruminococcus (T=0.6%; N=0.5%; P=0.02), and Fusobacterium (T=0.1%; N=0.0%; P=0.03) were over represented in CRC patients while Megamonas (T=0.0%; N=0.4%; P=0.04) was over represented in controls. This is the first study conducted in the Moroccan population that aimed to characterize the CRC gut microbiome. Data from this small cohort warrant a larger study that will include CRC patients from different Moroccan locations. Understanding the relationship between CRC and the intestinal microbiota will lead to the development of novel strategies for the diagnosis, treatment, and prevention of this disease.

P72 ORCAE: A wiki-style platform enabling efficient community curation of gene and genome annotations

Lieven Sterck, VIB-UGent BE
Thomas Van Parys, VIB-Ugent BE
Stephane Rombauts, VIB-Ugent BE
Pierre Rouzé, VIB-Ugent BE
Yves Van de Peer, VIB-UGent BE

Conducting gene and genome annotation typically relies on diverse information resources going from sequence to expression data depending on whether structural or functional annotation is performed. To help researchers doing gene annotation while having access to these different data types, we developed ORCAE (Online Resource for Community Annotation of Eukaryotes), a web-technology-compliant portal for use in community genome annotation efforts.

ORCAE allows browsing and on the fly editing of gene descriptions as well as gene structures, moreover all manual curations are immediately visible for other users. The portal will store all the modification from annotators in the database so for each locus a history of modifications is available.

Through its interface, ORCAE offers easy access to precomputed information that greatly facilitates the work of a curator. The gene page offers several informative graphics with a focus on the quality of the gene structure (eg. Multiple alignments of similar proteins, tiling array information …) helping the human annotators in improving the proposed automated annotation. Annotators can then use the build-in GenomeView interface to easily check/modify gene structures. A unique feature from ORCAE is that the portal is highly dynamic, all the available information (eg. protein similarity, transcript alignments, etc.) is immediately updated and presented on the gene page.

ORCAE can both be used to coordinate ongoing annotation efforts in the course of the project as well as to present published genomes to the publicby acting as a genome portal. Therefore it is equipped with all the necessary features to act as a public genome browser/portal: such as advanced text-search and Blast functionality as well as a genome browsing interface (AnnoJ).

Currently it offers public access to 16 eukaryotic genome projects and restricted access to another 32 genomes.

ORCAE is available at http://bioinformatics.psb.ugent.be/orcae/.

P73 de-novo annotation and accurate quantification of alternative splicing from RNAseq data

Panagiotis Papasaikas, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona ES
André Gohr, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona ES

Claudia Vivori, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona ES
Juan Valcarcel, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona ES

Alternative Spicing (AS) is prevalent among all metazoa and plays a key role in establishing the differential expression profiles that underlie cell differentiation while its misregulation is known to be involved in a wide array of human disease. Therefore, accurate and fast quantification of AS based on RNA sequencing (RNAseq) data is pivotal for monitoring isoform transition in different physiological or pathological contexts and for deciphering the mechanisms that underlie its regulation. To address limitations in existing methodologies, we developed SANJUAN (Splicing ANalysis and JUnction ANnotation), an integrated pipeline that efficiently identifies, quantifies and fully annotates differentially used splicing junctions in RNAseq samples. Our approach is insensitive to junction mappability biases and, importantly, does not rely on pre-existing transcriptome annotation for the identification or the quantification of AS events. This is especially critical in setups that focus on unexplored or insufficiently probed cellular contexts or experiments that perturb the splicing circuitry resulting in the utilization of multiple novel or non-canonical splice sites. Both simulated and experimental data on multiple cell-types and splicing-insult conditions demonstrate the sensitivity and accuracy of our approach. We expect our method to prove valuable for the exploration of the impact and the mechanistic elucidation of AS. SANJUAN is available at https://github.com/ppapasaikas/SANJUAN/

P74 Redefining a consistent microbial reference database

Zech Xu, University of California San Diego US
James Morton, University of California, San Diego US

With plunging cost of high-throughput sequencing, enormous amount of microbial genomic and metagenomic sequences are deposited into databases. The interpretation of these sequences requires consistent and high-quality annotation, which has been proven to be a major challenge in maintaining publically available databases such as Genbank. Here we have implemented a full, open-source annotation pipeline for bacterial and archaeal genomes, micronota. It wraps on the state-of-the-art tools to predict DNA features including coding genes, ncRNAs, CRISPR, prophage, and more. To provide consistent annotation across microbial genomes, micronota interfaces with multiple databases such as UniRef, UniProt, TIGRFam and PFam. Furthermore, micronota is actively developed and maintained with scalability and customizability in mind, which allows users to customize the individual tools and databases and annotate microbial genomes in their own laptop.

P75 An insides job: role of the human eye microbiome in development of conjunctivitis

Mariam Lotfy, Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cinical Operations Department - Ray-Clinical Research Organization, Dokki EG
Moamen Elmassry, Department of Biological Sciences, Texas Tech University, Lubbock, TX, US
Jarrad Marcell, Genomic and Systems Biology, Bioscience Division, Argonne National Laboratory US
Rania A. Khattab, Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University EG
Maha M. Abdelfattah, Department of Microbiology, Research Institute of Ophthalmology, Giza, EG
Jack A. Gilbert, Genomic and Systems Biology, Bioscience Division, Argonne National Laboratory-Department of Surgery, University of Chicago, Chicago, IL, US
Ramy K. Aziz, Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University EG

Since the Human Microbiome Project (HMP) was launched to explore human-associated microbes at different anatomic sites (mouth, nose, skin, colon, and vagina), major discoveries were made on the impact of the human microbiome on human health, disease, and immunity status. Although the human eye is in contact with the environment and is thus exposed to various types of microbes, studies on the eye microbiome are still lagging behind, perhaps because of dearth of cultured bacteria isolated from the eye. Here we aim to discover the role of resident eye microbes in the development of bacterial conjunctivitis, an infection that affects millions of people and ranges from self-limiting to severe infection that may lead to blindness. To this end, we collected 108 conjunctival swabs from 54 patients, and, to minimize interindividual variations, we compared infected and uninfected eyes of each patient. Of the collected samples, 64 were culture positive, and their commonly isolated bacteria were Staphylococcus aureus followed by S. epidermidis and Acinetobacter. Antimicrobial susceptibility profiles of the cultured microbial communities were determined, but were mostly similar in both eyes. The microbiomes of 48 samples (from 24 subjects) were further analyzed by high-throughput 16S sequencing. Many bacterial taxa were determined that were undetected by culture-based techniques, among which are Moraxella, Micrococcaceae and Corynebacteria. Overall, the eye microbiomes clustered together but were closest to skin and nose microbiome samples analyzed as a part of the HMP. Beta diversity between patients was higher than the diversity between eyes of the same patients; however, no significant differences were seen between the two groups. In conclusion, our findings suggest that conjunctivitis is an inside job, i.e., caused by resident eye microbes that expand subsequent to an eye injury or perhaps an undetermined viral infection.

P76 Mining Genomes: Using Statistical Learning Methods to Annotate Metagenomic Sequencing Data

Darryl Reeves, Weill Cornell Graduate School of Medical Sciences US
Christopher Mason, Weill Cornell Medical College US

Discovery of genomic signatures for discriminating between the sequences of organisms has a long history. These signatures are typically constructed using the k-mer (sub-sequences of length k) composition of genomic sequences. While k-mers are useful tools for the identification of organisms in sequencing data, the sensitivity increases as the value of k increases. Larger k-mers contain more useful discriminative information but the computational cost of managing these k-mers increases with the size of k. Utilizing feature hashing, KMerge is a framework which preserves the information available in these k-mers without the memory requirements needed to utilize raw k-mer sequences. This work explores the utility of k-mers and their frequencies as a tool for statistical learning on genomic sequencing data. KMerge is used as the foundation of a method for metagenomic sequence classification and abundance estimation named Rapid Identification and Quantification of Organisms (RIQO) which is a novel algorithm leveraging a bayesian network formalism to advance metagenomic research.

P77 A Deep Belief Network Classifier for Assignment of 16S rRNA Sequences Using Fourier Analysis

Guillermo Luque Y Guzmán Sáenz, Universidad de los Andes CO
Alejandro Reyes Muñoz, Universidad de los Andes CO

Bacterial life is able to develop in diverse ecosystems, and given its abundance it plays an essential role in multiple biochemical interactions[1]. Clustering of 16S rRNA amplicons to identify operational taxonomic units (OTU) is one of the most common approaches used nowadays to elucidate bacterial communities composition. Nevertheless, these methods can throw different results depending on either the way to measure genetic sequence similarity, chosen parameters, or the setting of a dissimilarity threshold, moreover in closely related organisms[2]. Here we present a new classifier not-based on clustering and trained with labeled sequences stored in the Greengenes reference database[3], that can be used to assign taxonomies to sequence fragments flanked by U515F and E333R/E529R primers, which have shown a good performance in the amplification of a broad range of phylotypes in varied community samples[4]. Each new sequence presented to the classifier is projected onto a orthogonal space where structural information might be preserved. In order to do that, the DNA sequence is considered as a composition of four binary signals with ones on the positions corresponding to each nucleotide and zeros on the other ones. Then we can get the spectrum of this compound applying a Discrete Fourier Transform (DFT)[5] and use it as an input for a first Deep Belief Network (DBN), previously trained in an unsupervised way and made of several layers of Continuous Restricted Boltzmann Machines (CRBM)[6], to get a new representation of the original sequence which attains information not only from the input but from the hidden

layers. The new obtained feature is then progressively classified using another set of DBN configured as Multi-layer Neural Networks trained to recognize a specific taxonomic rank. The precision and recall of the proposed algorithm, as well as its accuracy are compared to other commonly used classifiers in microbiome analysis.

P78 A new tool to harness annotated genomes for improved understanding genome function of investigated genome.

Yoram Shotland, Chemical Engineering, Shamoon College of Engineering IL
Daniel Khankin, Software engineering, Shamoon College of Engineering IL
Shlomo Mark, Software engineering, Shamoon College of Engineering IL

Nowadays when immense annotated genomes are available one of the challenges is to harness this vast amount of data to better and dipper understanding of genome function to improve its genome annotation. Towards this goal we developed a new tool that by comparing newly discovered and automatically annotated genome to other annotated genomes can sub group a set of genes required for a define function of the investigated organism.

The tool will be demonstrated when it was used to track functions related with arid lifestyle. We used the new bioinformatics tool to track possible enzymes and metabolic pathways, unique for arid inhabitant, in a cyanobacterium, Leptolyngbya ohadii, isolated from biological send crust. We compared genes found in the Leptolyngbya, with other cyanobacteria organisms, which occupied arid and non-arid environments. Genome comparison was performed in a two stage approach. Firstly, we grouped all homologous genes found in desiccation tolerant cyanobacteria. Then we used genomic information from freshwater organisms to deduct all housekeeping genes from our list of homologous genes. The sequence comparison was done using our newly developed tool, which preform round of iterations. In each iteration, by using NCBI tblastn, it compare between list of predicted Leptolyngbya ohadii protein and the genome currently investigated. Based on the criteria it omit all proteins under a defined threshold of similarity, while it keeps and transfer the remaining list of protein to the next iteration stage.

The analysis identified 52 genes found in the desiccation-tolerant but not in the freshwater cyanobacteria. Some of the genes found in this screen are located adjacent to each other, as a group, on the same region of the genome, which may indicate these genes are on the same operon. RNA-Seq study of the Leptolyngbya ohadii while exposed to dehydration condition supports the bioinformatics findings.

P79 Comparative Genomics of Klebsiella pneumoniae IIEMP-3 Isolated from Indonesian Tempeh with some Pathogenic Strains

Adi Yulandi, Faculty of Biotechnology Atma Jaya Catholic University of Indonesia ID
Mahaldika Cesrany, Department Biology, Faculty of Science and Mathematics, Bogor Agricultural University, Bogor ID
Antonius Suwanto, Department Biology, Faculty of Science and Mathematics, Bogor Agricultural University, Bogor ID

Tempeh is the most consumed traditional fermented soy food from Indonesia. During the fermentation process, Rhizopus and other microorganisms converted soybeans into a product, which is, not only increase in nutritional value, but also generate some important vitamins, minerals, and antioxidants. The present of vitamin B12 in some Indonesian tempeh makes tempeh as one of the world's first meat analogs. The presence of Klebsiella pneumoniae in Indonesian tempeh is intriguing since it produces vitamin B12 in tempeh, while also known as an opportunistic pathogen. K. pneumoniae IIEMP-3 isolated from tempeh exhibited different genetic profiles from strains known as human pathogens. Using NCBI Prokaryotic Genome Annotation Pipeline, a total of 5,285 genes were identified from Whole Genome Sequence (WGS) of K. pneumoniae IIEMP-3. Compared with the pathogenic strains K. pneumoniae subsp. pneumoniae HS11286 and NTUH-K2044, the IIEMP-3 genome did not harbor genes for Type IV secretion systems (virB3-4, virB5, virB6). This system commonly used by pathogenic Gram-negative bacteria to translocate a wide variety of virulence factors into the host cell

---

P80 pEffect predicts bacterial type III effector proteins

Tatyana Goldberg, Technical University of Munich DE
Burkhard Rost, Technical University of Munich DE
Yana Bromberg, Rutgers State University US

The type III secretion system transports effector proteins of pathogenic and endosymbiotic Gram-negative bacteria into the cytoplasm of host cells. During infection, effectors convert host resources to work to bacterial advantage. Existing computational methods for the prediction of type III effectors mainly employ information encoded in the N-terminal protein sequence. Here we introduce pEffect, a method that predicts type III effector proteins using the entire amino acid sequence. It combines homology-based inference with de novo predictions, reaching 87±7% accuracy at 95±5% coverage for a large non-redundant set of proteins. This performance is up to 3-fold higher than that of other methods. pEffect also sheds new light on effector secretion mechanisms. We establish that signals for the recognition of type III effectors are distributed over the entire protein sequence instead of being confined to the N-terminus. Our method, therefore, maintains high performance even when used with sequence fragments like metagenomic reads, and potentially facilitates studies of microbial community interactions. Explorations into the evolutionary origins of type III secretion identify a variety of recently evolved effectors and highlight the possibility of type III secretion ancestor dating to times prior to the archaea/bacteria split. pEffect is available at

http://www.bromberglab.org/services/pEffect.

P81 Genomic platform and custom bioinformatics to handle Next Generation Sequencing data and other data sets at the CBMSO

Ramón Peiró, Centro de Biologia molecular Severo Ochoa, Madrid ES
Maria Jose López-Sánchez, Centro de Biologia molecular Severo Ochoa , Madrid ES
Sandra Gonzalez-De La Fuente, Centro de Biologia molecular Severo Ochoa, Madrid ES
Sandra Gonzalo-Flores, Centro de Biologia molecular Severo Ochoa, Madrid ES
Manuel Belda, Centro de Biologia molecular Severo Ochoa, Madrid ES
Laura Tabera, Centro de Biologia molecular Severo Ochoa, Madrid ES
Fernando Carrasco-Ramiro, Centro de Biologia molecular Severo Ochoa, Madrid ES
Begoña Aguado, Centro de Biologia molecular Severo Ochoa, Madrid ES

The Genomics and Next-Generation Sequencing Facility at the Severo Ochoa Molecular Biology Center (CBMSO) brings together over 10 years of experience in the implementation and development of leverage cutting-edge technologies in molecular biology and genomics, and over 6 years in Next-Generation Sequencing (NGS).

In the Genomics and NGS core, we help to CBMSO internal researchers and, to both external academic and clinical users, with scientific and technical support in the experimental design, undertaking, and data analysis of real-time PCR, microarrays and NGS experiments. In addition, our service mediates between sequencing platforms and users with samples and data, monitoring the development of the projects. We are actively exploring, validating, optimizing, and implementing new NGS technologies and methods.

Concerning to NGS, we provide computational analysis of NGS experiments from a wide range of organisms that comprise model organisms (Drosophila, zebrafish, rat and mouse), bacteria, fungi, mammals (human and dog), algae, parasites, birds, and viruses, among others. Our service principally offers data analysis from RNA-seq, ChIP-seq, re-sequencing and the novo sequencing, amplicon (16S, 18S,…) and Clip-seq studies. We are actively learning and developing methodologies for high-throughput sequencing computational processing, and working on a variety of methods from the data assembly, gap filling, genome annotation and curation to understand the NGS data. In addition, we perform statistical analysis, and custom software. Finally, we generate back top-quality data and information and reports along the project, with rapid turnaround times at a competitive cost.

From 2014 the platform coordinates the participation of several CBMSO researchers in the MinION Access Program of Oxford Nanopore. In addition, the facility organizes seminars and on-demand courses for users in its areas of specialization.

P82 High-quality functional genome annotation through an intercampus competition initiative

Steve Caruso, University of Maryland Baltimore County US
James Hu, Texas A&M University US
Ivan Erill, University of Maryland Baltimore County US

Ensuring high-quality functional annotations in newly sequenced genomes has become a fundamental problem in next-generation sequencing genomics. This problem takes additional relevance in bacteriophage genomics, where well-annotated reference genomes are scarce and often highly diverged from newly sequenced genomes. This makes automated annotation pipelines especially prone to introduce and propagate functional annotation errors, devaluing the contribution of sequencing projects to the understanding of bacteriophage biology and decreasing the likelihood that novel genetic mechanisms with industrial and clinical application can be mined from the extremely diverse genetic repertoire of bacteriophage genomes. Genome annotation by expert biocurators ensures high-quality annotations, but such approaches are hard to scale and raise financial sustainability issues. Here we show how undergraduate biology courses devoted to genome annotation can be leveraged to generate high-quality functional annotations of bacteriophage genomes by integrating targeted training within the framework of an intercampus annotation competition. Students receive advanced training on biological ontologies, orthology assessment methods, biocuration standards and critical reading of scientific manuscripts. Working in small teams, students then participate in the Community Assessment of Community Annotation with Ontologies intercampus competition, generating Gene Ontology (GO) annotations of bacteriophage and bacterial gene products and challenging the accuracy of annotations made by other teams in a pre-established set of innings for annotation, challenge and revision. Revised annotations are reviewed by an expert panel before submission to the Gene Ontology Consortium. This approach yields tangible benefits on many fronts. Students develop critical reading skills, get exposed to key concepts in genome biology and obtain essential hands-on experience on the use of ontologies and bioinformatics methods to assess orthology, synteny and other genetic features. Team-work in a competitive framework and intercampus rivalry motivate students to improve the accuracy of their annotations, leading to high-quality functional annotations on both reference and newly sequenced bacteriophage genomes.

P83 Effect of rumen content exchange on gene expression in rumen epithelium of lactating cows

Daniel Fischer, Natural Resources Institute Finland (Luke) FI
Ilma Tapio, Natural Resources Institute Finland (Luke) FI
Seppo Ahvenjärvi, Natural Resources Institute Finland (Luke) FI

Kevin J. Shingfield, Aberystwyth University UK
Johanna Vilkki, Finland Natural Resources Institute Finland (Luke) FI

The effect of rumen digesta exchange on gene expression in rumen papillae was analysed from an experiment involving a total rumen content exchange between 3 pairs of lactating cows fed the same diet. Papillae samples were sequenced for both mRNA and miRNA at three time points: at the exchange, and one or two weeks afterwards. Papillae samples were obtained during rumen evacuation from the ventral sites of rumen, immersed in liquid nitrogen and submitted for RNA extraction (AllPrep DNA/RNA/miRNA Universal Kit, Qiagen). Sequencing libraries were prepared according to Illumina TruSeq® Stranded mRNA and TruSeq® Small RNA sample preparation for mRNA and miRNA, respectively. Paired-end sequencing with 2 x 150 bp read length and the Illumina HiSeq 3000 platform was used for mRNA (average 50.6 M reads per sample) and single-read sequencing with 1 x 50 bp read length using Illumina HiSeq 2500 for miRNA (average 2.5 M reads per sample).

From mRNA libraries 67.8% of reads were mapped. About 53.5% of the mapped reads were located to known genes, the remaining reads mapping to unannotated regions of the bovine genome. Novel gene candidates were analysed with our in-house R-package hoardeR to identify potential orthologs. From miRNA libraries 36.3% reads were uniquely mapped. From 811 annotated miRNAs, 382 miRNAs were expressed.

Differential expression (DE) analysis provided a set of genes found to be differentially expressed before and after the rumen exchange, indicating that specific genes respond to changes in the rumen contents and microbial communities. A GSEA identified the affected pathways. From the DE analysis two groups that respond differently to the rumen exchange were identified. The biological significance of these two groups was further confirmed by linking with rumen metabolic data. miRNA expression patterns between the two groups were compared and correlated to the expression of their known target genes.

P84 Development of a novel, integrative proteogenomics approach to discover the entire protein-coding potential of prokaryotic genomes

Ulrich Omasits, Agroscope & Swiss Institute of Bioinformatics CH
Adithi R. Varadarajan, Agroscope & Swiss Institute of Bioinformatics CH
Christian H. Ahrens, Agroscope & Swiss Institute of Bioinformatics CH

Large advances in Next Generation Sequencing (NGS) technologies have led to an enormous increase of genome sequences. However, our ability to accurately and comprehensively predict all protein-coding open reading frames (ORFs), has lagged behind: large discrepancies exist for the total number of ORFs predicted and the precise protein start sites. Moreover, the recent identification of unannotated short proteins that carry out key biological functions has further

emphasized these shortcomings. Today, we thus do not know the entire protein-coding potential of any organism, not even the best-studied prokaryotes.

Yet, the knowledge of all protein-coding genes is one crucial aspect to fully capitalize on the genome information and decode its function. It is relevant for all levels, from small, focused experiments, to large functional screens, systems biology studies that depend on complete quantitative data series, up to accurate prediction of gene regulatory, interaction and metabolic networks.

To address this knowledge gap, we have developed a novel, integrative proteogenomics approach capable of discovering the entire protein-coding potential of prokaryotic genomes. Testing our approach in pilot studies including a complete expressed proteome dataset [1] has, in each case, identified unannotated ORFs, including short proteins, differentially regulated proteins, membrane-associated lipo-proteins and metabolic enzymes.

[1]. Omasits U, et al. Directed shotgun proteomics guided by saturated RNA-Seq

identifies a complete expressed prokaryotic proteome. *Genome Research*. 2013;23:1916-27.

P85 Genomic activity changes in canine macrophages after infection with Leishmania infantum and stimulation with a Toll like receptor-2 agonist.

Fabiana Mayer, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, ES
Sara Montserrat, Universitat Autonoma de Barcelona ES
Anna Castello, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB ES
Lorena Alborch, Universitat Autonoma de Barcelona ES
Simon Heath, Centre Nacional d'Anàlisi Genómica CNAG-CRG ES
Laia Solano-Gallego, Universitat Autonoma de Barcelona ES
Alex Clop, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB ES

Leishmaniosis is a common zoonotic disease in dogs and humans caused by Leishmania infantum. The TLR2 agonist, PAM3CSK4 promotes an inflammatory response and reduces the parasite load in macrophages. In order to understand the molecular changes occurring in relation to L. infantum infection and after stimulation with PAM3CSK4, we compared two epigenetic features of a canine macrophage cell line (DH82) prior to infection (CTROL), after infection with L. infantum promastigotes (INFEC), and after L. infantum infection and stimulation with the TLR2 agonist (TLR2+). We performed ChIP-seq to profile the genomic location of two histones: H3K4me3 and H3K27ac, which mark promoters and genomic activity, respectively. We generated between 8.4 and 10.4 million 75bp x paired-end reads per sample in an Illumina HiSeq 2000 system. After read mapping and using MACS2, we identified between 66,828 and 70,869 and between 83,997 and 98,411 ChIP-seq peaks for H3K4me3 and H3K27ac, respectively. Within each histone mark the mean size (H3K4me3: 523-535 bp; H3K27ac: 347-

379 bp) and fold-enrichment (H3K4me3: 11.4-11.7; H3K27ac: 5.9-6.3) was very similar between the three conditions. To identify treatment-specific peaks we used the differential peak caller ODIN. As expected, ODIN showed that H3K4me3 is more resilient to changes than H3K27ac. For example, we found a novel H3K27ac peak when compared INFEC vs CTROL in TGFB, a key immune gene that has been associated to susceptibility to this infection in animals. We also found a differential H3K27ac peak in SERPINB3 - a gene that is upregulated in Toxoplasma gondii infected macrophages and which has been shown to activate T-cells - in the TLR2 treated cells (TLR2+ vs INFEC). These epigenetic results can contribute understanding of the molecular changes underlying leishmanial infection and its treatment with a TLR2 agonist.

P86 Designing Splice Isoform-Specific Nucleic Acid Based Biosensor for Detection of Breast Cancer

Urmila Saxena, National Institute of Technology Warangal IN
Asim Bikas Das, National Institute of Technology Warangal IN

Breast cancer is the most common cancer among women with about 1.6 million new cases diagnosed every year. Advancement in cancer research and therapy has augmented the survival rate of patients but only when diagnosed in early stage. It has been shown that alternative splicing of many genes plays a major role in the development and progression of all cancers including breast cancer. Splice variants that are found predominantly in tumors show stage-specific variation in their expression levels. Therefore, they have clear diagnostic and prognostic value. Identification of the breast cancer-specific splice isoforms and studying their expression pattern from mRNA sequence data during the cancer progression may aid in identifying novel biomarkers for diagnosis of breast cancer at an early stage. We have selected certain candidate genes whose alternative splicing plays a major role in breast cancer development including HER2, Progesterone receptor, CASC4, p53, BRCA1, and PTEN. The existence of the splice isoforms of these genes was confirmed by analyzing the publicly available RNA sequence data from cancer patients. We have measured the prognostic values by Cox regression analysis, Kaplan-Meier survival plot with hazard ratio and logrank P value. Based on these results, we have also proposed a design of a nucleic acid based biosensor to detect alternative splice isoforms involved in breast cancer progression.

P87 RNA sequencing data analysis for cancer cell line authentication

Erik Fasterius, KTH Royal Institute of Technology SE
Pär Lundin, Science for Life Laboratory SE
Mathias Uhlén, KTH - Royal Institute of Technology SciLifeLab SE
Cristina Al-Khalili Szigyarto, KTH Royal Institute of Technology SE

Cell lines constitute key model systems within cancer research, but their use raises questions regarding genetic changes due to adaptation to environmental conditions and consequently genetic drift during culturing. Efforts have been made to reach standardized procedures for validation of cell lines, using techniques such as short tandem repeat profiling, single nucleotide polymorphism arrays, chromosomal karyotyping or the polymerase chain reaction. To facilitate cell authentication we used RNA sequencing data to investigate differences between HCT116 and HKe3, two isogenic cell lines used as a model for colorectal tumours. Transcriptome analysis performed using next generation sequencing technology yields in-depth information not only regarding RNA expression but also sequence alterations in expressed transcripts, making comparisons with SNP databases such as COSMIC feasible. In this study we highlight the use of RNA sequencing for cell line authentication and show that HKe3 is expressing the KRAS-G13D transcript, indicating that it is a dosage effect mutant rather than a true isogenic derivative of HCT116 as expected. The authentication procedures presented could be used to revisit the numerous cell line-based RNA sequencing experiments performed to date and facilitate the comparison of data from different experiments, platforms and laboratories.

P88 Using de novo Transcriptome Assembly to improve Genome Annotation and Sequence in Dictyostelium discoideum

Reema Singh, University of Dundee IR
Hajara Lawal, University of Dundee IR
Pauline Schaap, University of Dundee IR
Geoffrey Barton, University of Dundee IR
Christian Cole, University of Dundee IR

The reduced cost and complexity of high throughput sequencing has put whole genome sequencing (WGS) within the reach of smaller groups. There are many stages to a genome sequencing project before it is considered finished with gene annotation being a key consideration. There are several complementary strategies for achieving gene annotation each with its own strengths and weaknesses. De novo transcriptome assembly is a technique which takes short read RNA-seq data and assembles it to recreate the full-length transcripts produced within the cell. This is achieved without alignment to any existing genome reference sequence. The independence of de novo transcriptome assembly from the genome makes it a good match for confirming annotation as well as finding errors. We have taken RNA-seq data from a Dictyostelium discoideum knock-out experiment and used it to generate a de novo transcriptome with Trinity. Following refinement of the transcriptome with PASA 2 and quality assessment with CEGMA and Transrate we have updated 7,182 gene models, including 554 modified protein translations, and found 187 new alternatively spliced transcript isoforms. We identified 119 genes with tiny (<6bp) introns and highlight examples where errors in the genome sequence introduced these likely spurious introns into gene models. The de novo assembly data was used to correct the spurious gene models. In conclusion, no single method

for genome annotation is perfect, but we show that, by combining the orthogonal technique of de novo transcriptome to existing data assembly, errors are reduced and an improved annotation can be achieved.