ECCB 2014 Accepted Posters with Abstracts

B: Gene Expression

B01: Huy Tran, Samuel Oliveira, Olli Yli-Harja and Andre Ribeiro. Inducer intake kinetics under high extracellular concentrations in live Escherichia coli cells

Abstract: In Escherichia coli, in optimal conditions, many genes are kept repressed, only expressing upon the appearance of inducers in the environment. The response time to such appearances depends both on the intake kinetics of the inducer and on the dynamics of transcript production of the active gene.
From temporal, live imaging of cells capable of producing an RNA target for MS2-GFP, under the control of lac/ara-1 promoter, it is extracted the moment of appearance of the first tagged RNA in each cell, following the introduction of large amounts of IPTG in the media. Based on the empirical data, the kinetics parameters of the intake mechanism are inferred from a model fitted to the data. The model consists of a deterministic intake process coupled with a stochastic, multi-step transcription process, able to match the kinetics of RNA production of an active promoter. Next, from the inferred parameters' values, it is identified the dominant intake mechanism of IPTG in the range of concentrations studied.
We find that, at high extracellular inducer concentrations, the inferred kinetics of intake of the inducer molecules through the outer membrane is in agreement with the existence of a positive feedback mechanism. Also, we show evidence that the transport through the inner membrane is a rate-limiting step of the activation of the target gene, as it creates a delay in the appearance of the first target RNA following the introduction of inducers in the media. The methodology followed here should be of use in obtaining more accurate characterizations of the various intake mechanisms in E. coli.

B02: Olivier Rué, Philippe Bardou, Jérôme Mariette, Sarah Maman, Matthias Zytnicki and Christine Gaspin. sRNA-PlAn : a workflow for the sRNAseq data analysis

Abstract: 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 throuput 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 the 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 their structural and functional annotation challenging. Small RNA-seq data analysis tools such as miRDeep, miRanalyzer and others focus on microRNAs annotation and prediction, neglecting other types of RNAs. Recently, web tools such as DARIO, Ncpro enlarged the structural and functional annotation. We present sRNA-PlAn, a software under development which aims at profiling and annotating sRNAseq data, considering the whole set of ncRNA families in plants and animals and differential expression in multiple conditions and/or tissues.
After having cleaned reads and collapsed them, they are mapped against ncRNA databanks. Available ncRNA databanks are miRbase, eukaryotik-tRNAs, Silva and Rfam. If a reference genome is available, the cleaned sequences are mapped against the genome by using bwa or

bowtie2. Then, loci are constructed as an overlap of mapped reads, strand dependently. miRNAs and tRNAs can then be predicted. The tRNA prediction is performed by using tRNAScanSE, a software dedicated to the prediction of tRNA genes. The miRNA prediction is based on the alignment of a mature miRNA on a miRNA*. So, we try to match the most expressed read of a locus on a near localization on the genome. The best alignment is selected for the locus. Each candidate locus is scored with regard to miRNA characteristics, including structure and expression profile.
Visualization relies on RNAbrowse (NGSPipelines) interface which allows to explore results including profile expression of isoforms along the locus candidate, the secondary structure associated to locus prediction and the expression profile of a candidate locus by sample.
Short term perspectives include the prediction of other ncRNA families from sRNAseq data, the visualization of different statistics with regard to cleaning and annotation processes and the integration of the pipeline in the Galaxy environment.
This work was supported by "France Genomique PIA".

B03: Marion Verdenaud, Michel Degueldre, Sheila Zuniga, Juan Carlos Triviño, Loïc Quinton, Edwin De Pauw, Pierre Escoubas and Frédéric Ducancel. VENOMICS : High-throughput peptidomics and transcriptomics of animal venoms for discovery of novel therapeutic peptides

Abstract: Venomous animals have developed a vast arsenal of proteins and reticulated peptides used in defense and predation. Based on various disulfide-linked molecular scaffolds, they represent an enormous structural and pharmacological diversity. The spatial structures and the pharmacological spectrum of venom peptides are very diverse, making them top candidates as innovative drug leads. Previous studies have demonstrated the presence of several hundreds of peptides in the venom of single species of cone snails, scorpions and spiders. The global animal venom resource can be seen as a collection of more than 40,000,000 peptides and proteins of which only ~3000 are known.
Nevertheless the use of venoms for drug discovery is a rapidly emerging but still mostly unrealized prospective, due to several major hurdles including the availability of material, sample size and the complexity of venoms. The European project VENOMICS proposes a new paradigm to access venom diversity by combining transcriptomics and proteomics technologies to uncover peptide sequences. This first step subsequently allows, in vitro production of peptides, to generate libraries to be used in drug screening programs.
One of the key features of this project is the ability to predict the exact sequence of the mature active peptides synthesized by the venom gland cell machinery. The exact sequence of N-terminal peptide toxin end has to be determined as well as the possible occurrence of post-translational modifications.
In the proteomic part of the workflow, peptide analysis is carried out on crude venom milked from venomous animals. After HPLC fractionation of the venom, peptide disulfide bridges are reduced prior to mass spectrometry analyses using MALDI-Post Source Decay TOF MS/MS.
In the transcriptomic workflow, mRNAs are extracted from venom gland tissues and sequenced using an Illumina platform. Reads are qualified, cleaned using in-house scripts and are then assembled into contigs using Oases with several k-mer values and CAP3. Open Reading frames are predicted using framedp and getorf software. The general annotation process is achieved by combining Blast and InterProScan annotations. Special attention has applied to the annotation of "toxin-like" sequences and mature sequence prediction. For that purpose in-house scripts, known tools and curated databases are used to predict signal, propeptide and mature sequences.
The combined results of these two parallel strategies allows for determination of the exact

sequence of mature peptide and the occurrence of possible post-translational modifications using Peaks software and Perl in-house scripts. A subset if the newly identified toxins is then selected for in vitro production by either chemical synthesis or recombinant production. The library of peptides generated will then be evaluated against various different biological targets to identify potential therapeutic leads.

B04: Yoichi Takenaka, Shigeto Seno and Hideo Matsuda. Chronological analysis of regulatory strength on gene regulatory networks

Abstract: Gene regulatory networks describe interactions among genes that control the expression levels of mRNAs. The interactions have been revealed through many researches and the achievements are stored in databases nowadays. With the increasing of the number of known interactions, they have been used as the previous knowledge to reach deeper knowledge. Once the network is fixed, for example, some researchers require analyzing the time when each interaction works in their experimental processes.
Given a gene regulatory network and a time-course expression profile, we proposed a method to analyze the strength of the regulations in the network using scores used in Bayesian network. We define node scores and edge scores that represents the regulatory strength of a time point in given expression profile. The node score represents the strength that the expression level of the node are controlled by other nodes and edge score represents the strength that from-node controls the expression level of the to-node. Information amount of each node and edge is calculated from the expression profile without the target time point. Then, the deviation from the average of all the time point is defined as the proposed score. The time series variation of proposed scores shows the start and end of the regulations.
We used two data to reveal the effectiveness of the proposed method is shown. One is the gene expression profile from the cell differentiation to adipocyte and osteoblast of mouse. Another is the data of E. coli and Yeast from DREAM4 challenge. All the results give plausible changes of the gene regulations.

B05: Liberata Mwita and Oleg Reva. Comparison of gene expression profiles of two plant growth promoting strains Bacillus atrophaeus UCMB-5137 and Bacillus amyloliquefaciens FZB42 stimulated by maize root exudate

Abstract: Bacillus atrophaeus UCMB-5137 and Bacillus amyloliquefaciens FZB42 are Gram positive plant growth promoting rhizobacteria (PGPR). They both belong to Bacillus subtilis taxonomic group but differ by plant colonization behavior as it was observed by using luminescent microscopy. UCMB-5137 formed thick colonies on root surface while strains of B.amyloliquefaciens were prone to endophytic colonization. It was interesting to study whether the difference in behavior rely upon different genes activated at the time of root colonization. Gene expression stimulation by maize root exudate on B.amyloliquefaciens FZB42 has been studied recently and published by Fan et al. (2012). In the latter study Bam4kOLI microarray technology and EMMA 2.8.2 software were used to compare gene expression profiles in experiment versus control. In the current study on Bacillus atrophaeus UCMB-5137 a sequencing of the whole RNA by Illumina MiSeq 500 was used instead, followed by a statistical analysis using CLC Genomics workbench 7. We report here that RNA-Seq approach showed much higher sensitivity in recognizing differences in gene expression and allow obtaining a more general picture of expression of all genes of the organism rather than only a set of pre-selected genes. One discovery that probably had been missed in the microarray experiment was a complete silencing of majority phage and horizontally acquired selfish genes by the root exudate signals. Another discovery was that UCMB-5137 showed an alternative profile of up and down regulated core genes compared to

that in FZB42. In both cases significant up-regulation was observed for genes involved in carbon utilization, biofilm formation and transcription regulation. However, genes for motility, chemotaxis and synthesis of secondary metabolites including polypeptide antibiotics were up regulated in FZB42 but down-regulated in UCMB-5137. On the other hand genes responsible for stress response were up regulated in UCMB-5137 while none of them was mentioned in FZB42. These observations illustrate different gene regulation in PGPR bacteria exploiting different strategies for plant colonization.
Fan B, Carvalhais LC, Becker A, Fedoseyenko D, von Wirén N, Borriss R. Transcriptomic profiling of Bacillus amyloliquefaciens FZB42 in response to maize root exudates. BMC Microbiol. 2012 Jun 21; 12:116.

B06: Kiyohiko Sakamoto and Y-H. Taguchi. Subtype specific promoter methylation in glioblastoma

Abstract: Abstract: Glioblastoma is known to be the most lethal glioma and five year survival rate is only 10 %. Thus, it is urgent to identify critical genes for glioblastoma therapy. In this paper, we have analysed promoter methylation profiles downloaded from The Cancer Genome Atlas using recently proposed principal component analysis based unsupervised feature extraction. New set of genes with aberrant promoter methylation associated with previously identified subtypes was identified.
Introduction: Glioblastoma is the most lethal glioma wish very little five year survival rate. Thus, identification of critical genes for therapy purpose of glioblastoma is urgent. In this paper, we applied recently proposed principal component analysis (PCA) based unsupervised feature extraction (FE) to promoter methylation profiles of glioblastoma downloaded from The Cancer Genome Atlas (TCGA). We have identified genes with aberrant promoter methylation associated with previously identified subtypes.
Results: By applying PCA based unsupervised FE, we have identified that the second principal component (PC2) exhibits distinction between subtypes. Fig. 1 shows the contribution to PC2 from samples (see Table 1 for P-values), but will not be shown here. Since PC2 turned out to exhibit distinction between samples, we have selected probes with larger absolute PC2 scores. Fig. 2 shows promoter methylation profiles of top ranked 12 outlier probes along PC2 in negative or positive direction, but will not be shown here. For most probes, only Proneural subtypes are significantly larger or less than other subtypes. Thus, PC2 possibly indicates the aberrant hyper/hypomethylation associated with Proneural subtype. Although Verhaak et al identified subtype with gene expression, in this study proneual subtype has also associated aberrant promoter meythylation.
Next, in order to confirm if genes associated with identified probes are related to cancers, we have used gendoo and DisGeNet (see Table 2, but will not be shown here.). Among 42 genes listed, 31 genes (74 %) were reported to be related to cancer related genes by either gendoo or DisGeNet. Two genes (ERBB2 and LRRC4) were reported to be related to glioblastoma, additional five genes (TTC12, LGALS3, WFDC2, BCAT2, ANK3) were related to other neuronal tumor. Thus, PCA based unsupervised FE seemed to work pretty well.
Conclusion: In this paper, PCA based unsupervised FE was applied to promoter methylation of glioblastoma. Many subtype specific aberrant promoter methylayion was identified.
Acknowledgments This study was supported by KAKENHI 23300357 and 26120528 and Chuo University Joint Research Grant.

B07: Mireya Plass, Simon H. Rasmussen, Lykke Pedersen and Anders Krogh. Computational analysis of RNA binding proteins in miRNA-mediated downregulation

Abstract: microRNAs (miRNAs) are endogenous short non-coding RNAs (18-22nt long) involved in the regulation of gene expression at the post-transcriptional level. To perform their regulatory function, miRNAs interact with AGO proteins to direct their binding on target mRNAs, promoting mRNA degradation and translation repression. Previous studies have reported that the efficacy of miRNA-mediated downregulation depends on the strength of miRNA binding on the target mRNA. Therefore, their function can be regulated by RNA binding proteins (RBPs) and RNA structures that enhance or block miRNA-mRNA interactions.
In this study, we performed a genome-wide analysis of the role of RBPs in miRNA-mediated gene regulation. We combined different types of high throughput data (RNA-seq and PAR-CLIP) to calculate enrichment values of RBPs in 3'UTRs and study their relation with miRNA-mediated downregulation. We found that the ability of miRNA targets to downregulate an mRNA depended on their ability to recruit AGO proteins. Furthermore, we noticed that other RBPs modulated this process. On the one hand, we observed that RBPs blocked miRNA regulation by direct competition with AGO2 for binding on target sites. On the other hand, we also noted that RBP binding on other regions of the 3'UTRs could enhance AGO2 binding on target sites. In summary, our results suggest that RBPs perform a crucial function for understanding miRNA-mediated regulation.

B08: Julien Roux, Irene Hernando-Herraez, Claudia Chavarria, Amy Mitrano, Jonathan Pritchard, Tomas Marques-Bonet and Yoav Gilad. A genomic study of the contribution of DNA methylation to regulatory evolution in primates

Abstract: A long-standing hypothesis is that changes in gene regulation play an important role in adaptive evolution, notably in primates. Yet, in spite of the evidence accumulated in the past decade that regulatory changes contribute to many species-specific adaptations, we still know remarkably little about the mechanisms of regulatory evolution. In this study we focused on DNA methylation, an epigenetic mechanism whose contribution to the evolution of gene expression remains unclear.
To interrogate the methylation status of the vast majority of cytosines in the genome, we performed whole-genome bisulfite conversion followed by high-throughput sequencing across 4 tissues (heart, kidney, liver and lung) in 3 primate species (human, chimpanzee and macaque). Because the 4 tissues are from the same individuals, we are able to monitor methylation differences between individuals, tissues and species. In parallel, we collected gene expression profiles using RNA-seq from the same tissue samples, allowing us to perform a high resolution scan for genes and pathways whose regulation evolved under natural selection.
We integrated these datasets to characterize better the genome features whose methylation status lead to expression changes, and we developed a statistical model to quantify the proportion of variation in gene expression levels across tissues and species which can be explained by changes in methylation. Globally, our study leads to a better understanding of the molecular basis for regulatory changes and adaptations in primates.

B09: Rim Zaag, Guillem Rigaill, Jean-Philippe Tamby, Véronique Brunaud, Zakia Tariq, Sébastien Aubourg, Etienne Delannoy and Marie-Laure Martin-Magniette. Global Analysis of coRegulation for the identification of functional modules

Abstract: One of the challenges faced by genomics currently is the understanding of gene function. Genome wide analysis of gene function mostly relies on guilt by association approaches through coexpression analysis taking advantage from the availability of transcriptome data. Indeed, cluster analysis of gene-expression profiles can be used to propose

functions based on the assumption that coexpressed genes have likely related biological functions (Eisen et al., 1998). Generally co-expression is performed by analyzing correlations between all the gene pairs from multiple microarray experiments collected from international repositories. Such approach has two drawbacks: First it leads to a local point of view about functional modules and second the dataset is composed of heterogeneous transcriptome results.
In constrast, we performed a global analysis of highly homogeneous transcriptome data extracted from CATdb (Gagnot et al., 2008). The whole dataset is composed of more than 18 000 genes described by 424 expression differences dealing with stress conditions. The coexpression analysis is performed through a model-based clustering method which allows the modelisation of the whole dataset by a mixture of distributions. A study of the Bayesian Information Criterion (Schwarz, 1978) as a function of the component number allows the evaluation of the fit between the data and the mixture. Once the assessment is done, the selected mixture is the one with the highest value of BIC.
Without a priori knowledge, the model has guided us to divide the whole dataset in twenty types of stresses leading to the identification of gene clusters having the same pattern of response under a single stress type. However coexpressed genes are not necessarily coregulated and then are less likely to be functional partners. To find groups of coregulated genes, we integrated these coexpression studies by calculating the occurrence number in a same cluster for each gene pair. Some pairs have a coordinated transcriptional response in up to 15 different types of stress and a resampling procedure showed that a gene pair observed in the same cluster in more than 4 stresses is significant. This approach allows us to focus our study on the potential key players of stress responses. Furthermore, the resulting coregulated gene network reveals an interesting topology of highly connected substructures. Preliminary analyses of these components containing orphan genes showed that they are more homogeneous than coexpression clusters highlighting probable functional modules.

B10: Valentin Voillet, Magali San Cristobal, Pascal G.P. Martin, Yannick Lippi, Louis Lefaucheur and Laurence Liaubet. Integrative approach to define biomarkers of piglet maturity

Abstract: In pigs, the perinatal period is the most critical time for survival. Piglet maturation, which occurs at the end of gestation, leads to a state of full development after birth. Therefore, maturity is an important determinant of early survival. Postnatal mortality is not an issue only in pigs but also affects other mammals like sheep or humans. The objective of our project is an integrated multi-omics approach (transcriptomics, proteomics, phenotypes...) with focus on muscle metabolism, because of its key role in adaptation to extra-uterine life, e.g. glycogen storage and thermoregulation. This involves substantial challenges due to the high-dimensionality of the data.
Progeny from two extreme purebreds (Large White and Meishan) for maturity were used. The Large White (LW) breed is a selected breed with an increased rate of mortality at birth, whereas the Meishan (MS) breed produces piglets with extremely low mortality at birth. Maturity of several tissues (e.g. muscle, liver and blood) is analyzed on the progeny from these two breeds (LW, MS and reciprocal F1) at two points during end of gestation (gestational days 90 and 110). Here, we focus on the integration of transcriptomics data of several tissues (muscle, liver and blood) to define biomarkers of piglet maturity.
First, muscle microarray data was analyzed to identify genes and biological processes involved in piglet muscle maturity. A specific pipeline was developed to operate with a high number of differentially expressed genes. Using functional analysis and relevance network, some key biological processes and crucial genes explaining the biological muscle maturity difference between the two extreme purebred piglets were highlighted. In particular, key

genes for gluconeogenesis/glycolysis are up-regulated in MS. These genes could be excellent candidates for a key role in the maturity.
After that, some integration statistical methods, e.g. sparse Partial Least Square (sPLS), network mining (Gaussian Graphical Model (GGM)) or multiple factors analysis (MFA), were performed between the three tissues: muscle, liver and blood. sPLS provides an extremely useful tool for the biologist in need of integrating two-block data sets and easily interpreting the resulting variable selections. Blood was used because of its strength of potential biomarkers, whereas liver and muscle were studied because of their known importance in the gluconeogenesis/glycolysis pathway. As expected, GGM networks showed that gene expression for the genes involved in gluconeogenesis/glycolysis was more shared between muscle and liver than between muscle and blood. sPLS selected few relevant genes linked between blood and muscle; these genes expressed in blood could be good biomarkers for piglet maturity.

B11: Nicola Voyle, Aoife Keohane, Stephen Newhouse, Katie Lunnon, Andy Simmons, Eric Westman, Hilkka Soininen, Iwona Kloszewska, Patrizia Meccoci, Magda Tsolaki, Bruno Vellas, Simon Lovestone, Angela Hodges, Richard Dobson and Steven Kiddle. Blood based gene expression markers of Alzheimer's Disease diagnosis: a pathway based approach.

Abstract: Background: Methods of diagnosing Alzheimer's Disease (AD) are invasive and expensive while treatments of AD only provide symptomatic relief. Discovery of a blood-based biomarker of disease would reduce the number of patients subject to such diagnostics and increase the likelihood of finding an effective treatment.
Gene expression levels in blood are potential markers of AD [1]. However, limited findings are replicated in independent datasets, possibly due to differences in platforms, technologies and study design. We hypothesized that considering changes in gene expression at the pathway level may create a more robust model.
Materials and Methods: 201 samples from the AddNeuroMed study were used to build and test a clinical diagnostic model, based on gene expression data, age, gender and ApoE4 genotype [2] . Genes were then grouped into pathways and scored using Gene Set Variation Analysis, to assign a pathway expression measure at the sample level [3]. The models were built in training data and tested on held out test data using Random Forests [4]. The model was compared with that built using age, gender and ApoE4 genotype alone.
The diagnostic model was assessed for stability in an independent dataset of 173 samples.
Results: In test data the model showed an accuracy of 0.60 using pathway predictors in comparison to 0.71 for gene level data. The model built using age, gender and ApoE4 genotype alone showed an accuracy of 0.49. Using a pathway approach enabled us to easily test the stability of the classifier in the independent dataset generated using a different version of the array.
Conclusions: A pathway level analysis of gene expression enables application of predictive models across different datasets, produced using different platforms, with ease. It also reduces the dimensionality of these very large datasets. We think the use of a pathway based approach will enable the development of more robust predictive models of Alzheimer's Disease diagnosis.
References:
[1] Han, G., Wang, J., Zeng, F. et al. Characteristic Transformation of Blood Transcriptome in Alzheimer's Disease Journal of Alzheimer's Disease 35 (2013) 373–386.
[2] Biomarkers in Brain Disease: Ann. N.Y. Acad. Sci. 1180: 36–46 (2009).
[3] Hänzelmann, S., Castelo, R. and Guinney, A. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics, 14:7, 2013.

[4] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18-22.

B12: Bart Cuypers, Manu Vanaerschot, Maya Berg, Pieter Meysman and Kris Laukens. Understanding Leishmania development and adaptation by using a systems biology approach

Abstract: Leishmania donovani is the causative agent of visceral leishmaniasis (VL) in the Indian subcontinent. Each year 500.000 people contract VL, which is lethal without treatment. With only four drugs available and rapidly emerging drug resistance, fundamental knowledge about the parasite is essential to boost the development of new drugs. Leishmania, however, is a challenging biological system. Little is known about its gene regulation and the few findings indicate major differences to known gene expression systems.
Indeed, no polymerase II promotors have ever been found in Leishmania1. Genes are constitutively transcribed in large polycistronic units and subsequently spliced into individual mRNAs (trans-splicing) (1). Gene expression is thus assumed to be regulated at the post-transcriptional level (mRNA stability, translation efficiency, etc) but evidence to support this is scarce1. Nevertheless, it is known that gene expression of adaptive genes can be increased over several generations by two mechanisms (2) 1) chromosome polyploidy of the complete chromosome 2) formation of circular or linear episomes (=Eukaryotic Plasmids).
To shed light on these mechanisms, we conducted for the first time a full systems biology study on this parasite. Genomic, transcriptomic, proteomic and metabolomic data was collected and added to a relational database, which was further complemented with publicly available data. The eventual end goal will be the construction of a network model that integrates all these different layers of information (gene, transcript, protein, metabolite) to gain new fundamental biological insight into the development and drug adaptation of Leishmania.
While this project has only just recently been initiated and most of the results are still pending, we could, however, already observe a fixed set of 5' UTR lengths for every mRNA, which is conserved in different strains, and that the usage of the 5' UTR is predominantly dependent on the conditions. We are currently investigating if there is a relation of 5'-UTR length with mRNA stability and/or translation efficiently and how to implement this in the compendium.
Apart from answering some of our research questions, the database compendium will be a treasure of organized information for every Leishmania researcher.
References
1. Donelson, J. (1999) PNAS. 96, 2579–258.
2.Downing, T. et al. (2011) Genome Res. 21, 2143–56.

B13: Pieter Meysman, Ehsan Sabaghian, Riet De Smet, Kristof Engelen, Yves Van de Peer, Yvan Saeys and Kris Laukens. A detailed comparison of seven prokaryotic transcriptional regulatory networks from an evolutionary perspective

Abstract: The regulation of gene expression is an important biological process that allows organisms to survive in the ever-changing environment. The interaction between genes and their regulators is frequently conceptualized as a network. Throughout evolution this regulatory network has undergone both minor alterations and massive restructuring events as species diverged and adapted to new conditions.
To further advance our understanding of prokaryotic gene expression, we have compiled expression compendia for seven model organisms. These compendia can be accessed through the Colombos web service (www.colombos.net). In a proof of concept, we have demonstrated that such expression compendia can be used to explore the difference in regulatory programs

between Escherichia coli and Salmonella Typhimurium that support their different life styles. In this study however we will focus on the characteristics and diversity of the transcriptional networks themselves.
Unfortunately for most prokaryotic species there is no reliable transcriptional network available. To circumvent this issue, we have predicted the regulatory networks based on the expression compendia for all seven species contained within Colombos using the GENIE3 inference algorithm. Comparison to those experimental transcriptional networks that do exist for some of these organisms, reveal that while we obtain a low recall (1% to 10%) for our predicted networks, the sensitivity is high (6% to 30%).
Orthology mapping based on protein sequence similarity and clustering with OrthoMCL was used to link the different transcriptional networks. This revealed that the edge overlap between the networks closely followed the known phylogenetic tree for these organisms, except for Helicobacter pylori. A more detailed analysis showed that while very little of the network is present across all the studied species, there are still some regulatory interactions that are conserved in most. These conserved interactions mostly involve DNA damage response, regulation of ribosomal proteins, nitrogen metabolism and stress regulation. Further there seem to be several global regulators that are present in a large fraction of the species and share very similar regulatory targets.
The goal of this study was not only to characterize the rewiring in the transcriptional regulatory network that has occurred between these species but also to evaluate if such a study is even possible based on predicted networks. This is further complicated by the fact that some of these organisms are only very distantly related and one would have to go back almost two billion years to find their common ancestors. However the biological relevance and significance of the predicted conserved interactions does seem to validate this kind of approach.

B14: Emmanuel Chaplais, Alice Talpin, Félicie Costantino, Clémence Desjardin, Nelly Bonilla, Ariane Leboime, Roula Said-Nahal, Franck Letourneur, Jacques Sébastien, Gilles Chiocchia, Maxime Breban and Henri-Jean Garchon. Multiway transcriptome analysis discriminates effects of disease and of HLA-B27 in Spondyloarthritis

Abstract: Objectives: Spondyloarthritis (SpA) is a prevalent chronic inflammatory rheumatism that affects axial skeleton, peripheral joints and may result in immobility of the spine and sacroiliac joints. Extra-articular features including uveitis, psoriasis and inflammatory bowel disease are frequent. Current SpA treatments are only symptomatic, relieving inflammatory symptoms. SpA etiology is largely multifactorial with a genetic component dominated by the long-known strong association with the HLA-B27 allele. This allele, however, is not sufficient for the disease to occur. Whereas dendritic cells are believed to play a key role in SpA pathogenesis, the precise mechanisms underlying disease development and particularly the role of HLA-B27 remain poorly understood. To shed light on the genes involved, we carried out a transcriptome analysis of dendritic cells from patients and healthy controls, also accounting for HLA-B27.
Methods: Transcriptomic profiles of monocyte-derived dendritic cells (MD-DCs) were obtained from 23 HLA-B27+ SpA patients, and from 44 controls (23 HLA-B27+ and 21 HLA-B27-). MD-DCs were stimulated or not with endotoxin for 6 or 24 hrs. We used the Affymetrix Human Gene 1.0 ST platform. Analysis of differentially expressed (DE) genes was conducted with LIMMA accounting for both the disease and the HLA-B27 status. The reliability of the linear model was confirmed by bootstrapping. Empirical p-values were also computed for each gene by status permutation and used to further filter LIMMA gene lists at the threshold of 5%. Gene set enrichment analyses were conducted with quSAGE using all MSigDB collections (10 295 genes sets).

Results: We performed three comparisons to identify DE genes related to HLA-B27 or SpA status. We thus generated three lists: A, including 794 DE genes between HLA-B27+ SpA patients and HLA-B27- healthy controls; B, including 675 DE genes between HLA-B27+ controls and HLA-B27- controls; and C, including 464 DE genes between HLA-B27+ patients vs HLA-B27+ controls. Subtracting A–B left 653 genes, of which 63 are in list C, thus yielding a robust list of genes affected by SpA and filtering out irrelevant genes affected by HLA-B27.
Running quSAGE revealed that DE genes associated with SpA are mainly involved in metabolic processes, immune system and cellular process. There is however substantial redundancy of genes across gene sets, as shown by a network analysis of genes shared by the genes sets involved. We are now investigating how to exploit this redundancy to increase the power of our analysis.
Conclusion: Our study demonstrates for the first time that HLA-B27, the primary genetic factor in SpA, appears to alter expression of an unexpectedly large number of genes in a manner unrelated to the disease. Accounting for this observation will be therefore critical to identify the genes whose expression in MD-DCs is more specifically altered in relation to the disease process.

B15: Alexandra Popa and Rainer Waldmann. Genome-wide characterization of translation by ribosome profiling – bioinformatics challenges

Abstract: Open reading frames are frequently found outside of known coding sequences (i.e. long non coding RNAs, untranslated regions) raising the question of whether those ORFs are actively translated. An in silico definition of the protein coding potential of genes is just insufficient since the features that distinguish a translated from an untranslated ORF are not always defined. A recently developed high throughput sequencing technique, ribosome profiling was a big step towards a genome-wide map of translation. Ribosome profiling takes advantage of the fact that ribosomes protect their associated RNA fragments against RNAse digestion. In consequence, most of the RNAse-resistant signal originates from ribosome protected, translated sequences. Yet, although most of the signal is translation related, there is a significant amount of noise, since any other RNA binding proteins or RNA secondary structure can protect RNA against RNAse-digestion.
Different bioinformatics ribosome profiling analysis pipelines have been previously used on the same dataset to distinguish coding from noncoding sequences. Surprisingly both approaches led to quite opposite conclusions regarding the translation of long noncoding RNAs and 5' UTRs [Ingolia NT et al., Nat Protoc. 2012 Jul 26; Guttman M et al., Cell. 2013 Jul 3].
We show here that integration of ribosome profiling data generated after treatments with several different translation inhibitors leads to a much more reliable definition of translated sequences. We discuss the bioinformatics challenges and describe several possible solutions that allow a reliable definition of bona fide coding sequences.
This work was supported by Fondation pour la Recherche Médicale (DEQ20130326464), Agence Nationale pour la Recherche (ANR-10-INBS-09-03), Canceropole PACA (2013-09, 2014-07), and Fondation ARC (SFI20121205973).

B16: Axel Rasche, Matthias Lienhard and Ralf Herwig. ARH/ARH-Seq: Discovery tool for differential splicing in High-throughput data

Abstract: Alternative splicing (AS) is a key mechanism for generating the complex proteome of an organism. AS has been observed within a variety of biological conditions, for example, in tissue expression, with respect to human diseases and in protein modification. The

computational prediction of alternative splicing from high-throughput data is inherently difficult and necessitates robust statistical measures because the differential splicing signal is overlaid by influencing factors such as gene expression differences and simultaneous expression of multiple isoforms amongst others. We propose ARH, a discovery tool for differential splicing in case control studies that is based on the information-theoretic concept of entropy. ARH-seq works on high-throughput sequencing data and is an extension of the ARH method that was originally developed for exon microarrays. We show that the method has inherent features, such as independence of transcript exon number and independence of differential expression, what makes it particularly suited for detecting alternative splicing events from sequencing data. In order to test and validate our workflow we challenged it with publicly available sequencing data derived from human tissues and conducted a comparison with eight alternative computational methods. In order to judge the performance of the different methods we constructed a benchmark data set of true positive splicing events across different tissues agglomerated from public databases and show that ARH is an accurate, computationally fast and high-performing method for detecting differential splicing events.

B17: Claudia Coronnello, Giovanni Perconti, Patrizia Rubino, Flavia Contino, Serena Bivona, Salvatore Feo and Agata Giallongo. Statistical Validation of a Comprehensive Gene/miRNA Expression Profile Dataset for miRNA:mRNA Interaction Analysis

Abstract: MicroRNAs (miRNAs) are small non-coding RNA molecules mediating the translational repression and degradation of target mRNAs in the cell. Mature miRNAs are used as a template by the RNA-induced silencing complex (RISC) to recognize the complementary mRNAs to be regulated. Up to 60% of human genes are putative targets of one or more miRNAs. Several prediction tools are available to suggest the miRNA targets , however, only a small part of them has been validated by experimental approaches. In addition, none of these tools does take into account the network structure of miRNA:mRNA interactions, which we believe is crucial to efficiently predict the miRNA regulation effects in a specific cellular context.
We aim to model the miRNA:mRNA interaction network, by including all the miRNAs and mRNAs endogenously expressed in any cellular condition. We started by using as test bed the breast cancer MCF-7 cells. In order to build the miRNA:mRNA interaction model, we collected several miRNA and mRNA expression profiles, by using the Agilent microarray platforms . We analyzed samples derived from the immunoprecipitation (IP) of two RISC proteins, AGO2 and GW182. Specifically, we considered the input, the IP and the flow through samples. We also collected and analyzed miRNAs and mRNAs from polysomal/non polysomal fractions separated through sucrose gradient, as completion of a dataset useful to investigate on miRNA function. The expression level of the top expressed miRNAs has been validated by real time PCR.
Due to the peculiarities of our dataset, we used non-standard bioinformatics techniques to preprocess and analyze the obtained expression profiles. As result, we validated the sample extraction techniques (both RISC proteins IP and polysomes isolation), by obtaining expression profile clustering and regression results consistent with the experimental design. Our dataset can then be used to further investigate on miRNA:mRNA interactions, and here we also show our preliminary results in this direction.

B18: Tom Lesluyes, Gaëlle Pérot, Marine R. Largeau, Céline Brulard, Pauline Lagarde, Jean-Michel Coindre, Agnès Neuville, Carlo Lucchesi and Frédéric Chibon. From laboratory bench to patient, micro-arrays to NGS: clinical transfer of a gene expression signature.

Abstract: Soft tissue sarcomas are a heterogeneous group of tumors with a wide spectrum of clinical behavior. Histological type and grade are used to determine therapeutic management of patients.
Histological grade defined by the French FNCLCC group is the current international standard but suffers from two major limitations: it is not informative in 40% of cases, which are classified as intermediate grade, and it is difficult to apply to microbiopsies, which are the commonest type of diagnostic sample taken before surgery. To overcome these problems, our group identified, validated and patented a prognostic expression signature (CINSARC) that is more effective than FNCLCC grade, but whose clinical applicability is limited by the need to analyze RNA from frozen samples on micro-arrays.
We have initiated a technology transfer program (Aquitaine Science Transfert: SATT Aquitaine) to test the existing micro-array signature on RNA-seq (Next Generation Sequencing) expressions. The objective is to analyze RNA extracted from formalin-fixed paraffin-embedded (FFPE) tumors, since this is the material routinely used by pathologists for diagnostic and prognostic examination.
We performed Illumina sequencing on a set of 100 frozen tumors to compare the performance of RNA-seq and micro-arrays. We then sequenced a set of 40 sarcomas with paired frozen and FFPE samples to make sure that gene expression profiles were similar, regardless of the type of material used. The analysis pipeline we have developed has four steps: pre-alignment (Sickle, SeqPrep, FastQC), alignment (TopHat2), post-alignment (SAMtools, MarkDuplicates) and gene expression (Cufflinks, HTSeq-count).
We present our results, showing that 1) the CINSARC signature correctly classifies sarcoma grade on RNA-seq and 2) FFPE expressions are similar to frozen ones. We also present an experimental protocol suitable for transcriptomic RNA-seq analysis of clinical samples.

B19: Metsada Pasmanik-Chor, Shay Ben Shachar, Henit Yanai, Liran Baram, Hofit Elad, Amos Ofer, Eli Brazowski, Noam Shomron, Hagit Tulchinsky and Iris Dotan. Gene and microRNA expression as tools to infer spectrums of Inflammatory Bowel Diseases

Abstract: Pouchitis is inflammation of the previously normal small bowel reservoir (ileal pouch) which may develop in ulcerative colitis (UC) patients undergoing large bowel resection and pouch surgery. We aimed to characterize pouch disease behavior using a molecular approach. UC pouch patients were prospectively stratified according to disease behavior into normal pouch (NP), chronic pouchitis (CP), and Crohn's-like disease of the pouch (CLDP) groups. These were compared to Crohn's disease (CD). Gene expression analysis of intestinal mucosal biopsies was performed using Affymetrix microarrays, in sixty six subjects, validated by real-time PCR. MicroRNA expression was performed by Illumina miR-Seq. Gene ontology was studied using Bioinformatics tools. While in UC ileum there were no significant gene or microRNA expression alterations, NP patients had 168 differentially-expressed genes (fold change $\geq|2|$, corrected p value ≤ 0.05). In CP and CLDP 490 and 1152 gene expression alterations were detected, respectively. Gene expression and microRNA profiles reflected disease behavior. CD ileitis had 358 alterations, with a 96% overlap with the various pouch groups. Gene ontology analyses revealed multiple biological processes associated with pouch inflammation, including response to chemical stimulus, small molecule metabolic and immune system processes and specific infectious-related pathways such as staphylococcus aureus, leishmaniasis and tuberculosis. There were 190 genes with significant negative correlation with 63 microRNAs. Interestingly, only 6% of these genes were up regulated, while the majority were down-regulated. Gene and microRNA alterations in pouch inflammation and CD overlap, suggesting that IBD is a spectrum, rather than distinct diseases. Altogether, our work shows that, gene and microRNA expression patterns could be used to characterize IBD subgroups.

The authors would like to thank Dr. Varda Oron-Karni for the most professional performance of microarrays. This study was partially supported by a generous grant from the Leona M. and Harry B. Helmsley Charitable Trust.

B20: Laurence Josset, Lisa Gralinski, Amie Eisfeld, Ralph Baric, Yoshihiro Kawaoka and Michael Katze. Analysis of a cellular gene response network to SARS-CoV and influenza A virus infection identifies specific virus-host dynamics.

Abstract: Influenza A virus (IAV) and SARS-coronavirus (SARS-CoV) are highly transmissible respiratory viruses, which both can cause death by acute respiratory distress syndrome in humans. Despite the similar outcome and tropism, they have important differences in structure, epidemiology characteristics and dynamics of infection.
To better define determinants of the host response to these viruses, we performed a systematic analysis on a compendium of more than 500 transcriptomic profiles, which included microarrays from Calu-3 cells infected with IAV [the highly pathogenic H5N1 avian influenza A/Viet Nam/1203/2004 (VN1203), several VN1203 mutants, and two strains of 2009 H1N1 virus] or SARS-CoV [the lethal mouse-adapted strain (MA15), urbani infectious clone or several mutants].
We used the maximal information coefficient (MIC) to explore relationships between viral replication and gene-expression profiles. Regulatory relationships between genes responding to either SARS-CoV and/or IAV replication were inferred by gaussian graphical modeling. Topological analysis of the host-response network (HRN) revealed that both common and virus-specific gene co-expression modules were activated in response infection. Common modules were enriched in genes coding for innate viral sensors or in various metabolic processes. A highly SARS-CoV specific module contained genes related to fatty acid biosynthesis, while a highly IAV specific module was enriched in genes involved in positive regulation of inflammatory response. Interestingly, the HRN was induced with different dynamics following infection with SARS-CoV or IAV. Module analysis also highlighted specific effects of viral factor mutations. Finally, key points of the HRN were defined as bottleneck and intramodular hubs and could represent targets for antiviral therapy targeting specifically IAV or SARS-CoV, or both respiratory viruses.

B21: Mitra Barzine and Alvis Brazma. Integration of independent RNAseq datasets

Abstract: In the last few years, many gene expression studies have been performed and released. These studies often have overlaps across some of their variables. A method that allows the integration of gene expression levels across different RNAseq experiments and conditions would be a useful asset; either as a baseline expression reference or as a way to find new gene expression correlations. Arguably, if the integration is carried out directly on the data (as opposed to meta-analyses) weaker effects will be also detected. Some resources already attempt to achieve this for public microarray datasets (for instance, combining experimental data obtained on a particular microarray platform).
Our study aims to understand to what extent, for a given condition, different RNAseq datasets could be used to infer gene expression information, regardless of library preparation and sequencing platform. This task is challenging because RNAseq datasets generated in different labs or at different times are not directly comparable.
We have focused our analyses on human data. Four different datasets, which represent (at least) four common conditions, have been used for this study. Different approaches were used to compare the datasets, such as correlations between samples of different datasets – either as a whole or on particular subsets (e.g. most expressed genes, more variant genes, etc).
Preliminary results are encouraging because the same tissues in different datasets present the

same gene expression profiles globally. However, while de novo discoveries or differential expression studies are mildly affected by normalization, in the case of integrating several datasets together, normalization becomes one of the main issues. Figuring out the genes whose expression is more sensitive to library preparation than to biological conditions is another challenge.

B22: Emilie Chautard, Clara Benoit-Pilven, Vincent Lacroix and Didier Auboeuf. Development of a new bioinformatics pipeline to annotate, quantify and visualize alternative splicing events from human and mouse RNA-Seq data

Abstract: Recent genome-wide analyses reveal that more than 90% of all human genes produce at least two transcripts through alternative splicing. We have previously shown than epithelial, endothelial and fibroblast cells exhibit specific splicing programs independently of their tissues of origin (Mallinjoud et al. 2014). Our aim is now to determine on a large number of normal and tumor cell types how splicing programs are co-regulated and which splicing events are involved, which may participate in explaining their phenotypes.
To meet our objectives, we developed a bioinformatics pipeline to annotate, quantify and visualize alternative splicing events (ASEs) from RNA-Seq data. Our pipeline is able to handle both short reads and long reads of variable lengths. The input is a read alignment file in bam format, which can be produced from various spliced-aware aligners (e.g.: TopHat, STAR, GMAP). In parallel, we will analyze the same datasets with KisSplice (Sacomoto et al. 2012), a local transcriptome assembler developed in the LBBE team, that has been shown to produce complementary results to the mapping approach.
The annotation step of our pipeline is using by default FasterDB as a reference, a web interface to a database that describes known splice variants of human and mouse protein coding genes. However, our pipeline is also able to identify new unannotated events, non-protein coding ASEs, or work with other external annotations databases (e.g. EnsEMBL, UCSC). New ASEs annotated using our pipeline and KisSplice will be progressively added to FasterDB in order to enrich its annotations.
The quantification step enumerates read counts mapping to exon-exon junctions corresponding to the known and newly annotated ASE events. Different statistical models are then applied to detect ASEs depending on the type of event (e.g.: exon cassette, 3' or 5' alternative sites, mutually exclusive exons) and on the experimental design (e.g.: replicate number, paired or unpaired data). Significant ASEs are then displayed in tables in a comprehensive manner for biologists. We also developed a new visualization mode in FasterDB to help interpreting the results and selecting the best candidates for validation.
Our pipeline has already been challenged against various types of samples, including several datasets already analyzed and validated extensively by our group at CRCL using exon arrays.
Mallinjoud P, Villemin JP, Mortada H, Polay Espinoza M, Desmet FO, Samaan S, Chautard E, Tranchevent LC, Auboeuf D. (2014) Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. Genome Res. 24:511-521.
Sacomoto GA, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot MF, Peterlongo P, Lacroix V. (2012) KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics. 13 Suppl 6:S5.
Acknowledgments: Plan Cancer Inserm 2009-2013, ARC SANTE 1

B23: Liliana Greger and Alvis Brazma. Characterizing transcriptome diversity by RNA chimeras across human populations

Abstract: Fusion genes originated from exons of two or three different genes are known to play a major role in cancer development, where they are mostly result from chromosomal

rearrangements. However, chimeric transcripts formed by post-transcriptional events are also present in normal tissues, where they may increase the protein complexity and contribute to the human evolution. The exact mechanism of the RNA chimeras formation is still not known. Here, we used RNA sequencing data from 462 healthy individuals from the 1000 genomes project to systematically find and characterize 81 RNA tandem chimeric genes, 13 of which were novel. In addition, we discovered that prevalence of long introns at the fusion breakpoint is associated with RNA chimeras formation. Furthermore, we observed that the RNA chimeras are lowly expressed genes. Finally, by combining our results with genomic data from the same individuals we associated the mechanism of RNA chimeras formation to specific intronic genetic variants.

B24: Julie Aubert, Christelle Hennequet-Antier, Cyprien Guérin, Delphine Labourdette, Anne de La Foye, Nathalie Marsaud, Fabrice Legeai, Frédérique Hilliou and Brigitte Schaeffer. How to design a good RNA-Seq experiment in an interdisciplinary context ?

Abstract: The development of high-throughput sequencing has revolutionized the study of genomics and molecular biology so that a lot of various biological questions may now be addressed. Among these technologies RNA-seq is a powerful tool for characterizing and quantifying transcriptome. Upstream careful experimental planning is necessary to pull the maximum of relevant information and to make the best use of these experiments.
We propose to share our experience and give some practical advice to improve communication between biologists, bioinformaticians and statisticians and to propose an optimal experimental design adapted to each biological question.
1. Try to harmonize your vocabulary and share a minimal common background !
To be able to plan in an effective way, biologists, bioinformaticians, and statisticians must be consulted and must share a common vocabulary. After a reminder of the RNA-Seq technology, its vocabulary, and the scopes of its use, we will define the major statistical principles of experimental design. Randomization, replication and blocking (arrangement of experimental units into homogenous blocks) principles proposed by Fisher from 1935 are always valid.
2. What is the biological question ?
We will explain the importance of a clear formulation of biological questions as well as the following setting up of experimental design, taking into account biological, technical and financial constraints. We will adapt the basic principles of experimental design in the case of the identification of differentially expressed genes using RNA-Seq technology.
3. Anticipate difficulties !
We will discuss the potential sources of variability and give a check-list to ensure that all the needed elements will be collected before the execution of the experiment.
4. Make good choices !
We will finally give some recommendations based on recent bibliography concerning RNA-seq, particularly on two main points : replication choices, and optimum compromise between replication number and sequencing depth depending on the biological questions.

B25: Martina Sattlecker, Hamel Patel, Susie Humby, Richard Dobson and Stephen Newhouse. Cross-disorder assessment of a diagnostic Alzheimer's disease gene expression signature

Abstract: Background: In recent years several gene expression patterns have been identified for diagnosis of various disorders, for instance we identified a 48 gene signature in peripheral blood predicting Alzheimer's disease (AD) with 75% accuracy. However, no attempt has been made to assess if diagnostic models are disease specific or weather they or share a

common aetiology amongst disorders disorders. We tested our Alzheimer's disease gene expression patterns for disease specificity across neurological and age related disorders. Material and methods: We obtained 17 data sets, including neurological, autoimmune, diabetes and arterial related disorders from public depositories. Each dataset varied from one another by either microarray-platform used to generate the data (Illumina or Affymetrix), expression chip (2 different expression chips used for Illumina and 4 different expression chips used for Affymetrix), sample size (33 to 222) and sample type (blood or brain). For each dataset 1000 Random Forest classification models using the genes from the AD gene expression pattern were built by bootstrapping 75% of the smallest group from case/control, using equal bootstrapped complimentary case/control numbers and building a model with the probes to which the 48 genes mapped to.
Results: The average accuracy, sensitivity and specificity over the 1000 models were calculated to demonstrate these 48 genes are not specific to AD and can be implemented as a diagnostic classifier in all 17 datasets to achieve an accuracy range of 89% (Rheumatoid arthritis) to 72% (Parkinson's disease), sensitivity range from 92% (Rheumatoid arthritis) to 71% (Parkinson's disease) and a specificity range of 74% (Parkinson's disease) to 91% (Amyotrophic Lateral Sclerosis). We also compared the gene expression in blood to brain expression and an accuracy range of 73% (Major Depressive Disorder) to 82% (schizophrenia). Heat maps of the most influential genes in each classifier model across disorders indicated that no single gene was the driving force behind the classifiers and differential expression analysis suggests a different subset of genes in each disorder are most likely driving the classifier. Permutation tests verified the classifier is not influencing the results, and correlation tests showed no association with dataset sample size to results.
Conclusion: Our results suggest that the AD gene expression signature might not be AD specific, rather an indication of ill health. Further investigations and comparisons of signatures for other disorders are required.

B26: Bernard Fongang and Andrzej Kudlicki. Conserved transcriptional regulatiory modules in mouse, chicken and zebrafish somitogenesis networks.

Abstract: The metameric segmentation of the vertebrate body is established during somitogenesis, when a cyclic spatial pattern of gene expression is created within the mesoderm of the developing embryo. Although several studies have been devoted to the subject, the mechanism controlling vertebrate somitogenesis is still poorly understood. A recent study shows that the generally accepted "clock and wavefront" model involving Notch, Fgf and Wnt signaling pathways is partially redundant and some of its components may be unessential. To identify the essential genes and interactions in the somitogenesis network, we combined two approaches. First, we use model-based timing of transcriptional regulation to precisely identify the moments in time, when the particular genes are active. This method, based on maximum entropy deconvolution, allows selecting potential causal dependencies within the underlying genetic, signaling and transcriptional networks. Second, we use a maximum-likelihood approach to compare results of such analyses from three different vertebrate species: mouse, chick and zebrafish. As a result, we obtained the list of causal interactions in the somitogenesis network that are evolutionarily conserved and are thus the primary candidates for the most essential, core modules of the regulatory network. We expect that our results will lead to identifying the regulatory interaction between somitogenesis and other conserved developmental processes, including morphogenesis and regulation of genes in Hox clusters.

B27: Xiaobei Zhou, Helen Lindsay and Mark Robinson. Robustly detecting differential expression in RNA sequencing data using observation weights

Abstract: A popular approach for comparing gene expression levels between (replicated) conditions of RNA sequencing data relies on counting reads that map to features of interest. Within such count-based methods, many flexible and advanced statistical approaches now exist and offer the ability to adjust for covariates (e.g. batch effects). Often, these methods include some sort of 'sharing of information' across features to improve inferences in small samples. It is important to achieve an appropriate tradeoff between statistical power and protection against outliers. Here, we study the robustness of existing approaches for count-based differential expression analysis and propose a new strategy based on observation weights that can be used within existing frameworks. The results suggest that outliers can have a global effect on differential analyses. We demonstrate the effectiveness of our new approach with real data and simulated data that reflects properties of real datasets (e.g. dispersion-mean trend) and develop an extensible framework for comprehensive testing of current and future methods. In addition, we explore the origin of such outliers, in some cases highlighting additional biological or technical factors within the experiment. The paper has been published on Nucleic Acids Research in April 2014; our robust framework is available in R/Bioconductor edgeR package.

B28: Chakravarthi Kanduri, Minna Ahvenainen, Anju K Philips, Tuire Kuusi, Harri Lähdesmäki and Irma Järvelä. Transcriptional modulation of neurotransmission by music performance

Abstract: Listening to and practicing music is common in all societies. In biological terms, music perception and practice represent a fusion of all the cognitive functions of the human brain. Musical training exerts multiple positive effects on the human brain's structure and function. Practicing music in particular is known to induce neuroplasticity and enhance cognitive performance, and memory. However, the molecular mechanisms and biological pathways mediating the effects of music performance remain unknown so far.
Here, utilizing a combination of genomics and bioinformatics methods, we compared the human blood genome-wide transcriptional responses of professional musicians after performing in a 2-hour concert (classical music) and after a 2-hour 'music-free' control session. In the concert, peripheral blood samples were collected from 13 professional musicians just before and after their performance. 10 other professional musicians volunteered to a control session, where periheral blood samples were collected just before and after a 'music-free' leisure activity of 2 hours that includes conversation or reading a magazine or taking a walk outside. Total RNA that was isolated from the peripheral blood samples was then assayed on the Illumina HumanHT-12 v4 beadarray (Illumina Inc.; San Diego, CA, USA), which targets more than 47000 probes. The Ethical Committee of Helsinki University Central Hospital approved this study.
Relative to the control session, the transcriptional responses of professional musicians after musical performance showed an up-regulation of genes that participate in synaptic neurotransmission, neurotransmitter uptake, neurotransmitter biosynthesis and transport, and associative learning. We also found the up-regulation of genes that participate in calcium signaling pathway, especially those that maintain calcium ion homeostasis. In addition, the response elements of glucocorticoid/corticosteroid hormones, cAMP, and oxidative stress were also found with elevated activity. On the other hand, the down-regulated genes are known to enhance neuronal apoptosis, the characteristic of neurodegeneration. Upstream regulator analysis provided novel insights about the mechanisms underlying music performance. In line with the neuroscientific literature, all these findings shed light on the molecular mechanisms behind musical training-induced neurotransmission, cognitive enhancement, and neuroprotection.

B29: Rafael Takahiro Takahiro, Pedro Rafael Costa and Ney Lemke. Transcription of the ribosomal DNA in density traffic of RNA polymerases in Escherichia coli using a stochastic model sequence-dependent

Abstract: Ribosomal genes are essential in cellular metabolism, as a consequence of their importance; they are exquisitely controlled and highly transcribed. Multiple rounds of transcription initiation are the main mechanism to guarantee a fast enough ribosomal RNA synthesis supplying ribosomes to assure cell growth. Condon et al. estimated the transcription rate of these genes studying the elongation of the ribosomal RNA operon of E. coli. In their experiments, they measured the incorporation of [3H] adenine into the final tryptophan - tRNA (tRNATrp) after addition of rifampicin by hybridization to filters containing an excess of tryptophan probe. The time evolution of the tRNATrp activity shows a plateau achieved when the last RNA polymerase finish the elongation and indicate the required time for the process. Based on this result, they determined the RNAP elongation rate: 90 bp/s. Quan et al. analyzed the density of RNAPs transcribed the rrnB operon by electronic-microscopy (EM). They divided the rrnB operon in 20 sections and count the number of RNAPs in each one and showed a non-uniform distribution along the operon. In this work we propose a model based on Costa et al. that considering pauses and RNAPs collisions in a multiple round of transcription sequence-dependent simulations. We divided the rrnB operon in 20 sections and simulated 1000 times using our stochastic sequence-dependent model for multiple rounds of transcription to obtain precise results. We estimated the NTP concentration, the average stalling force of RNAP and the number of active RNAPs during fast exponential growth in E. coli. In case of collisions between RNA polymerases, the trailing RNAP "pushes" the leader, which induces to a cooperative behavior that attenuates the backtracking. Our model reproduces the time dependent behavior of [3H]- tRNA-Trp transcript observed experimentally. When we increased the concentration of the nucleoside triphosphates the transcription rate increased by a factor of two. The results showed that NTP concentration increases the velocity of the processes and corroborates Schneider et al. experimental, however the transcription rate remains too small in comparison with experimental results. Our results for RNAPs density are consistent with Quan et al. experiments. But our preliminary results indicate that other factors besides multiple rounds of transcription and nucleoside triphosphates concentration are required to explain the highly transcribed rate observed in vivo and RNAPs density near the genes 16S and 23S rRNA. We are currently considering extending the model to include: interactions between the nascent mRNA and the RNAP, the influence of transcription factors and the influence of the cellular environment on the parameters we are using, since as a rule our parameters are based on in vitro studies.

B30: Feargal Ryan, Ian Jeffrey and Marcus Claesson. Metatranscriptomics of colonic biopsies in Inflammatory Bowel Disease

Abstract: Crohn's disease and ulcerative colitis are Inflammatory Bowel Diseases (IBD) characterized by chronic and relapsing inflammation of the gastro-intestinal tract. They cause lifelong suffering, as well as considerable drainage of health care resources. Although their aetiology is still unclear there is a growing body of evidence for a significant microbial factor. Previous research in this area has focused on examining the microbial and gene composition. While this has produced some interesting results, there has been no consistent population of bacteria found to be associated with health or disease in IBD. In this study we take a new approach and focus on the global gene expression of these communities through RNA sequencing of colonic biopsies. Biopsies were collected from inflamed and non-inflamed colonic mucosa in 19 IBD patients. Using RNA-Seq with unprecedented depth we compared microbial metatranscriptomes in these colonic biopsies. This was done using 600Gb of

Illumina HiSeq RNA-Seq technology (15Gb/sample). Raw reads were quality filtered and trimmed using trimmomatic before aligning to the human genome (hg20) with STAR. The SILVA database along with Bowtie2 was used for identifying and removing rRNA sequences. The remaining reads were aligned using bowtie2 against a non-redundant gene catalogue constructed from multiple previously published metagenomic studies of the human gastrointestinal tract. DESeq2 was subsequently used to analyse the count data and identify differentially expressed genes. This led to the finding of microbial genes which are significantly differentially expressed between inflamed and non-inflamed mucosa in bacterial species. Furthermore, the count data from these samples show a clear distinction between bacterial gene expression of ulcerative colitis and Crohn's disease. Thus, our analysis has revealed a clear difference in the gene expression of bacteria in the colon of ulcerative colitis and Crohn's disease patients, and demonstrated that novel approaches are required in order to understand complex multi-factorial diseases.

B31: Francisco J Altimiras, Barbara Uszczynska, David E Loyola, Emilio Palumbo, Anna Vlasova, Robert Mj Deacon, Rodrigo A Vasquez, Roderic Guigó and Patricia Cogram. Whole Transcriptome Analysis by RNA-sequencing Reveals Novel Alzheimer's Disease Biomarkers in Natural Population of the Rodent Octodon degus

Abstract: The Octodon degus, a South American rodent endemic to Chile, has been recently found to naturally develop histopathological signs of Alzheimer's Disease: accumulation of soluble Aβ oligomers and tau protein phosphorylation, as well as cognitive decline in spatial memory (T-maze) and object recognition memory (ORM). O. degus can live approximately 9-10 years in captivity, and from 3 years of age O. degus has been found to spontaneously develop an AD-like neuropathology. The natural onset and development of neurodegeneration, without the need of genetic manipulation, validate O. degus as a suitable animal model for studying AD. However, it is necessary to asses other events linked to AD - neuroinflammation, oxidative stress, immunological response, as well as an impairment to develop "activities of daily living" (ADL) - to further assess the validity of O. degus as an animal model for AD. The present work investigates the global gene expression profile of O. degus at the onset of the AD-like neuropathology. Understanding how common genetic and immunological variants function in AD is of great importance. For this purpose, we developed a whole transcriptome analysis using RNA-sequencing on wild-type Octodon degus. Animals were captured from a natural population in central Chile at Rinconada de Maipú, 30 km west of Santiago. To assess the quantity of soluble Aβ oligomers (Aβ1-42) we used MALDI-Tof mass spectrometry. For whole transcriptome analysis, we used whole brain samples from two groups (n=8 animals), according to their brain Aβ1-42 quantification. Total RNA was isolated using PureLink RNA Mini kit and mRNA enrichment was carried out using MicroPoly(A)Purist kit. RNA-sequencing of paired-end libraries was done using ABI SOLID 5500xl system. For reads alignment was used Lifescope genomic analysis pipeline. For read counts per gene were used HTSeq-count. For differential gene expression analysis was used edgeR package in R. For gene ontology classification analysis was used clusterProfiler package in R. We obtained an average throughput of 150 million reads per sample with a 150X estimated depth coverage. Multi dimensional scaling plot were used to assess the biological coefficient of variation in the samples. The criteria used to identify differential expressed genes were: count per million reads (CPM)>1, LogFC [1,-1] and False Discovery Rate (FDR) < 0.05. It was obtained 822 differential expressed genes according these criteria, 371 up-regulated genes and 451 down regulated genes. Biological processes such us neurodevelopment and immune system response were enriched in GO-terms analysis. This work is the first genome-wide analysis in Octodon degus. We built a reference transcriptome

for genetic O. degus research. The gene expression analysis reveals novel tentative biomarkers that could be useful for drug discovery in the fight to treat AD.

B32: Henry Han. Feature Selection for RNA-Seq Data Analysis

Abstract: RNA-Seq data are challenging existing omics data analytics for its volume and complexity. Although quite a few computational models were proposed from different standing points to conduct differential expression (D.E.) analysis, almost all these methods do not provide a serious feature selection for high-dimensional RNA-Seq count data. Instead, most or even all genes are invited into differential calls no matter they have real contributions to data variations or not. Thus, it would inevitably affect the robustness of D.E. analysis and lead to the increase of false positive ratios.
In this study, we presented a novel feature selection method: nonnegative singular value approximation (NSVA) to enhance RNA-Seq differential expression analysis by taking advantages of RNA-Seq count data's built-in characteristics. Mathematically, nonnegative singular value approximation is built upon the Perron-Frobenius theorem, which has been widely used in Google webpage ranking, and singular value decomposition (SVD) method. As a data driven feature selection method, it does not require any priori data distribution assumption for RNA-Seq count data. On the other hand, as a variance based feature selection method, it selects genes according to its contribution to the whole data variance of RNA-Seq count data. Combining with classic differential expression analysis algorithms, our nonnegative singular value approximation demonstrated the advantage in identifying differentially expressed genes with high sensitivities on benchmark data. Especially, it contributes to overcoming the sequencing dependence of parametric differential expression analysis methods, in addition to decreasing the false positive rates for the non-parametric differential expression analysis methods.
In addition, we compared our feature selection with count-based naive feature selection, principal component analysis, nonnegative matrix factorization and other competing methods to further demonstrate its advantage in D.E. analysis. It is interesting to see that our purely additive variance-based feature selection method achieved statistically significant advantages than those methods in selecting potentially differentially expressed genes under a same D.E. analysis model. Our results demonstrated that our proposed feature selection method is an effective way to enhance D.E. analysis by lowering false discovery rates and maintaining sequencing depth independence. Finally, we proposed a NSVA-based biomarker discovery algorithm for RNA-Seq data to explore its potential in capturing meaningful biomarkers.

B34: Anita Lerch, Cristian Koepfli, Zbynek Bozdech, Ivo Mueller, Liam O'Connor and Ingrid Felger. Inferring gametocyte stage specific transcriptomes from mixed life stages of Plasmodium vivax field samples

Abstract: Plasmodium vivax is the second most common cause of malaria in humans and threatens almost 40% of the world's population. P. vivax parasites cannot be cultured continuously in vitro and have to be studied in human blood samples collected in the field. Analysis of gene expression is complicated by the fact that field samples contain a mixture of life stages, despite attempts to enrich for specific stages. Gametocytes are the sexual blood stage of P. vivax and the only stages relevant for transmission to the mosquito vector. Thus gametocytes are an important target of transmission interrupting interventions, but the transcriptome of gametocytes is not available.
Here we present an approach to infer the transcriptome of P. vivax gametocytes by applying a multivariate regression model. The model uses gene expression data from RNAseq experiments and relative abundance of each stage in these samples to estimate stage-specific

gene expression levels, e.g. rings, trophozoites, schizonts and gametocytes. The RNAseq data was obtained from gametocyte enriched field samples with known stage composition and from synchronized blood stage parasites in short term culture. A major challenge for the model is to accurately predict stage-specific gene expression despite high variation of gene expression within samples. Preliminary results of a limited number of samples suggest that we can consistently infer a set of gametocyte specific expressed genes, including the known gametocyte specific genes pvs25 and pvs28. The results also suggest that among gametocyte specific genes, pvs25 and pvs28 are highest expressed.
We conclude that it is possible to infer high expressed gametocyte specific genes from P. vivax field samples with known stage composition. This data will greatly enhance our understanding of the biology of this neglected parasite.

B35: Arthur Chun-Chieh Shih, Ling Li, Ya-Ting Chang and Chien-Chang Chen. Co-expressed Regulatory Functional Modules in Pressure overload-induced Cardiac Hypertrophy

Abstract: Heart disease has been one of the leading causes of people death in worldwide every year. Many of the diseases are accompanied by cardiac hypertrophy. Previous studies have identified and validated couples of key transcription factor (TF) and microRNA (miRNA) regulators that are involved in the process of cardiac hypertrophy. However, most of the studies were only based on the samples obtained at one or two time points. Only a few genes have been surveyed their whole dynamic regulations. Thus, a comprehensive study of the dynamic regulation among the whole process during cardiac hypertrophy is still lack.
In this study, we first used microarray analyses to estimate the expression changes of mRNAs and miRNAs isolated from hypertrophic murine hearts subjected to transverse aorta banding surgery (TAB) at five time points among the four weeks, respectively. Then, we normalized the microarray intensity data between different time points and identified a set of genes and miRNAs with significantly up- or down-regulated at least at one time point. These genes were divided into TFs and non-TF genes according to a murine TF database. Instead of using the fold change, we used the intensity profiles to calculate the correlations between genes and regulators in each condition. Based on the statistical significances of correlation preferences with the up/down-regulated genes and those also belonging to cardiovascular (CV) genes against the overall total correlations and those by only regulated genes, respectively, almost all the TFs and miRNAs can be classified to five groups. We identified the TFs and miRNAs in the first two groups there were positively or negatively co-expressed with the CV genes more significantly than with all up-/down-regulated genes. These regulators should be more related to CV functions. Third, we integrated several TF-Gene and miRNA-Gene databases and examined whether the co-regulated pairs were also co-expressed. From the total correlation distributions, we found that the expressions of the TFs and their target genes were significantly positive correlated in TAB-induced heart while no preference in sham-operated hearts. Moreover, those of the miRNAs and the predicted target genes were negatively correlated in the TAB condition but also no preference in the sham condition. It indicates that these regulator and target pairs were not only co-regulated but also co-expressed. Finally, we selected dozens of cardiovascular-related functions and used the up/down-regulated genes involved in each of the functions to identify the interacting regulators. Combining the TF-gene and miRNA-gene clustering results and also integrating other regulatory information together, we not only reconstructed the regulatory network but also created the coexpression regulatory maps for the selected cardiovascular functions. In short, our conducted results can enhance the understanding in the TF- and miRNA-regulated gene networks in cardiac hypertrophy.

B37: Enrica Calura, Gabriele Sales, Paolo Martini and Chiara Romualdi. Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles

Abstract: The production rate of gene expression data is nothing less than astounding. However, with the benefit of hindsight we can assert that, since we completely ignored the non-coding part of the transcriptome, we spent the last decade to study cell mechanisms having few data in our hands. In this scenario, microRNAs, which are key post-trascriptional regulators, deserve special attention. Currently, miRNA and gene circuits are identified through the combination of binding prediction and expression correlation analyses, MAGIA, the web tool we developped, is an example to feel this aim (Sales et al NAR 2010, Bisognin et al NAR 2012). Although effective in many cases the simple correlation does not imply a causal relationship and a lot of false positive miRNA-mRNA interactions are still found. Moreover, miRNA and target genes are characterized by many-to-many relationships and they should be considered as part of a much more complex system of cellular interactions. Recently, to analyze the cellular circuits we developped a new web tool dedicated to topological pathway analyses called Graphite Web (Sales et al NAR 1013). Given the state of knowledge about the biogenesis of miRNAs, their mechanisms of action and the numerous experimentally validated target genes, miRNAs are also gradually appearing in the formal pathway representations such as KEGG and Reactome maps. However, the number of miRNAs annotated in pathway maps are very few and pathway analyses exploiting this new regulatory layer are still lacking. To fill these gaps, we developed micrographite a new pipeline to perform topological pathway analysis integrating gene and miRNA expression profiles. Micrographite analysis of gene and miRNA integrated transcriptome is used to study and dissect the epithelial ovarian cancer gene complexity and miRNA transcriptome defining and validating a new regulatory circuits (Calura et al CCR 2013 and Calura et al., NAR 2014).

B38: Owen Dando, Peter Kind and Ian Simpson. Piquant: a pipeline for assessing the performance of transcriptome quantification tools

Abstract: RNA-sequencing has become an important technique for characterising and quantifying the transcriptome, and many computational methods have been described to reconstruct transcripts from RNA-seq data and subsequently estimate their abundances. At the level of genes, expression estimates calculated by software implementations of these methods have been shown to be relatively robust. However, at the level of transcripts, problems arising from the ambiguous origin of short RNA-seq reads and from bias in their sequence composition are compounded, and thus estimates of isoform abundance may be less accurate. It is therefore important to understand the conditions under which different transcriptome quantification tools perform well or more poorly, and how the many optional parameter choices available for each tool may affect their performance.
Here we present a software pipeline to assess the accuracy of transcriptome quantification tools. In the first stage of the pipeline, RNA-seq reads are simulated from a starting set of transcripts under specified combinations of sequencing parameters: different read lengths and sequencing depths, single- and paired-end reads, reads with or without sequencing errors, and reads with or without sequence bias. In the second stage, a number of transcriptome quantification tools (or the same tool with different optional parameter choices) estimate isoform abundances for each set of simulated reads. In the final stage, the isoform expression estimates calculated by each tool for each data set are compared to the known transcript abundances used to generate the reads. The comparative accuracy of expression estimates calculated by each tool can then be assessed as sequencing parameters change, or for different groups of transcripts segregated by particular transcript classification measures, via a wide range of automatically generated statistics and graphs. The pipeline is easy to use and extend,

and its three stages can be executed with a minimum of intervention.
As an example of the use of the pipeline, we compare the performance of four popular transcriptome quantification tools - Cufflinks, RSEM, eXpress and Sailfish - in calculating abundances of transcripts of human protein-coding genes (as defined in Ensembl release 75), under different combinations of sequencing parameters. We analyse how their accuracy in calculating isoform expression levels varies as read length and sequencing depth increase, and for single- versus paired-end reads. Moreover, we inspect how their performance is differentially affected by particular properties of transcripts - for example, their length, real abundance, and uniqueness of sequence.

B39: Mei-Ju May Chen, You-Yu Lin, Wen-Hsiung Li and Chien-Yu Chen. Transcriptional Regulation of Long Non-coding Gene Expression in Drosophila melanogaster: A Genome-wide study using RNA-seq

Abstract: By the advance of next-generation sequencing technology, RNA sequencing (RNA-seq) uncovers a novel insight that long non-coding RNAs (lncRNAs) are not transcriptional noises but play essential roles in several biological processes. Over the past years, most studies have focused on investigating the function of lncRNAs. For example, lncRNAs might be essential regulators of diverse cellular functions such as epigenetic silencing and transcriptional regulation. However, only a few studies went upstream to ask how lncRNAs are regulated. In fact, it is quite challenging to study this issue in a genome-wide level owing to the fact that transcription factor binding sites (TFBSs) with experimental validation are currently scarce. In this regard, in silico predictions of TFBSs may be needed to investigate the regulation of lncRNA expression. This study incorporates de novo motif discovery to systemically investigate the presence of TFBSs in lncRNA promoters and how it is related to the regulation of lncRNA expression. This study adopted time-course RNA-seq data of Drosophila melanogaster, including 30 developmental stages (modENCODE IDs: 4433-4462). Co-expressed coding and long non-coding (LNC) gene clusters were constructed by applying hierarchical clustering on the expression profiles that were quantified using eXpress on the transcripts collected from FlyBase and UCSC genome browser. To identify potential TFBSs for each cluster, de novo motif discovery (eTFBS) was conducted on the promoters (upstream 500 bps from the transcriptional starting sites) of coding genes in a cluster. Then, the discovered motifs were mapped onto the LNC gene promoters in the same cluster to see whether the discovered motifs from the coding gene promoters could be also found in the co-expressed LNC gene promoters. To validate the discovered sites, 815 annotated TFBSs were collected from the JASPAR, the TRANSFAC and the Fly Factor Survey databases. The results of the clusters with highly correlated expression (Pearson's correlation coefficient >0.9) showed that more than 85% of the discovered motifs were similar to the annotated TFBSs, revealing the potential functional roles of the discovered motifs. Moreover, ~90% of lncRNAs were found to contain at least one discovered motif in their promoters. To confirm the results were not random events caused by genome-wide motif mapping, we further mapped the discovered motifs onto 3' untranslated regions (3' UTRs) and introns of coding genes. The frequency of motif hits in LNC gene promoters was significantly higher than 3' UTRs and introns by using paired t-test, while it was not different from the coding gene promoters. In summary, this study provided a genome-wide evidence to show that same TFBSs were usually co-occurred in the promoters of coding and LNC genes in a co-expressed cluster. This suggested that the regulatory mechanism of lncRNA expression is generally similar to the coding genes in the fruit fly system.

B40: Audrey Bihouée, Erwan Delage, Sébastien Charneau, Abdelhalim Larlhimi, Audrey Donnart, Damien Eveillard, Jérémie Bourdon, Pierre Lindenbaum, Gilles Toumaniantz,

Flavien Charpentier, Richard Redon and Géraldine Jean. A combined approach to identify therapeutic targets involved in Progressive Cardiac Conduction Defect

Abstract: Sudden cardiac death (SCD) claims almost one million deaths annually in industrialized countries and results most frequently from ventricular fibrillation. Five to ten percent of cases of SCD occur in the absence of structural heart abnormality: such cases have been associated with inherited arrhythmias that can be related to alterations in the heart electrical activity.
In this study, we propose an original bioinformatic protocol to find therapeutic targets involved in early stages of Progressive Cardiac Conduction Defect (PCCD) pathology for which it currently does not exist preventive treatment. For the sake of application, one investigates two transcriptomes as obtained from a heterozygous mouse invalided for Scn5a gene (the first gene associated to PCCD). This model presents a progressive deterioration of conduction during aging. This phenotypic feature is characterized by an ECG evaluation on young adults that distinguishes between mice touched by mild disorders of conduction and others that precociously show much stronger defects. In order to perform a global analysis of the transcriptome profile from cardiac cells at different stages of PCCD development, we compare 3 groups of mice (wild type, mild, strong) at different ages: 6 weeks (the earlier stage where variability of cardiac conduction defect appears), 30 weeks (when restructuring markers differ according to ECG phenotype) and 45 weeks (when TGF-beta pathway is activated). We obtain 27 samples (3 individuals per group) of RNA-seq data extracted from the free wall of left ventricle. For methodological validation, we also produce micro-array data for the same groups of mice.
After showing that micro-array data are less sensible than RNA-seq data but still highly correlated, we perform 2 different analysis: differential gene expression and weighted correlation network analysis.
The first analysis is a pipeline of bioinformatic tools for computing differentially expressed genes: TopHat2 for the alignment of the reads on the reference genome, HTseq/DESeq for a count-based differential expression analysis at gene level. The results show there is a strong age effect on gene expression compared to the phenotype. The second approach applies WGCNA to emphasize highly correlated modules within transcriptomic gene correlation network. Some modules are related to phenotypic traits (age, ...). Moreover, GO analysis applied to the identified modules highlights particular modules with enriched GO terms related to channel activity, immune response or signaling pathways. Results from both approaches are then projected into common graphical representation that depicts modules that contain differentially expressed genes under a given condition, which represent fruitful insights for future experiments. The interesting genes we extract are potential therapeutic targets that need to be biologically validated.
This approach is generic and is part of heterogeneous data integration process.

B41: Sepideh Babaei, Ahmed Mahfouz, Boudewijn P.F. Lelieveldt, Marcel Reinders and Jeroen De Ridder. Multi-scale chromatin interactions are predictive for spatial co-expression patterns in the mouse cortex

Abstract: The three dimensional conformation of the genome in the cell nucleus influences important biological processes such as gene expression regulation. Recent studies have shown a strong correlation between chromatin interactions and gene co-expression. However, determining whether the gene co-expression is predictable from frequent long-range chromatin interactions remains challenging. We derived scale-aware topological measures of the chromatin interaction network based on the Hi-C data of mouse cortical cells. We used these measures to evaluate how the cortical genome-wide chromatin interactions predict spatial co-expression between genes in the mouse cortex. Consistent with the strong

correlation, we found that the chromatin interaction profile of a gene-pair is a good predictor of their spatial co-expression. The best prediction performance is obtained using scale-aware topological measures with an AUC prediction performance of 0.84. Our results suggest that the topological description of the network in the multi-scale fashion is important to capture co-expression relationships between genes. Hence, for co-expression prediction it is necessary that different levels of chromatin interactions to be taken into account ranging from direct interaction between genes (i.e. small-scale) to chromatin compartment interactions (i.e. large-scale).

B42: Frédéric Fer, Julie Aubert and Jean-Marie Beckerich. Combining kinetic modeling and transcriptomic analysis to study the resilience on a model of microbial ecosystem

Abstract: RNA-sequencing(RNAseq) has enabled great advance both in microbiology and ecology. It consists in sequencing RNA transcripts of a a biological sample to estimate genes activities. RNAseq experiments on ecosystems produce tremendous amount of data, typically tables of thousands of rows (genes) and tens of columns (samples).
In microbial ecology, one problematic is to link these genetic activities and major ecosystem functions to study the maintenance of this functions following a perturbation (resilience). The aim of the work is to combinate kinetic modeling and transcriptomic analysis to study the resilience of the proteolysis function on a cheese model ecosystem.
A reduced ecosystem composed of six bacteria species and three yeast species has been developed to perform the the ripening of a washed rind cheese. We used this ecosystem as a model to study the resilience properties of a microbial ecosystem. To this end this controlled ecosystem was challenged by modifying either the salinity of the curd or by omitting one of the two major yeast species. Metatranscriptomics and physiochemical data was acquired during each ripening.
In order to identify the role of each species in the proteolysis resilience, we have developed a kinetic model that integrate the biochemical data of proteolysis and the population dynamics. Model parameters were estimated by Metropolis-Hastings algorithm and MCMC chains using the MCMC matlab Toolbox [2]. With this approach,we have highlighted the major role of the yeasts Debaryomyces hansenii and Geotrichum candidum in this phenomenon. Then we have focused transcriptomic analysis on these major species.
We use a normalization method inspired from TMM to correct simultaneously libray size and variations of species concentration. We selected genes declared differentially expressed using the R package DESEQ2 [3] between a disturbed condition and the normal one at 5% after correcting for multiple testing. As the number of selected genes was exceeding the number of observations, we have used lasso penalized regression to identify genes potentially predictors for the proteolysis.
Our approach has identified a set of implicated genes. This work is a example of approach which integrate, analyze and link diverse data (microbial growth, biochemical, physiochemical and genomic data) in order to understand the microbial ecosystem resilience. This will provide to the scientific community a set of reusable and flexible tools to study with high-throughput methods microbial ecosystems.

B43: Philipp Senger and Shweta Bagewadi. Automatic Quality Assessment Of Microarray Datasets Using Ensemble Methods

Abstract: Gene expression profiling is one among the powerful high-throughput technologies advancing the understanding of patho-mechanisms in complex diseases. Gene expression data (e.g. microarray and next generation sequencing) are made available through public databases such as Gene Expression Omnibus and ArrayExpress. Although, these databases provide a

semi-structured meta annotation of the data, there is no standardized labeling of the underlying quality. Estimation of the quality level is a challenge due to different platforms and experimental setup followed by the laboratories. Different (non-)commercial tools provide the ability to calculate the quality level but they are often not applicable, easy to use, or freely available. As a consequence, machine learning techniques are used for quality classification.
Here, we propose an approach that uses artificially generated microarray data of pre-defined quality levels to train an ensemble SVM-based system, in order to predict the quality of real world data. Using, SyNTren generator we produce matrices encoding the expression levels of genes in this network with pre-defined noise levels. The quality measure is an automatically assigned value encoding three categories “good”, “medium”, and “bad”, discretized based on the input noise values. We validated the direct influence of different noise parameters on the data quality by measuring correctly identified differentially expressed genes. By using a developed parser, we are able to export the data into chip layouts of well known microarray vendors like Affymetrix. The artificially generated data repository contains in total 1,623 different samples equally covering all quality levels.
Using several R packages (e.g. arrayQualityMetrics), a set of 26 features for each of the artificially generated sample are calculated. These features are based on the expression data, stored as a simple matrix. This makes the approach easily applicable for other platforms. An ensemble of SVM-based classifiers are trained on respective quality level of a randomly chosen subgroup of samples. Several optimization steps are performed to find the best suiting architecture of the classifier ensemble. The results show a robust classification of correct quality bins of about 86.9%, on 10-fold cross validation on artificial validation data.
In addition, a set of 166 manually annotated samples from different platforms are taken as an external real world validation dataset. The ensemble classifier achieves a prediction accuracy of 84.8% showing that the classifiers trained on artificial data are also able to efficiently classify the quality label of real world data.
With the ability to generate large sets of training samples, and integration of features from different quality control (QC) tools, the proposed workflow shows to be better than the conventional approaches for microarray QC with the possibility to extend for other gene expression technologies.

B44: Marcelo Segura, Hector Keun and Tim Ebbels. Pathway based models have similar predictivity and robustness to models based on random collections of genes

Abstract: Predictive models based on “omic” data sets are prone to technical and biological noise. Existing methods such as overrepresentation analysis (ORA) suggests that pathway approaches could be a robust alternative to the analysis of ungrouped genes. However, this has not being clearly demonstrated. Using a pathway representation of the data, we compared the performance of pathway and gene based models under simulated noise conditions. Pathway based models showed superior robustness. Surprisingly, the analysis of robustness also showed that the true set of pathways does not perform better than random pathways. However, we have found that models based on true pathways have a characteristic distribution the contribution of true pathway other features proved to be specific to the true pathways set.

B45: Candida Vaz, Choon Wei Wee, Gek Ping Serene Lee, Vivek M Tanavde and Sinnakarupan Mathavan. Next generation sequencing reveals tissue and sex specific known and novel miRNAs in zebrafish

Abstract: Background: The zebrafish, Danio rerio is a prime model organism for the study of vertebrate development and human diseases owing to their extensive reproductive capacity, short generation time, rapid embryogenesis and transparent embryos and comparable organs and tissues to humans. MicroRNAs (miRNAs) have emerged as a major class of small non coding RNAs that regulate several cellular functions. Various studies on miRNA profiling and function have led to the therapeutic application of miRNAs as biomarkers.
Next Generation Sequencing technology offers a sensitive method to detect miRNAs and are being widely used for miRNA profiling and novel miRNA detection.
Principal Findings: This project involves next generation sequencing of miRNAs from different tissues of zebrafish such as the Brain, Gut, Liver, Ovary, Testis, Eye, Heart as well as its Embryo. For some tissues such as the Brain, Gut, Liver, miRNA expression patterns were determined for the male and female zebrafish separately. Deep sequencing data was obtained from Illumina HiSeq 2000 platform comprising of : 11 libraries, approximately 20 million reads and 3 biological replicates per library. Around 92-98% of the sequences mapped to the zebrafish genome (Zv9). The known miRNAs (344 precursors, 247 mature) comprised a wide range of 16-62% of the mapped reads, with the exceptions of Ovary (5.71%) and Testis (7.80%). The un-annotated pool of all the samples (7.57 to 23.03%) with exceptions of Ovary (51.39%) and Testis (55.17%) was used for novel miRNA prediction using miRDeep2.
Of the known miRNAs, a few of them, such as the members of the let-7 family, miR-21, miR-92a were highly expressed in all the samples indicating their importance as housekeeping miRNAs. Comparison of the known miRNA expression data of the tissues showed the presence of tissue and sex specific miRNAs that could serve as important biomarkers. The brain showed the highest number of tissue specific known miRNAs (23) wheras the liver showed the highest number of sex specific known miRNAs (38). Finally, a total of 459 novel pre-miRNAs were detected, among which some appeared to be tissue specfic.
Conclusions: This study involves an extensive analysis of the miRNA expression patterns of several tissues to identify tissue and sex specific known miRNAs that could serve as biomarkers. Addition of novel miRNAs that are nearly double the number of known miRNAs will not only serve as a good resource for further investigation, but will also provide a complete miRNA transcriptome essential for getting a complete picture of miRNA regulation in zebrafish.

B46: Luiz Augusto Bovolenta, Danillo Pinhal, Simon Moxon, Arthur Casulli Oliveira, Pedro Gabriel Nachtigall, Marcio Luis Acencio, Cesar Martins and Ney Lemke. Nile tilapia miRNAs: characterization and target prediction

Abstract: MiRNAs have being identified as important elements in gene regulation. They are small non-coding RNAs containing about 22 nucleotides that pair with complementary sequences within mRNA molecules and this process mediates gene expression silencing either by degradation or translation inhibition of target mRNAs. Many important biological processes are affected by post-transcriptional repression, such as proliferation, differentiation and cellular death. For this reason, we sought to identify and quantify the miRNAs in Nile tilapia (Oreochromis niloticus) whose genome has been fully sequenced. Moreover this species is an important economic fish species in Brazil's aquaculture. In this sense, the study of miRNAs in Nile tilapia is potentially valuable for understanding the genomic organization, biological process inference and the expression levels of miRNAs in vertebrates, as well as for identifying new miRNAs with unknown functions. The objective of the present work was to determine the genomic organization, expression levels and to predict targets of known and putative new miRNAs identified in Nile tilapia samples. Next-generation sequencing together with qPCR recovered 149 known and 84 novel candidate mature miRNAs from 371 hairpins

expressed in 15 samples of adult female and male red and white muscle, eyes, gonads and brain, as well as of female liver, and miRNAs from early developmental stages at 2, 3, 4, 5 and 10 days post fertilization. Then the respective pre-miRNAs were taken from miRBase (using zebrafish Danio rerio as reference), blasted against the Nile tilapia genome (version 1.1) and their physical location obtained. From this mapping we could determine that 183 miRNAs are intergenic, 187 intronic and 1 exonic. At the same time, we performed a first target prediction analysis for novel miRNAs using TargetScan, miRanda and RNAhybrid target prediction tools. Preliminary results have shown many possible targets: while 12291 miRNA-binding target sites were detected by TargetScan and MiRanda, 11202 sites were predicted by RNAhybrid. Moreover, by merging results from the three tools, we found 543 common predicted target sites from 455 target genes. The ongoing target prediction coupled with a Gene Ontology-based functional enrichment analysis will contribute to an extensive characterization of the metabolic pathways and to the understanding of the evolutionary dynamics of miRNAs in vertebrate genomes.

B47: Sandra Koser, Jan-Philipp Mallm, Sabrina Schumacher, Stephan Wolf, Stephan Stilgenbauer, Karsten Rippe, Daniel Mertens and Benedikt Brors. Differential expression analysis of microRNA in CLL and their influence on mRNA expression

Abstract: MicroRNAs are known to play an important role in regulation of gene expression, mainly by post-transcriptional mechanisms. Chronic lymphocytic leukemia (CLL) was recently found to be related with microRNA deregulation. We aimed to characterize the differential expression of microRNAs and other small RNA species in CLL. A further point of investigation is the influence of deregulated microRNA on the splicing patterns of certain genes.
A pipeline was developed to investigate the expression of long and small RNAs from next-generation sequencing data. Eleven CLL samples, four B-cell pools and one T-cell sample were used for the analysis. Two technical replicates were produced for each. A stranded ribo-zero protocol was used to get a single-end 50bp library.
The analysis workflow includes adapter trimming before the reads are mapped to the human genome. The reads are annotated in a strand-specific manner. The read counts are normalized according to the transcript lengths, the library size and to the variance within the samples. Differential expression between the tumor samples an the controls is computed based on the negative binomial distribution; the mean and variance are linked by local regression. The p-values were adjusted to a false discovery rate (FDR) of 10%. A clustering of the samples was performed to test for batch effects and to identify potential sub-groups.
Findings were filtered for negative correlation between miRNA and target mRNA expression. Messenger RNAs found to be differentially expressed included prominent transcripts that were known to be involved in leukemogenesis or other malignant B-cell disorders. Several microRNAs showed strong dysregulation, which is currently studied in further detail with respect to its regulatory consequences. We will also integrate this data with information on histone modifications and DNA methylation as obtained from ChIP-seq and whole-genome bisulfite sequencing, respectively.

B48: Chee Lee, Yanxiao Zhang and Maureen A. Sartor. RNA-Enrich: A cut-off free gene set enrichment testing method for RNA-seq that adjusts for gene read count

Abstract: Tests for differential expression in RNA-seq data are more likely to identify longer and highly-expressed transcripts as significant. Because these transcripts have more reads, programs such as EdgeR and DESeq have greater statistical power to detect differential expression (DE). This bias in the ability to detect DE as a function of read counts per gene has

been shown to affect down-stream analysis, such as gene set enrichment testing, which is used to find over- or under-represented biological functions in the data. We present a new method for enrichment testing of RNA-seq data, RNA-Enrich, that accounts for this bias by adjusting for average read count per gene. Unlike previous enrichment tests designed specifically for RNA-seq, our method does not require a p-value cut-off to define differentially expressed genes (DEGs). Rather, the statistical model underlying RNA-Enrich directly uses p-values from a DE test for enrichment testing. Parallel methods for microarray data have shown improved ability to detect gene sets enriched with either a few very strong DEGs or many only moderately DEGs. Our method is a modified version of the Random Sets method, with the addition of empirically estimated gene-specific weights to control for average read count. The weights are calculated with a smoothing spline using average read count per gene. To test for enrichment of each gene set, weights of genes in the gene set are normalized to sum to 1 and used to calculate a z-score and corresponding p-value.
We first verify that our method results in a well-calibrated type I error rate in multiple RNA-seq datasets. We show that in datasets where DE is positively correlated with average read count of a gene, adjusting for average read count yields more significant gene sets and greatly improves discovery of known biology. Our results suggest that examining one's data for this bias is important for choosing the correct gene set enrichment test, as some DE test results did not exhibit a relationship between average read count per gene and p-values of DEGs; in these datasets, RNA-Enrich results were highly correlated ($r>0.99$) with a previous method developed for microarray data, LRpath, which does not account for this relationship. Our method has several advantages over GOseq, an R package for enrichment testing of RNA-seq data: (1) GOseq requires a p-value cut-off, whereas our method uses p-values from DE testing. (2) GOseq is either run with (a) the Wallenius approximation, which only corrects for the overall locus length or read counts in a gene set as opposed to gene-wise corrections; or (b) permutations, which result in very long run time if accurate p-values are desired and often lead to many tied top-ranked gene sets (with p-value=0). RNA-Enrich does not have these disadvantages, and has a fast run time. In summary, RNA-Enrich fills a void in the available gene set enrichment tests for RNA-seq data by allowing to correct for bias due to average read count.

B49: Marta Rosikiewicz and Marc Robinson-Rechavi. Benchmarking of quality control parameters for RNA-seq data

Abstract: Next generation sequencing technology has become a standard tool for gene expression analysis and over the past few years thousands of RNA-seq datasets have become available in public repositories. The quality control step is essential for developing valuable datasets for exploratory studies. The classical quality assessment methods do not give clear clue which parameters are crucial in identification RNA-seq samples which should be eliminated because gene expression profiles derived from them do not sufficiently represent real biological expression profile of analyzed tissue.
In the current study we tested several parameters for quality assessment of sequencing data like quality score of raw reads, percent of reads mapped, GC bias, coverage uniformity, saturation of splicing junction detection and others. As an independent measure of quality we specified how well the gene expression profile from each library correlates with reference expression profile of homologous genes in the same organ from different species. We discovered that samples, which show lower correlation with the reference, could be identified as of poor quality on the basis of some of analyzed quality metrics.

B50: José Luis Gaete and Marta Fernandez. RNA-Seq applications for plant transcriptome analysis in response to cold acclimation by Ion Torrent technology

Abstract: RNA-Seq technology has become widely used as a tool to understand the transcriptome of a given experimental system. It uses deep-sequencing technologies to sequence a cDNA library to achieve information about the RNA content and transcriptional status of a sample of interest, generating an expression profile, which enables interpreting the functional elements of the genome, and thus reveal the molecular constituents of cells and tissues. The Ion Torrent sequencing technologies simplified sequencing technology translating directly the chemical information codified into digital data on a semiconductor chip, making it faster, simpler and, more scalable than any other available technology.
This study presents a transcriptome analysis of mRNA libraries from the woody plant Eucalyptus nitens, sequenced on the Ion Proton instrument, to identify genes associated to cold acclimation and gene expression quantitation. Twelve expression libraries prepared from leaves of plants exposed to four acclimation conditions to fake the annual seasonal variation in a growth chamber (NA: Non-acclimated, CABF: Cold acclimated before frost, CAAF: Cold acclimated after frost and DA: De-acclimated). The bioinformatics analysis included: i. data preprocessing: trimming of the sequences in 3'-ends and a subsequent filtering, ii. mapping reads to the reference genome of a related species (Eucalyptus grandis) using TopHat2 and, iii. the generation of the table of reads counts using an HT-Seq script, followed by the analysis of differential gene expression in silico with the R-package edgeR for the six pairs of comparisons (NA/CABF, NA/CAAF, NA/DA, CABF/CAAF, CABF/DA and CAAF/DA), to identify differentially expressed genes involved in the cold acclimation process in E. nitens. Additionally, genes showing the most stable expression levels were selected as reference genes for further validation by qRT- PCR.
After the 3'-end trimming and filtering of reads, ~49% of reads were removed. To assess the effect of preprocessing, the filtered and non-filtered reads were compared after mapping. Approximately %20-30% of non-filtered reads were mapped, while ~45% to 57% of filtered reads were mapped. A total of 17,794 genes were identified and specific genes responding to each acclimation condition were observed. The gene expression analysis shown 100 to 600 differentially expressed genes with a Fold Change (FC) cut off of FC < -1.5 and FC > 1.5 and a p-value<0.05 for each pair of comparison. Twenty genes as candidate reference genes with a coefficient variation (CV) from 0.09 to 0.13 were selected. These results reveal Ion Torrent sequencing system as a viable platform for RNA-Seq and transcriptome analysis of woody plants to reveal the dynamic and complex nature of gene expression occurring during cold acclimation.