EMBL-EBI

9th Annual EMBL-EBI PhD Seminars Day
Thursday 28 April 2016
Kendrew Lecture Theatre, EBI South Building

09:40 - 09:45 **Nick Goldman**
Welcome

		Host
09:45 - 10:00	**Damien Arnol** Computational Analysis of space-resolved single cell proteomics data obtained with mass cytometry imaging	*Saez/Stegle*
10:00 - 10:15	**Claudia Hernandez** Predicting changes in kinase activities from conditional phosphoproteomics data	*Beltrao*
10:15 - 10:30	**Paolo Casale** Joint local testing of variant-sets improves power and interpretation of gene-context interactions	*Stegle*
10:30 - 10:45	**Rachel Spicer** The Lipidome in Weight Loss	*Steinbeck*
10:45 - 11:00	**Michael Schubert** Expression footprinting outperforms pathway mapping to generate signatures predictive of cancer drug sensitivity and patient survival	*Saez/Marioni*
11:00 - 11:30	*Refreshments (outside the Kendrew)*	
11:30 - 11:45	**Tom Leonardi** Positional conservation identifies topological anchor point (tap)RNAs linked to developmental loci	*Enright*
11:45 - 12:00	**Mitra Barzine** Comparison and integration of independent high-throughput Transcriptomic and Proteomic studies in Human	*Brazma*
12:00 - 12:15	**Maria Xenophontos** NuRD-mediated regulation of gene expression in pluripotent cells	*Bertone/Flicek*
12:15 - 12:30	**Sergio Santos** Assessing changes in alternative splicing from RNA-seq	*Brazma*
12:30 - 13:45	*Lunch (not provided)*	
13:45 - 14:00	**Omar Wagih** Predicting mechanistic consequences of single nucleotide variants in yeast	*Beltrao*
14:00 - 14:15	**Dhoyazan Azazi** Using High-throughput Genomics To Investigate the Evolution & Function of CTCF Binding in Closely-related Mouse Species	*Flicek*
14:15 - 14:30	**Billy Coleman-Smith** Whole genome diagnosis of bacterial infection types	*Goldman*
14:30 - 14:45	**Matthew Jeffryes** Community Classification of the Protein Universe	*Bateman*
14:45 - 15:00	**Nils Eling** Ageing of the immune system from a single cell perspective	*Marioni*
15:00 - 15:15	**Greg Slodkowicz** Patterns of adaptive evolution: a structural perspective	*Goldman*
15:15 - 16:00	*Refreshments (outside the Kendrew)*	

Abstracts

Speaker: Damien Arnol

Computational Analysis of space-resolved single cell proteomics data obtained with mass cytometry imaging

Mass cytometry imaging is a novel technique that allows sub-cellular resolution measurements of the abundance of dozen proteins directly in tissues (Giesen et al., "Highly multiplexed imaging of tumour tissues with sub cellular resolution by mass cytometry", *Nature Methods*, 2014). The proteins are labelled with antibodies coupled with metal isotopes of specific masses. The tissues are laser ablated into a sub-cellular size, and subsequently injected into a mass spectrometer for measuring the protein abundances.

Preserving the tissue integrity while being able to perform many measurements is a significant improvement to classical proteomics techniques. As the spatial information is not lost, it can be used for instance when modelling intercellular signalling pathways. Furthermore, it also outperforms classical imaging techniques with respect to the number of protein markers (up to 35 in the current state of the art). In a collaboration with the Bodenmiller group in the University of Zurich the aim is to utilise this technique to analyse Breast cancers tumour micro environment (TME). Here, the intention is to better understand the interactions between cancer and stromal cells and this environment.

However, in order to fully exploit the data being generated by this approach, accurate computational methods are required. First, the normalisation and processing is non-trivial and poses several key challenges that are specific to the measurement method. These specificities include cell to cell signal contamination due to cell segmentation errors. It also includes signal variability due to the different antibody binding affinities. Other problems include the instability of the measurement device's sensitivity over the time of the measurement. Besides these specific issues, the usual caveats lying in single cell data analysis, due to cell to cell variability exist. When comparing two single cells with each other, confounding factors due to different cell cycle states or different cell types are to be taken into account. This motivates an integrative approach based on hierarchical Bayesian modelling to address this range of issues, which we will present here.

With Mass Cytometry Imaging, each single cell expression profile comes enriched with several features, such as the type, shape and size of the cell and its surrounding environment, including the expression profiles of the neighbouring cells and the density of surrounding cells. Therefore, the next challenge is to find how these multiple environmental features impact the expression profile at the single cell level, in particular in the Tumour Micro Environment. To model this effect, we use different techniques, such as simple correlation measures, LASSO linear models and random forests. We hope that these models will unveil some of the determinants of single cell heterogeneity and protein expression stochasticity in the Tumour Micro Environment. We will present some of the methods we have used, our preliminary results and our plans for future developments.

Speaker: Dhoyazan Azazi

Using High-throughput Genomics to Investigate the Evolution & Function of CTCF Binding in Closely-related Mouse Species

CTCF binding motifs have undergone the most active extent of expansion in rodent lineages via a particular family of transposable elements: short interspersed elements B2 (SINE-B2). In this talk, I will give an overview of the current state of my project which aims to map how highly active, expanding repeats have rapidly remodelled CTCF binding, and thus chromatin and transcription, in two Mus genus mouse subspecies, separated by one million years of evolutionary divergence: Mus musculus domesticus (C57BL/6J) and Mus musculus castaneus (CASTEiJ).

Speaker: Mitra Barzine

Comparison and integration of independent high-throughput Transcriptomic and Proteomic studies in Human

While we know that the central dogma of molecular biology is not as linear as it could have been perceived, it is still widely accepted that a protein is the translational product of a transcript. Hence, studying jointly Transcriptomics and Proteomics should give us insight on the regulatory processes involved while and after the translation. However, some recent cell studies have failed to show high or good correlations between these two biological layers and therefore the focus shifted to qualitative comparison instead. Here instead, I show that when comparing the same 12 tissues from independent transcriptomic (RNA-Seq) and proteomic (Mass Spectrometry) studies we observe good correlations, even though the samples have different biological sources. Moreover, we found that there is a high similarity between tissue-specific mRNAs and tissues specific proteins.

Speaker: Paolo Casale

Joint local testing of variant-sets improves power and interpretation of gene-context interactions

Understanding the genetic architecture of complex traits remains a central question in quantitative genetics. One of the major avenues is to understand how contextual variables, such as environmental factors, impact the genetic basis of complex traits. To address this, we here present a new statistical approach based on linear mixed models (iSet) that i) jointly models the effect of multiple variants from a genetic region and ii) accounts for categorical context variables. iSet increases power compared to previous interaction tests and enables discerning classical interactions from situations where the configuration of causal variants is altered between contexts. We apply the model to a stimulus eQTL study and a genome-wide analysis of sex-specific regulation of lipid levels. Across these applications, iSet markedly increases statistical power and enhances the interpretability of interaction effects when compared to previous methods.

Speaker: Billy Coleman-Smith
Whole genome diagnosis of bacterial infection types

Due to the diversity of bacterial species, bacterial infections can take many forms. For example, infections can be classified based on which part of the organism is infected (e.g. UTI, Bacteraemia, Bacterial Pneumonia).

Bacterial species can also be divided into two general groups – gram negative and gram positive. However, genotypic factors have a large impact on treatment outcomes, particularly in the case of antibiotic resistance.

There are various PCR and DNA microarray based techniques which can be used to measure these genetic factors, but these come with the down side of only being able to search for what is already known. With prices dropping and speed increasing, whole genome sequencing looks like a promising source of high quality data, which can be used for genetic screening, as well as providing the information needed to continuously improve our knowledge base. Furthermore, the data can be used for molecular epidemiology studies. In my talk, I will briefly discuss the effectiveness of the 'Mykrobe predictor' software package, which uses raw sequence data to predict antibiotic resistance.

Speaker: Nils Eling
Ageing of the immune system from a single cell perspective

The precise molecular mechanism affecting the T cell pool during ageing is not well characterized yet. Age related changes in T cells comprise T cell production, maintenance, function and response to persistent infections. We particularly focus on naive CD4$^+$ T cells, which are characterized as being a homogenous population of cells and remain predominately in a quiescent state. Single cell RNA-sequencing was performed to dissect transcriptional changes in naïve/stimulated CD4$^+$ T cells during ageing in 3 evolutionary related mouse species. This comprehensive dataset establishes the basis for a thorough analysis of several biological components (e.g. CD4$^+$ T cell activation, inter-species comparison of gene expression) across the lifespan of mice. We use robust methods for differential variability testing in order to select differentially variable and differentially expressed genes independently. Our results indicate that the core activation process in CD4$^+$ T cells involves a transcriptional switch from stochastic to tightly regulated gene expression. Furthermore, this activation program is conserved across related mouse species and shows no global but little alterations only on a single gene level during ageing.

Speaker: Claudia Hernandez
Predicting changes in kinase activities from conditional phosphoproteomics data

Cellular decisions rely on the activation or inhibition of signalling pathways regulated by protein post-translational modifications (PTMs). Protein phosphorylation is one of the most abundant PTMs that is catalysed by kinases and is often used by the cellular machinery to decide which processes must be regulated given a particular stimulation. Studying the coordinated activation of kinases across different conditions is therefore an important requirement for the understanding of phospho-regulation in cell signalling.

Previous work on cell decisions has been limited by the capacity to measure only a small number of kinase activities in a single experiment using phosho-specific antibodies. Efforts to study kinase signalling from a systems wide perspective include the use of large scale conditional phosphoproteomics experiments to predict the profile of kinase activities, based on the idea that the activation state of kinases is reflected in the changes of phosphorylation levels for the known substrates. During my talk, I will discuss the performance and comparison of the following methods to estimate conditional changes in kinase activities: GSEA algorithm, non parametric tests and multiple linear regression. Strategies to benchmark the predictions, and the integration of kinase-substrate specificities will also be described.

Speaker: Matt Jeffryes

Community Classification of the Protein Universe

Pfam is a protein family database. Families are described by an HMM-profile which has been generated using a set of representative "seed sequences". Pfam is a human-curated resource and therefore its growth is limited by the ability of its curators to build new families and annotate them. As the 'low hanging fruit' of the protein universe have already been added to Pfam, it has become increasingly hard to make great gains in coverage.

The HMMER web service provides a number of tools to perform protein sequence similarity search, including jackhmmer, which iteratively generates hidden Markov models by aligning the result of sequence similarity searches. Families generated in this way often coincide with Pfam families, but using the information from only a single exemplar protein, instead of multiple seed sequences. Users who perform sequence similarity search with jackhmmer may produce a hidden Markov model which describes a protein family unknown to Pfam, or one which includes proteins contained in an existing family, and other proteins which ought to be included in the family, but aren't. We would like to identify when this occurs, and provide the facility for users to submit this information to Pfam.

By analysing user searches, we have determined that this is a viable approach to expand Pfam. An outstanding challenge is identifying searches which could provide novel additions to Pfam fast enough that users can be prompted to submit the family at the same time as their search results are displayed to them. We have adapted techniques used by web search engines in order to rapidly compare novel families to Pfam.

We aim to integrate these techniques into a curation interface to the HMMER web service, to enable evaluation of their effectiveness for improving curator productivity, before moving on to testing with protein sequence similarity search users.

Speaker: Tom Leonardi

Positional conservation identifies topological anchor point (tap)RNAs linked to developmental loci

The mammalian genome is transcribed into large numbers of long noncoding RNAs (lncRNAs), but the definition of functional lncRNA groups has proven difficult, partly due to their low sequence conservation and lack of identified shared properties. Here we consider positional conservation across mammalian genomes as an indicator of functional commonality. We identify 665 conserved lncRNA promoters in mouse and human genomes that are preserved in genomic position relative to orthologous coding genes. The identified ‘positionally conserved’ lncRNA genes are primarily associated with developmental transcription factor loci with which they are co-expressed in a tissue-specific manner. Strikingly, over half of all positionally conserved RNAs in this set are linked to chromatin organization structures, overlapping the binding site for the CTCF chromatin organizer and located at chromatin loop anchor points and topologically associating domains (TADs). These **t**opological **a**nchor **p**oint (tap)RNAs possess conserved sequence domains that are enriched in potential recognition motifs for Zinc Finger proteins. Characterization of these non-coding RNAs and their associated coding genes shows that they are functionally connected: they regulate each other’s expression and influence the metastatic phenotye of cancer cells *in vitro* in a similar fashion. Thus, interrogation of positionally conserved lncRNAs identifies a new subset of tapRNAs with shared functional properties, and allows us to propose a model of the “extended gene” in which conserved developmental genes are genomically linked to regulatory lncRNAs across mammalian evolution.

Speaker: Sergio Santos
Assessing changes in alternative splicing from RNA-seq

Alternative splicing is an important process in the regulation of gene expression in eukaryotes. Through this process a single gene can be transcribed into different RNA sequences and, therefore, be translated into different proteins. With the advances in sequencing technology, we can now use RNA-seq data to identify the isoforms of each gene that are being expressed in a specific condition. By quantifying the expression level of each isoform of a gene and by comparing two different tissues, we can assess changes in alternative splicing.

I have been developing methods to do this type of analysis over a dataset of 32 human tissues. Preliminary results show that, although alternative splicing can lead to the expression of different transcripts of a gene, many genes have a n-fold dominant transcript, a transcript that is expressed n-fold more than the second most expressed. On average 63% of the expressed genes in a given tissue have a 2-fold dominant transcript and 41% of the expressed genes have a 5-fold dominant transcript. There are genes that express different dominant transcripts in different tissues. For a given pair of tissues, the number of genes that switch dominant transcripts can vary from 1 to ~200. This provides new insights into the understanding of alternative splicing.

Speaker: Michael Schubert
Expression footprinting outperforms pathway mapping to generate signatures predictive of cancer drug sensitivity and patient survival

Numerous pathway methods have been developed to quantify the signaling state of a cell, mostly from mRNA abundance due to the amount of data available. These methods treat pathways either as gene sets whose expression level is tested for different samples, or incorporate pathway structure or correlation of its components. However, these approaches are fundamentally at odds with the notion of tight post-translational control of signal transduction.

Here, we analyzed the predictiveness of downstream mRNA as readout of signaling pathway activity instead of mapping it to the pathway components. Specifically, we created a platform which infers signaling activity from gene expression by identifying genes that are up- or down-regulated upon stimulation with a known pathway modulator in a wide range of conditions. We applied this method to primary tumor and cell line cancer data, and compared it to state of the art pathway mapping methods. We found that our method (1) is the only one that can recover pathway activations mediated by known driver mutations, (2) it provides stronger associations with cancer cell line drug response where it is the only pathway method to recover known oncogene addiction associations, and (3) yielded better biomarkers of patient survival where it is the only method to recover the expected effect on survival of oncogenic and apoptotic pathways.

However, pathway methods in general have taken a back seat compared to gene mutations in terms their importance as biomarkers for drug sensitivty and patient survival. This is why we investigated whether our method is able to further stratify cell lines and tumor samples with a given mutation into more and less sensitive subsets. We found that it is indeed able to do that, which leads us to the conclusion that it is the first pathway method to show a measurable improvement over using mutated genes as predictor of drug sensitivity and survival alone.

Speaker: Greg Slodkowicz

Patterns of adaptive evolution: a structural perspective

Protein structure is a major cause of site-to-site evolutionary rate variation. Many structural features such as solvent accessibility, local packing density and proximity to active sites or interfaces have been shown to modulate the evolutionary rate. It is, however, not well understood how these features affect the prevalence of adaptive evolution. Most codon-based models, which are commonly applied for detecting sites under positive selection, do not incorporate any information about the protein structure.

In this study, we attempted to form a better view of adaptation on molecular level by asking whether residues under positive selection are close to each other on the protein structure. We generated a large dataset of trees and alignments for 39 mammalian species (covering over 80% of human genes) and calculated sitewise values of selective constraint (dN/dS). We then mapped positively-selected sites onto available crystal structures and analysed whether they tend to be co-located by statistically assessing the distribution of pairwise distances between them.

We find that positively-selected sites frequently form tight clusters on protein structures and that this conclusion is robust to low alignment quality and other technical issues. Identified clusters can be assigned into one of several categories: we find that groups of positively-selected residues can surround active sites, occur in binding regions, and form small, linear clusters in the N-termini of proteins. To our knowledge, the last of these findings has not been previously reported. Additionally, the prevalence of clustering varies in different enzyme classes, with oxidoreductases exhibiting the most evidence for clustering.

Speaker: Rachel Spicer
The Lipidome in Weight Loss

In the UK 64% of adults are now either overweight or obese. Referral to commercial programs for weight management has been found to be an effective treatment in previous trials,. During the weight loss referrals for adults in primary care (WRAP), 1200 overweight and obese participants, from 3 centres across the UK, were randomly assigned to three treatment groups: a standardised brief intervention (BI) or a referral to a commercial programme for 12 (CP12) or 52 weeks (CP52). Participants allocated to the commercial treatment groups received vouchers to attend Weight Watchers sessions for either 12 or 52 weeks; those in the BI group were provided with a booklet on self help weight loss strategies. Fasting blood samples were then taken from participants at 0 and 12 months and blood plasma was extracted. Lipids were recovered from these samples using methyl-tert-butyl (MTBE) extraction. Following extraction, samples were then analysed using direct infusion mass spectrometry (DI-MS). This study aims to identify changes in individuals' lipidome profile composition between the inception of the study and a 12 month follow up appointment, as well as differences between the intervention groups.

Speaker: Omar Wagih
Predicting mechanistic consequences of single nucleotide variants in yeast

Abstract: A substantial amount of effort has gone into QTL mapping and GWAS-type studies, which provide us with potential causal variants that are associated with changes in phenotype. However, this does not provide us with the perturbed mechanism, leading to such changes. In this talk, I will give a brief overview of a pipeline we have been developing, which will allow annotation of variants that perturb cellular mechnanisms. These mechanisms include transcription factor binding, post-translational modifications, protein stability and protein-protein interactions. I will also discuss efforts that we have put into predicting phenotypes using using predicted deleterious variants.

Speaker: Maria Xenophontos
NuRD-mediated regulation of gene expression in pluripotent cells

The Nucleosome Remodelling and Deacetylation complex (NuRD) regulates gene expression in a chromatin-dependent manner and is required for differentiation of pluripotent cells. To enhance our understanding of NuRD's effect on developmental pathways we analysed the transcriptome and epigenome of wild-type and NuRD-depleted (*Mbd3*$^{-/-}$) embryonic (ESCs) and epiblast stem cells (EpiSCs). We find that in self-renewing ESCs loss of NuRD reinforces the core pluripotency network (*Nanog, Klf4* and *Klf5*), whereas in EpiSCs it results in down-regulation of endoderm and ectoderm markers (*Otx2, Eomes*, *Foxa2*). We also observe that in both ESCs and EpiSCs up- and down-regulation of transcription is a direct effect of NuRD depletion. Our analysis of the binding sites for two of the NuRD subunits, Mbd3 and Mi2b, revealed occupancy of active promoter and enhancer marks (H3K4me3 and H3K27ac) in pluripotent cells. Furthermore, in ESCs NuRD co-localises with core pluripotency transcription factors Oct4, Nanog, Sox2 and the transcriptional co-activator p300. These results point to the role of the NuRD complex as a modulator of transcription.

To better understand how NuRD function impacts gene expression we also study the effect of acute

NuRD recruitment by inducing Mbd3 expression in NuRD-depleted ESCs. In this system, NuRD reforms and associates with chromatin within 30 minutes. Induction of NuRD activity resulted in rapid changes in H3K27ac and H3K4me3 levels, however changes in gene expression are only detected 4 hours upon tamoxifen treatment.